First Year Econometrics

Charles Christopher Hyland

July 13, 2026

Contents

1 Probability Theory **6**
1.1 Axioms . . . 6
1.2 Distribution Functions and Densities . . . 7

2 Transformations and Expectations **11**
2.1 Transformations of Distributions . . . 11
2.2 Expectations of Distributions . . . 12
2.3 Multivariate Normal Distributions . . . 14
2.4 Normal Sampling Distributions . . . 15
2.4.1 Chi-Squared Distribution . . . 15
2.4.2 T and F Distributions . . . 18

3 Linear Regression Model and Linear Projection Model **20**
3.1 Conditional Expectation Function . . . 20
3.2 Linear Regression Model . . . 23
3.3 Linear Projection Model . . . 23

4 Regression and Matrix Algebra **26**
4.1 Ordinary Least Squares . . . 28
4.2 Partition Regression . . . 33
4.3 Frisch-Waugh-Lovell Theorem . . . 34
4.4 Residual Variance Decomposition . . . 35

5 Asymptotic Theory **37**
5.1 Convergence in Distribution and Probability . . . 37
5.1.1 Weak Law of Large Numbers . . . 38
5.1.2 Continuous Mapping Theorem . . . 40
5.1.3 Central Limit Theorem . . . 42
5.1.4 Delta Method . . . 44
5.2 Stochastic Order . . . 46

6 Maximum Likelihood Estimation **51**
6.1 Introduction and Set Up . . . 51
6.2 Score Function and Fisher Information Matrix . . . 52
6.3 MLE for Regression . . . 57
6.3.1 Homoskedastic Gaussian Regression . . . 57
6.4 Asymptotic Theory . . . 59
6.5 Hypothesis Testing . . . 60
6.5.1 Likelihood Ratio Test . . . 60
6.5.2 Wald Test . . . 61
6.5.3 Lagrange Multiplier Test . . . 61

7 The General Regression Model **62**
7.1 Sample and Population Models . . . 62
7.2 General Regression Model . . . 64
7.3 Classical Linear Regression Models . . . 66
7.4 Finite Properties in Regression Models . . . 68
7.5 Interpretation of Coefficients in Regression Models . . . 70
7.6 Inference in Regression Models . . . 71
7.6.1 Classical Linear Regression Model with Normally Distributed Error Terms . . . 71
7.6.2 Hypothesis Testing . . . 78
7.6.3 Using Linear Transformations to Simplify Hypothesis Tests . . . 83
7.6.4 Goodness of Fit . . . 84
7.6.5 Using R^2 to test exclusion restrictions . . . 86
7.7 Asymptotic Properties in Regression Models . . . 88
7.7.1 Consistency of OLS Estimator . . . 88
7.7.2 Linear Projection Model Revisited . . . 90
7.7.3 Asymptotic Normality . . . 91
7.7.4 Asymptotic Inference . . . 95

8 Heteroskedasticity and Generalised Least Squares **100**
8.1 Conditional Heteroskedasticity . . . 100
8.2 Generalised Least Squares . . . 102
8.2.1 Feasible Generalized Least Squares . . . 105
8.2.2 Weighted Least Squares . . . 106
8.3 Clustering . . . 108

9 Endogeneity **111**
9.1 Endogenous Explanatory Variable . . . 111
9.2 Measurement Error . . . 111
9.2.1 Multiple regression with errors in variables . . . 115
9.3 Omitted Variable . . . 116
9.3.1 Single Linear Regression with Omitted Variables . . . 116
9.3.2 Multiple Regression with Omitted Variables . . . 117

10 Instrumental Variables **119**
10.1 Two Stage Least Squares . . . 119
10.1.1 Just-Identified System . . . 125
10.1.2 L Instruments . . . 126
10.2 Control Function . . . 128
10.3 Indirect Least Squares . . . 129
10.4 Consistency of 2SLS Estimator . . . 130
10.5 Asymptotics . . . 132
10.6 IV with Multiple Regressors . . . 133
10.7 Testing . . . 134
10.7.1 Testing for Informativeness of Instrumental Variables . . . 135
10.7.2 Test for Validity of Instrumental Variables . . . 135
10.7.3 Test for Endogeneity . . . 136
10.8 Finite Sample Problems . . . 136
10.8.1 Overfitting . . . 137
10.8.2 Weak Instrumental Variables . . . 137
10.8.3 Anderson-Rubin Test . . . 137

11 Generalised Method of Moments **139**
11.1 Motivating Examples . . . 139
11.1.1 GMM Applied to Linear Regression . . . 139
11.1.2 GMM Applied to Linear Regression with Instrumental Variables . . . 142
11.2 GMM General Framework . . . 147
11.2.1 GMM Applied to IV Model . . . 148
11.3 Asymptotics . . . 149
11.3.1 Consistency of GMM Estimator . . . 149
11.3.2 Asymptotic Normality of GMM Estimator . . . 151
11.4 Identification . . . 155
11.4.1 Identification in Linear Model with Instrumental Variables . . . 156
11.5 Asymptotic Efficiency . . . 157
11.5.1 Efficiency in Linear IV Model . . . 158
11.6 Test For Overidentifying Restrictions . . . 159
11.7 Hypothesis Testing . . . 161
11.7.1 Linear Testing . . . 161
11.7.2 Nonlinear Testing . . . 162
11.8 Criterion-Based Test . . . 164

12 Time Series **166**
12.1 Introduction . . . 166
12.2 Cross Sectional Data Review . . . 166
12.2.1 Ordinary Least Squares . . . 166
12.2.2 Maximum Likelihood Estimation . . . 166
12.2.3 Partial Least Squares . . . 168
12.2.4 Correlation . . . 170

12.3 Autocorrelation . . . 171
12.4 Partial Autocorrelation . . . 174
12.5 Linear Projection in Time Series . . . 177
12.5.1 Linear Time Series Models . . . 180
12.6 Autoregression . . . 184
12.7 Autoregressive Model With Intercepts and Time Trends . . . 188
12.7.1 Random Walks . . . 192
12.8 Stationarity . . . 194
12.9 Unconditional Stationary Autoregression . . . 198
12.10 Asymptotic Theory for Time Series . . . 203
12.11 Limit Theorems for Martingale Difference Sequences . . . 205
12.12 Limit Theorems for AR(1) Model . . . 207
12.13 Higher Order Autoregression . . . 215
12.13.1 AR(2) Model . . . 215
12.13.2 AR(p) Model . . . 221
12.14 ADL Models . . . 225
12.15 VAR Models . . . 229
12.15.1 Granger Causality . . . 235
12.16 Misspecification Testing in Time Series . . . 237
12.17 Unit Root Testing . . . 238
12.17.1 Deterministic Terms . . . 241
12.18 Cointegration . . . 244
12.18.1 Beveridge-Nelson Decomposition . . . 244
12.18.2 Cointegration . . . 247
12.18.3 Cointegration for VAR(2) . . . 249
12.18.4 Cointegration for VAR(p) . . . 251
12.18.5 Granger Representation Theorem . . . 254

13 Panel Data **256**
13.1 Two-Way Error Components Model . . . 256
13.1.1 Time Dummies . . . 257
13.2 Pooled OLS in Linear Panel Model . . . 258
13.3 Fixed Effects Model . . . 263
13.3.1 Within Groups Estimator . . . 264
13.3.2 Least Squares Dummy Variables Estimator . . . 267
13.3.3 First-Differenced OLS . . . 268
13.3.4 Efficiency Comparison of Fixed Effects Estimators . . . 271
13.3.5 Large T Asymptotics of Fixed Effects Estimators . . . 272
13.4 Random Effects Model . . . 274
13.4.1 Between Groups Estimator . . . 274
13.4.2 Prereqs For Random Effects Generalised Least Squares Estimator . . . 276
13.4.3 Random Effects Generalised Least Squares Estimator . . . 278
13.4.4 Theta Differencing . . . 282
13.5 Testing for Random Effects . . . 285

13.6 Time-Invariant Explanatory Variables 287
13.7 Models with Predetermined Covariates 288
13.7.1 First Differenced 2SLS Estimator 288

1 Probability Theory

1.1 Axioms

We define Ω to be the **sample space** containing all possible outcomes. However, we cannot simply assign probabilities to every subset of Ω; we first specify a collection of measurable events. That is where a $\sigma - algebra$ comes in.

Definition 1 ($\sigma - algebra$)

A collection of sets $\mathcal{A}$ is a $\sigma - algebra$ if it satisfies the following properties:

1. $\Omega \in \mathcal{A}$;
2. If $A \in \mathcal{A}$ then the complement $A^c \in \mathcal{A}$;
3. If $A_1, A_2, ... \in \mathcal{A}$, then
$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}.$$

Now with a $\sigma - algebra$, we can define probability measures.

Definition 2 (Probability Measure)

A **probability measure** $\mathbb{P} : \mathcal{A} \to [0, 1]$ is a function satisfying:

1. $\mathbb{P}(\Omega) = 1$.
2. $\mathbb{P}(A) \in [0, 1]$ for all $A \in \mathcal{A}$.
3. If $A_1, A_2, ... \in \mathcal{A}$ are disjoint events, then
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

A triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is called a **probability space**.

Example 3

(Real Line). Let $\Omega = \mathbb{R}$. Then, the $\sigma - algebra$ is given by the **Borel field**, which is the $\sigma - algebra$ generated by intervals of the form $(-\infty, x]$. A probability measure over the Borel field induces the **cumulative distribution function** $F(x) = \mathbb{P}((-\infty, x])$.

We can now introduce random variables.

Definition 4 (Random Variable)

A random variable X is a measurable mapping $X : \Omega \to \mathbb{R}$ such that, for every Borel set $B \in \mathcal{B}$,
$$X^{-1}(B) \in \mathcal{A}.$$

1.2 Distribution Functions and Densities

We let X be a **random variable**. Every random variable has a distribution, which can be described by its **cumulative distribution function** $F(x) = \mathbb{P}(X \leq x)$. A discrete random variable has a **probability mass function**; when the distribution is absolutely continuous, it has a **probability density function** $f(x)$. In the latter case,

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Definition 5 (Cumulative Distribution Function)

A function $F(x)$ for $x \in \mathbb{R}$ is a **cumulative distribution function** if it satisfies:

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$;
2. F is non-decreasing: $F(x) \leq F(y)$ for $x \leq y$;
3. $F(x)$ is right-continuous: $\lim_{h\downarrow 0} F(x+h) = F(x)$.

The cumulative distribution function uniquely describes the probability distribution of a random variable X. One way to think about this is that a distribution is the blueprint for a house whilst a random variable is a particular instance of a house.

A distribution can have many random variables associated with it but a random variable can only have one distribution associated to it.

$$\text{Random Variable} \xrightarrow{\text{Induces}} \text{Distribution} \xrightarrow{\text{Described by}} \text{CDF (and, when it exists, a PMF or PDF)}$$

Example 6

(Normal Random Variable). When we say that X is a normal random variable, what we are actually saying is that the random variable X has a **probability density function** of the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

and the associated cumulative distribution function

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

Definition 7 (Atom)

An **atom** is defined as:

$$\mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = F(x) - \lim_{h\downarrow 0} F(x-h).$$

Definition 8

(Random Sample). The random variables $X_1, ..., X_n$ are called a **random sample** of size n from a population if they are independent and identically distributed. If their common distribution has a density, then $f_i(x) = f(x)$ for $i = 1, ..., n$.

Definition 9 (Joint Distribution Function)

A function $F(x_1, x_2) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2)$ for $x_1, x_2 \in \mathbb{R}$ is a **joint cumulative distribution function** if:

1. $F(x_1, x_2) \to 0$ if either coordinate tends to $-\infty$, and $F(x_1, x_2) \to 1$ as both coordinates tend to ∞;
2. $F(x_1, x_2)$ is non-decreasing in x_1, x_2;
3. $F(x_1, x_2)$ is right continuous in x_1, x_2;
4. the probability assigned to every rectangle $(a_1, b_1] \times (a_2, b_2]$ is non-negative.

The **marginal distribution function** is defined as:

$$F_1(x_1) = \lim_{x_2 \to \infty} F(x_1, x_2). \tag{1}$$

Definition 10 (Independence)

Random variables $X_1, ..., X_p$ are **independent** if their joint distribution function satisfies:

$$F(x_1, ..., x_p) = \prod_{i=1}^{p} F_i(x_i) \tag{2}$$

for all $x_1, ..., x_p$.

Proposition 11 (Joint Density Function of Random Sample)

With a random sample $(X_1, ..., X_n)$, we can use the independent and identically distributed assumption to write the **joint density** of our random sample as:

$$f(x_1, ..., x_n) = f_1(x_1) f_2(x_2) ... f_n(x_n) = \prod_{i=1}^{n} f(x_i) \tag{3}$$

Proof. The first equality follows from independence and the second equality follows from being identically distributed. □

Generally, when we have a random variable X with a probability density function $f(x)$ which is parameterised by θ, we usually write the probability density function as $f(x; \theta)$.

Example 12

(Joint density of exponential distribution). The density function of an exponential distribution is

$$f(x;\beta) = \frac{1}{\beta} e^{-x/\beta}, \qquad x \geq 0,\ \beta > 0.$$

Then, the joint distribution is

$$\begin{aligned} f(x_1, ..., x_n; \beta) &= \prod_{i=1}^{n} f(x_i; \beta) \\ &= \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-(x_1 + ... + x_n)/\beta} \end{aligned}$$

Now, suppose that we have our random sample $X_1, ..., X_n$, we are interested in **functions** of our random sample.

Definition 13 (Statistic)

Let $X_1, ..., X_n$ be a random sample of size n from a population. Let $T(x_1, ..., x_n)$ be a measurable real-valued or vector-valued **function** defined on the sample space. Then,

$$Y = T(X_1, ..., X_n)$$

is called a **statistic**. The probability distribution of a statistic Y is called the **sampling distribution** of Y.

Remark 14. *Since $X_1, ..., X_n$ are random variables, the statistic $Y = T(X_1, ..., X_n)$ is also a* **random variable**. *That is why the statistic has a distribution (and a density when one exists).*

Example 15

(Sample Mean) The sample mean is the classic example of a statistic

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Example 16

(Sample Variance) The sample variance is given by

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \tag{4}$$

Definition 17 (Conditional Density)

For $f_X(x) > 0$, the conditional density of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{5}$$

We can rewrite this as:

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x)$$

If x and y are independent, then:

$$f_{Y|X}(y|x) = f_Y(y)$$

since $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Definition 18 (Bayes Formula)

Bayes formula is given by:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{\mathbb{R}} f_{Y|X}(y|t)f_X(t)dt} \tag{6}$$

We now introduce some important probability density functions and cumulative distribution functions.

2 Transformations and Expectations

2.1 Transformations of Distributions

We state a useful theorem that we will use quite a few times.

Theorem 19 (Independence under Transformations)
Let $X_1, ..., X_n$ be independent random variables. Suppose that $g_1, ..., g_n$ are measurable functions from $\mathbb{R}$ to $\mathbb{R}$ and define $Y_i = g_i(X_i)$ for $i = 1, ..., n$. Then, the random variables $Y_1, ..., Y_n$ are independent.

Proof. Measurable functions of independent random variables are independent. □

Suppose we have a random variable X with a density f_X. Furthermore, suppose that we have a monotone function $g(\cdot)$ such that we construct the new random variable $Y = g(X)$. What is the density function of Y? The following theorem tells us this answer.

Theorem 20 (Change of Variables Formula)
Let X be a continuous random variable with density f_X and support $\mathcal{X}$. Let $g(\cdot)$ be a monotone mapping with a continuous derivative. Define the random variable $Y = g(X)$ with support $\mathcal{Y} = g(\mathcal{X})$. Then, the density of the random variable Y is given by:

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{\partial}{\partial y}g^{-1}(y)\right| \tag{7}$$

for all $y \in \mathcal{Y}$.

Proof. If $g(\cdot)$ is strictly monotone, then its inverse exists. For an increasing transformation,

$$\frac{\partial}{\partial y}(g^{-1}(y)) = \frac{1}{g'(g^{-1}(y))}$$

We then have that:

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(g(X) \leq y) \\ &= \mathbb{P}(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \end{aligned}$$

So we have:

$$F_Y(y) = F_X(g^{-1}(y))$$

Take derivatives of both sides with respect to y. Use the chain rule and plug in our expression for the derivative of the inverse:

$$f_Y(y) = \frac{d}{dy}\left[F_X(g^{-1}(y))\right]$$

and hence

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{1}{g'(g^{-1}(y))}\right|.$$

For a decreasing transformation the inequality is reversed, and the absolute value gives the same formula. □

2.2 Expectations of Distributions

This section will essentially state a bunch of results we will need throughout the course.

Definition 21 (Expectation)

If $\mathbb{E}[|X|] < \infty$ then the expectation of the random variable X is:

$$\mathbb{E}[X] = \int_{\mathbb{R}} xf(x)dx.$$

Remark 22. *The Cauchy distribution has no expectation.*

Proposition 23 (Linearity)

The expectation operator is a linear operator:

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Definition 24 (Moments)

The r-th **moment** of a random variable X is given by:

$$\mathbb{E}[X^r]$$

for $r > 0$. The r-th **central moment** is given by:

$$\mathbb{E}\big[(X - \mathbb{E}[X])^r\big]$$

for $r > 0$.

Definition 25 (Variance)

The variance is given by:

$$Var[X] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{8}$$

Standard deviation is the square root of variance:

$$SDV(X) = \sqrt{Var[X]}.$$

Proposition 26 (Quadratic Variance)

The variance is a quadratic function:

$$Var[\mu + \sigma X] = \sigma^2 Var[X].$$

Definition 27 (Jensen's Inequality)

If g is convex, then:

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

Proposition 28 (Linear Expectation)

Let X be a random vector. Let A and B be matrices. The expectation is a linear function:

$$\mathbb{E}[AX + B] = A\mathbb{E}[X] + B.$$

Proposition 29 (Quadratic Variance)

Let X be a random vector. Let A and B be matrices. The variance is a quadratic function:

$$Var[AX + B] = AVar[X]A^T$$

Definition 30 (Covariance)

The covariance between X and Y is:

$$Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{9}$$

Proposition 31 (Covariance Properties)

Let X, Y be random variables. Let a and b be scalars. Then:

$$Cov[aX + b, cY + d] = ac \cdot Cov[X, Y]$$

If X and Y are independent, then

$$Cov[X, Y] = 0.$$

The converse is **not** true.

Proposition 32 (Variance of Two variables)

Let X, Y be random variables. Let a and b be scalars. Then:

$$Var[aX + bY] = a^2Var[X] + b^2Var[Y] + 2abCov(X, Y).$$

Definition 33 (Correlation)

Correlation is a normalised covariance:

$$Corr[X, Y] = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Definition 34 (Cauchy-Schwarz Inequality)

The Cauchy-Schwarz inequality is:

$$|\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

Definition 35 (Conditional Expectation)

Conditional expectation of Y given X is:

$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} yf(y|x)dy.$$

Definition 36 (Law of Iterated Expectations)

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$$

$$Var[Y] = \mathbb{E}[Var[Y|X]] + Var[\mathbb{E}[Y|X]]$$

$$Cov(X, Y) = \mathbb{E}[Cov(X, Y|Z)] + Cov(\mathbb{E}[X|Z], \mathbb{E}[Y|Z])$$

2.3 Multivariate Normal Distributions

We now wish to define the multivariate normal distribution. This is a distribution we will use a lot throughout the course. First, we can define the standard multivariate normal distribution.

Definition 37 (Standard Multivariate Normal Distribution)

The random vector $\boldsymbol{X} = (X_1, ..., X_p)^T$ is a standard multivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ if it has the density function:

$$f(\boldsymbol{x}) = (2\pi)^{-p/2} \exp(-\boldsymbol{x}^T\boldsymbol{x}/2). \tag{10}$$

With the standard multivariate normal distribution, we can use the properties of the expectation and variance operators to define the regular multivariate normal distribution.

Definition 38 (Regular Multivariate Normal Distribution)

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. Define the random variable $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{AX}$ for $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{A} \in \mathbb{R}^{p\times p}$. Denote the covariance matrix $\Omega = \boldsymbol{AA}^T$ and assume that it is invertible.

Then, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Omega)$ with the density function:

$$f(\boldsymbol{y}) = \det(2\pi\Omega)^{-1/2} \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T\Omega^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\} \tag{11}$$

Notice that in the definition of a regular multivariate normal distribution, we required that the inverse of the covariance matrix Ω^{-1} to exist. However, what happens if this inverse does not exist? Therefore, we can define what is known as a **singular** multivariate normal distribution to handle this.

Definition 39 (Singular Multivariate Normal Distribution)

Let Ω be a symmetric positive semidefinite covariance matrix. We say that $\boldsymbol{Y}$ follows a **singular normal distribution** $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Omega)$ if, for every vector $\boldsymbol{V} \in \mathbb{R}^p$, one of the following holds for $\boldsymbol{V}^T\boldsymbol{Y}$:

1. $\boldsymbol{V}^T \cdot \boldsymbol{Y}$ has a regular univariate normal density function:

$$\boldsymbol{V}^T \cdot \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{V}^T\boldsymbol{\mu}, \boldsymbol{V}^T\Omega\boldsymbol{V}) \tag{12}$$

2.

$$\boldsymbol{V}^T \cdot \boldsymbol{Y} = \boldsymbol{V}^T \cdot \boldsymbol{\mu} \tag{13}$$

This definition also applies when the covariance matrix Ω is not invertible.

The multivariate normal has a few nice properties.

Proposition 40 (Expectation and Variance)

Let $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Omega)$. Then:

$$\mathbb{E}[\boldsymbol{Y}] = \mu$$

$$Var[\boldsymbol{Y}] = \Omega$$

Proposition 41 (Linear Transformations)

Let $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Omega)$. Then:

$$\boldsymbol{B} + \boldsymbol{C}\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{B} + \boldsymbol{C}\boldsymbol{\mu}, \boldsymbol{C}\Omega\boldsymbol{C}^T).$$

Proposition 42 (Independence and Zero Covariance for Jointly Normal Variables)

Let $(Y_1, Y_2)^T$ be jointly normally distributed. Then, Y_1 and Y_2 are independent if and only if $Cov(Y_1, Y_2) = 0$.

Proposition 43 (Convolution)

If $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then:

$$Y_1 + Y_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

2.4 Normal Sampling Distributions

2.4.1 Chi-Squared Distribution

We now move onto one of the most useful distributions besides the normal distribution in this course.

Definition 44 (Chi-squared Random Variable)

A random variable χ_p^2 is a chi-squared random variable with p degrees of freedom if it has the probability density function

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}$$

for $x \in (0, \infty)$. The p here is the **degrees of freedom** of the χ^2 distribution.

Remark 45. *We never really work with the density function of the χ^2 random variable in practice, but do know that it has one!*

We come to a **very important** result we will need throughout the course.

Theorem 46 (Multivariate Normal and Chi-Squared Distribution)
Suppose that we have a random variable $\boldsymbol{X}$ which has a multivariate normal distribution

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

whereby $\boldsymbol{X}$ has dimension d and the covariance matrix $\boldsymbol{\Sigma}$ has full rank d.

Then, we have that the following quadratic form of $\boldsymbol{X}$ has the distribution:

$$\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \sim \chi^2(d) \tag{14}$$

which is a χ^2 distribution with d degrees of freedom.

Proof. Recall from linear algebra.

Theorem 47 (Eigendecomposition of a matrix)
Let $\boldsymbol{\Sigma}$ be a real symmetric matrix. Then, we can express this matrix in terms of its eigendecomposition:

$$\boldsymbol{\Sigma} = \boldsymbol{V} \boldsymbol{D} \boldsymbol{V}^T \tag{15}$$

where $\boldsymbol{V}$ is an orthogonal matrix containing eigenvectors in every column and $\boldsymbol{D}$ is a diagonal matrix containing the corresponding eigenvalues.

We also need another little fact about orthogonal matrices.

Lemma 48
If $\boldsymbol{V}$ is an orthogonal matrix, then

$$\boldsymbol{V} \boldsymbol{V}^T = \mathbb{I}$$

whereby the transpose of $\boldsymbol{V}$ **is** the inverse of $\boldsymbol{V}$.

By our result on eigendecomposition, we can write the inverse of the covariance matrix as:

$$\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{V} \boldsymbol{D}^{-1/2} \boldsymbol{V}^T$$

whereby $\boldsymbol{D}^{-1/2}$ is simply taking the inverse square root of the diagonal entries of $\boldsymbol{D}$.

We now apply a change of variables whereby for $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, we can write the rescaled random variable:

$$\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{X}$$

We now want to derive the expectation and variance of this new random variable $\boldsymbol{Z}$.

$$\mathbb{E}[\boldsymbol{Z}] = \mathbb{E}[\boldsymbol{\Sigma}^{-1/2} \boldsymbol{X}] = \boldsymbol{\Sigma}^{-1/2} \mathbb{E}[\boldsymbol{X}] = 0$$

Since the mean of $\boldsymbol{Z}$ is zero, its covariance matrix is its second-moment matrix:

$$\begin{aligned}\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T] &= \boldsymbol{\Sigma}^{-1/2}\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\boldsymbol{\Sigma}^{-1/2} \\ &= \boldsymbol{\Sigma}^{-1/2}Var[\boldsymbol{X}]\boldsymbol{\Sigma}^{-1/2} \\ &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2} \\ &= \boldsymbol{V}\boldsymbol{D}^{-1/2}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}^{-1/2}\boldsymbol{V}^T \\ &= \boldsymbol{V}\boldsymbol{D}^{-1/2}\boldsymbol{D}\boldsymbol{D}^{-1/2}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{V}^T = \mathbb{I}\end{aligned}$$

We therefore have that $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I})$. Therefore, the sum of squares of $\boldsymbol{Z}$ is a χ^2-distribution. We can therefore conclude that:

$$\begin{aligned}\boldsymbol{Z}^T\boldsymbol{Z} &= \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X} \\ &= \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \sim \chi^2(d).\end{aligned}$$

□

We can extend this result to when the mean $\boldsymbol{\mu} \neq \boldsymbol{0}$.

Corollary 49

Suppose that we have a random variable $\boldsymbol{X}$ which has a multivariate normal distribution

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

whereby $\boldsymbol{X}$ has dimension d and the covariance matrix $\boldsymbol{\Sigma}$ has full rank d.

Then, we have that the following quadratic form of $\boldsymbol{X}$ has the distribution:

$$[\boldsymbol{X} - \boldsymbol{\mu}]^T\boldsymbol{\Sigma}^{-1}[\boldsymbol{X} - \boldsymbol{\mu}] \sim \chi^2(d) \tag{16}$$

which is a χ^2 distribution with d degrees of freedom.

Proposition 50 (Quadratic Form and Chi-squared Distribution)

If the $n \times 1$ vector $z \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and the non-stochastic $n \times n$ matrix $\boldsymbol{G}$ is symmetric and idempotent with $rank(\boldsymbol{G}) = r \leq n$, then the scalar:

$$w = \boldsymbol{z}^T\boldsymbol{G}\boldsymbol{z} \sim \chi^2(r). \tag{17}$$

The following result is a useful relationship between standard normal random variables and χ^2 random variables.

Proposition 51 (Chi-squared is the sum of squared standard normals)

Let $Z_i \sim \mathcal{N}(0, 1)$ for $i = 1, ..., n$ be independent standard normal random variables. Then

$$\boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

That is, the sum of n squared standard normal random variables is a χ^2 random variable with n degrees of freedom.

We get the useful corollary.

Proposition 52 (Chi-squared is the square of a standard normal)
If $Z \sim \mathcal{N}(0,1)$ then

$$Z^2 \sim \chi^2(1)$$

That is, the square of a standard normal random variable is a χ^2 random variable with 1 degree of freedom.

Proposition 53 (Chi-Squared and Orthogonal Projections)
Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be symmetric, idempotent matrices. Then,

$$\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X} \sim \chi^2(tr(\boldsymbol{P}))$$

$$\boldsymbol{X}^T\boldsymbol{Q}\boldsymbol{X} \sim \chi^2(tr(\boldsymbol{Q}))$$

Furthermore, if $\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{0}$, then $\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{Q}\boldsymbol{X}$ are independent.

2.4.2 T and F Distributions

Definition 54 (Student's t distribution)
The random variable T has a Student's t distribution with p degrees of freedom if T has the probability density function

$$f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{(p\pi)^{1/2}} \frac{1}{(1+t^2/p)^{(p+1)/2}}$$

for $t \in (-\infty, \infty)$.

We have the following result that relates a standard normal random variable, χ^2 random variable, and t-distributed random variable.

Proposition 55 (Deriving a t-distribution from a normal and chi-squared distribution)
Let $X \sim \mathcal{N}(0,1)$ be a standard normal random variable and let $Y \sim \chi^2(p)$ be independent of X. Then,

$$\frac{X}{\sqrt{Y/p}} \sim t_{(p)} \tag{18}$$

which is a t distribution with p degrees of freedom. The ratio of a standard normal random variable to the square root of an independent chi-squared random variable divided by its degrees of freedom has a Student t-distribution with those degrees of freedom.

Remark 56. *This says that the probability density function of the ratio $\frac{X}{\sqrt{Y/p}}$ is the Student's t probability density function. Note that we need to divide the χ^2 random variable by its degrees of freedom for this result to hold.*

Example 57

We know that if $X_1, ..., X_n$ are from a $\mathcal{N}(\mu, \sigma^2)$ sample, then we have that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \tag{19}$$

In practice, we do not know σ. Let $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ be the sample variance, which we can plug in:

$$\frac{\overline{X} - \mu}{\hat{\sigma}/\sqrt{n}} \tag{20}$$

The question we want to ask is what is the distribution of this new ratio. We can multiply and divide by σ to get:

$$\frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{\hat{\sigma}^2/\sigma^2}}$$

Now, we know that the numerator is:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

and the rescaled sample variance satisfies

$$\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1).$$

Equivalently, $\sqrt{\hat{\sigma}^2/\sigma^2}$ is the square root of this chi-squared variable divided by $n-1$. It is independent of the numerator, so the Student-t result implies that:

$$\frac{\overline{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t(n-1).$$

Definition 58 (F-distribution)

If the random variables $w_1 \sim \chi^2(m)$ and $w_2 \sim \chi^2(n)$ and w_1 and w_2 are independent, then the scalar:

$$\nu = \frac{(w_1/m)}{(w_2/n)} \sim F(m, n). \tag{21}$$

The ratio of two independent chi-squared random variables, each divided by their respective degrees of freedom, has a F-distribution with the degrees of freedom of the numerator and denominator respectively.

3 Linear Regression Model and Linear Projection Model

3.1 Conditional Expectation Function

We give a motivation for why we care so much about conditional expectations and how linear projection relates to them. Generally, we have a dependent variable of interest Y and want to learn about the factors driving it. For example, Y could be an individual's wage and X could be an individual's sex. We may then ask what the average wage of a female employee is and compare it with the average wage of a male employee. Using mathematics, we can express this using conditional expectations:

$$\mathbb{E}[\text{wage}|\text{sex} = \text{female}]$$
$$\mathbb{E}[\text{wage}|\text{sex} = \text{male}]$$

To generalise this, we are interested in the conditional expectation of Y given values that the variables $X_1, ..., X_k$ can take on:

$$\mathbb{E}[Y|X_1 = x_1, X_2 = x_2, ..., X_k = x_k] = m(x_1, ..., x_k).$$

We can write the vector $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_k \end{bmatrix} \in \mathbb{R}^k$ and from this, write the conditional expectation as:

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = m(\boldsymbol{x}).$$

Definition 59 (Conditional Expectation Function)
The conditional expectation function $m(\boldsymbol{x})$ is a function of $\boldsymbol{x} = (x_1, ..., x_k)$:

$$m(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]. \tag{22}$$

Remark 60. *We can interpret this as when $\boldsymbol{X}$ takes on the value $\boldsymbol{x}$, then the average value of Y is $m(\boldsymbol{x})$.*

We can also write the random variable:

$$m(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}].$$

Now, given the joint density $f(y, x)$, recall that the conditional density is given by:

$$f_{Y|X}(y|x) = \frac{f(y, x)}{f_X(x)}.$$

Then, this means we can write the CEF of Y given $\boldsymbol{X} = \boldsymbol{x}$ as:

$$m(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \int_{-\infty}^{\infty} y \cdot f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})dy.$$

which looks at all possible values of y given the probability of that value y, $f_{Y|\boldsymbol{X}}(y|x)$. We can interpret $m(\boldsymbol{x})$ as the

expectation of Y for a subpopulation whereby $\boldsymbol{X} = \boldsymbol{x}$. We can plug in $\boldsymbol{X}$ to get:

$$m(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}] = \int_{-\infty}^{\infty} y \cdot f_{Y|\boldsymbol{X}}(y|\boldsymbol{X})dy.$$

To recap, we have the variable Y and $\boldsymbol{X}$. From this, we constructed the CEF function $m(\boldsymbol{X})$. We can then define the CEF error.

Definition 61 (CEF Error)
The conditional expectation function error $\boldsymbol{e}$ is defined as the difference between Y and the CEF:

$$e \equiv Y - m(\boldsymbol{X}). \tag{23}$$

From this, we get the formula:

$$Y = m(\boldsymbol{X}) + e. \tag{24}$$

Proposition 62 (Properties of CEF Error)
Let e be the CEF error. Then,

1. $\mathbb{E}[e|\boldsymbol{X}] = 0$
2. $\mathbb{E}[e] = 0$
3. For any function $h(\boldsymbol{X})$, $\mathbb{E}[h(\boldsymbol{X})e] = 0$.

Proof. To show the first result:

$$\begin{aligned}\mathbb{E}[e|\boldsymbol{X}] &= \mathbb{E}[(Y - m(\boldsymbol{X}))|\boldsymbol{X}] \\ &= \mathbb{E}[Y|\boldsymbol{X}] - \mathbb{E}[m(\boldsymbol{X})|\boldsymbol{X}] \\ &= m(\boldsymbol{X}) - m(\boldsymbol{X}) = 0.\end{aligned}$$

To show the second result, we use the law of iterated expectations:

$$\mathbb{E}[e] = \mathbb{E}[\mathbb{E}[e|\boldsymbol{X}]] = \mathbb{E}[0] = 0.$$

For the last one:

$$\mathbb{E}[h(\boldsymbol{X})e] = \mathbb{E}[\mathbb{E}[h(\boldsymbol{X})e|\boldsymbol{X}]] = \mathbb{E}[h(\boldsymbol{X})\mathbb{E}[e|\boldsymbol{X}]] = \mathbb{E}[h(\boldsymbol{X}) \cdot 0] = 0.$$

□

This third property has a well-known name.

Definition 63 (Orthogonality Condition)
Let e be the CEF error. Then the condition:

$$\mathbb{E}[h(\boldsymbol{X})e] = 0 \tag{25}$$

for any measurable function $h(\cdot)$ is called the **orthogonality condition**.

The orthogonality condition implies that the covariance of $\boldsymbol{X}$ and e is zero.

Proposition 64 (Covariance of Regressor with CEF Error is Zero)
Suppose that the orthogonality condition holds $\mathbb{E}[h(\boldsymbol{X})e] = 0$. Then, the covariance of the regressor $\boldsymbol{X}$ with the CEF error is zero:

$$Cov[\boldsymbol{X}, e] = 0. \tag{26}$$

Proof. The covariance is:

$$Cov[\boldsymbol{X}, e] = \mathbb{E}[\boldsymbol{X}e] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[e] = 0 - 0 = 0$$

as $\mathbb{E}[\boldsymbol{X}e] = \mathbb{E}[h(\boldsymbol{X})e] = 0$ by the orthogonality condition where $h(\boldsymbol{X}) = \boldsymbol{X}$. □

We have a special name for the variance of the error term e.

Definition 65 (Unconditional Variance of Error Term)
The **unconditional variance of the error term** is:

$$\sigma^2 = Var[e]. \tag{27}$$

Proof. We have that:

$$\begin{aligned}\sigma^2 &= Var[e] \\ &= \mathbb{E}[e^2] - \mathbb{E}[e]^2 \\ &= \mathbb{E}[e^2] - 0 \\ &= \mathbb{E}[e^2]\end{aligned}$$

where e^2 is the second moment of error. □

Remark 66. *We can interpret σ^2 as the amount of variation in y that is not accounted for in the conditional expectation $\mathbb{E}[Y|\boldsymbol{X}]$.*

Intuitively, as you include more variables in your conditional expectation function, then the variance of your CEF error should drop.

Now, suppose we have the variables y and $\boldsymbol{X}$, and we want to predict Y using a function $g(\boldsymbol{X})$. We want to know what is the function $g(\cdot)$ that we can use as our best predictor (under mean squared error). We state the optimal function for this task.

Proposition 67 (Conditional Expectation Minimises the MSE)
The conditional expectation $\mathbb{E}[Y|\boldsymbol{X}]$ minimises the mean-squared error:

$$\mathbb{E}[(Y - g(\boldsymbol{X}))^2] \tag{28}$$

where $g(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$.

Proof. We want to find a $g(\boldsymbol{X})$ that minimises the mean squared error. Expand out the square:

$$\begin{aligned}\mathbb{E}[(Y - g(\boldsymbol{X}))^2] &= \mathbb{E}[(m(\boldsymbol{X}) + e - g(\boldsymbol{X}))^2] \\ &= \mathbb{E}[e^2] + 2\mathbb{E}[e(m(\boldsymbol{X}) - g(\boldsymbol{X}))] + \mathbb{E}[\big(m(\boldsymbol{X}) - g(\boldsymbol{X})\big)^2] \\ &=_{(1)} \mathbb{E}[e^2] + 0 + \mathbb{E}[\big(m(\boldsymbol{X}) - g(\boldsymbol{X})\big)^2] \\ &= \mathbb{E}[e^2] + \mathbb{E}[\big(m(\boldsymbol{X}) - g(\boldsymbol{X})\big)^2]\end{aligned}$$

where $=_{(1)}$ uses the orthogonality condition. Now, if we look at the last equation:

$$\mathbb{E}[e^2] + \mathbb{E}[\big(m(\boldsymbol{X}) - g(\boldsymbol{X})\big)^2]$$

we can see that we cannot choose a $g(\boldsymbol{X})$ to alter the first term, only the second term. The most we can do to minimise this equation is to set:

$$g(\boldsymbol{X}) = m(\boldsymbol{X})$$

as this sets the second term to be 0 (cannot be negative due to the square). □

Hence, we have shown that the conditional expectation function is the function that minimises the mean squared error. That is, the best predictor of Y is $E[Y|\boldsymbol{X}]$.

3.2 Linear Regression Model

We can now define a special model. Now, if we impose the assumption that the conditional expectation function **has** to be a linear function of $\boldsymbol{X}$, which we can denote by $\boldsymbol{X}^T\boldsymbol{\beta}$, then we get the linear regression model.

Definition 68 (Linear Regression Model)
The Linear Regression model is given by:

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + e \tag{29}$$

$$\mathbb{E}[e|\boldsymbol{X}] = 0. \tag{30}$$

Remark 69. *This linear conditional-mean specification is a special case of the general regression model discussed later.*

Under the linear regression model, $\mathbb{E}[e|\boldsymbol{X}] = 0$. We have already shown that this implies the error term is also unconditionally mean zero and orthogonal to the regressors:

$$\mathbb{E}[e] = 0$$
$$\mathbb{E}[\boldsymbol{X}e] = \boldsymbol{0}.$$

3.3 Linear Projection Model

We now want to derive a different model that is related to the linear regression model. We shall motivate it from another perspective. We know that $\boldsymbol{X}^T\boldsymbol{\beta}$ is a linear predictor of Y. The question we now want to ask is: what value of $\boldsymbol{\beta}$ minimises the mean squared error of predicting Y? This value is known as the **linear projection coefficient**.

Proposition 70 (Linear Projection Coefficient)

Let $\boldsymbol{X}^T\boldsymbol{\beta}$ be our linear predictor for Y. The **linear projection coefficient** minimises the mean squared error of predicting Y:

$$\min_{\boldsymbol{\beta}} \mathbb{E}\big[(Y - \boldsymbol{X}^T\boldsymbol{\beta})^2\big]. \tag{31}$$

The expression for $\boldsymbol{\beta}$ that minimises this is:

$$\boldsymbol{\beta} = (\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T])^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}]. \tag{32}$$

Proof. First, write out the mean squared error:

$$\mathbb{E}\big[(Y - \boldsymbol{X}^T\boldsymbol{\beta})^2\big] = \mathbb{E}[Y^2] - 2\boldsymbol{\beta}^T\mathbb{E}[\boldsymbol{X}Y] + \boldsymbol{\beta}^T\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\boldsymbol{\beta}.$$

Now, we take first order conditions of this. To do so, recall from matrix calculus:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\beta}^T = \boldsymbol{I}$$
$$\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta} = 2\boldsymbol{A}\boldsymbol{\beta}.$$

So taking first order conditions, we have that:

$$2\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\boldsymbol{\beta} = 2\mathbb{E}[\boldsymbol{X}Y]$$

Now, assuming invertibility, we can conclude that:

$$\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\mathbb{E}[\boldsymbol{X}Y].$$

□

Hence, we have argued that the best linear predictor of Y is:

$$\boldsymbol{X}^T\boldsymbol{\beta} = \boldsymbol{X}^T\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\mathbb{E}[\boldsymbol{X}Y]$$

We now notice that under this definition, we still have the orthogonality condition:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{X}e] &= \mathbb{E}[\boldsymbol{X}(Y - \boldsymbol{X}^T\boldsymbol{\beta})] \\
&= \mathbb{E}[\boldsymbol{X}Y - \boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\beta}] \\
&= \mathbb{E}[\boldsymbol{X}Y] - \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\beta}] \\
&= \mathbb{E}[\boldsymbol{X}Y] - \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\boldsymbol{\beta} \\
&= \mathbb{E}[\boldsymbol{X}Y] - \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\mathbb{E}[\boldsymbol{X}Y] \\
&= \mathbb{E}[\boldsymbol{X}Y] - \boldsymbol{I}\mathbb{E}[\boldsymbol{X}Y] \\
&= \boldsymbol{0}
\end{aligned}$$

Now, we can define our alternative model.

Definition 71 (Linear Projection Model)

The **linear projection model** is defined by:

$$\begin{aligned} Y &= \boldsymbol{X}^T\boldsymbol{\beta} + e \\ \mathbb{E}[\boldsymbol{X}e] &= \mathbf{0} \\ \boldsymbol{\beta} &= (\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T])^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}]. \end{aligned}$$

When you have a linear projection model, you are always guaranteed by definition:

$$\mathbb{E}[\boldsymbol{X}e] = \mathbf{0}.$$

Therefore, we still have that result that every regressor x_j is orthogonal to the error term e. Now, if we include an intercept $x_j = 1$, then we get that:

$$\mathbb{E}[1 \cdot e] = \mathbb{E}[e] = 0.$$

Hence, we can see now why we include an intercept in a linear regression model. By including an intercept term, we are guaranteed that the error term has mean 0 by the orthogonality condition:

$$\mathbb{E}[e] = 0.$$

If we don't have this, we no longer have any guarantees that the error term is mean zero.

Now, you should notice that strict exogeneity $\mathbb{E}[e|\boldsymbol{X}] = 0$ implies the orthogonality condition $\mathbb{E}[\boldsymbol{X}e] = \mathbf{0}$. This gives us the useful result that relates the linear regression and linear projection models we have defined.

Proposition 72 (Linear Regression and Linear Projection Model)

The linear regression model implies the linear projection model.

Remark 73. *The converse is not necessarily true. The projection error may not necessarily satisfy:*

$$\mathbb{E}[e|\boldsymbol{X}] = 0.$$

Therefore, the linear projection model is a weaker model compared to the linear regression model.

4 Regression and Matrix Algebra

Suppose we have a sample of n outcome variables (such as wage) we are interested in analysing. We can stack these outcome variables into a $n \times 1$ vector, which we denote by $\boldsymbol{y}$ as:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}.$$

Suppose that we have K explanatory variables of interest and n observations that may explain the outcome variable $\boldsymbol{y}$. We can stack these variables into a $n \times K$ data matrix:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \dots & x_{1K} \\ x_{21} & \dots & x_{2K} \\ . & \dots & . \\ . & \dots & . \\ . & \dots & . \\ x_{n1} & \dots & x_{nK} \end{bmatrix}$$

We will be working with this matrix a lot. We now state some definitions and properties of matrices that will be useful to us.

Definition 74 (Symmetric)
A matrix $\boldsymbol{A}$ is symmetric if:

$$\boldsymbol{A}^T = \boldsymbol{A}.$$

Proposition 75 (Symmetry of Matrices)
For any matrix $\boldsymbol{A}$, the matrices $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^T$ are **symmetric matrices**.

Definition 76 (Idempotent)
A square matrix $\boldsymbol{A}$ is idempotent if:

$$\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}.$$

Proposition 77 (Symmetric and Idempotent Matrices)
If $\boldsymbol{A}$ is a symmetric and idempotent matrix, then

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}.$$

Proof.

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$$

where the first equality uses symmetry and the second equality uses idempotency. □

Going back to our data matrix $\boldsymbol{X}$, a key assumption we make throughout the course is that we have more observations than explanatory variables. This makes sense or else we will not have enough degrees of freedom to estimate the parameters on our K explanatory variables if we have more parameters than observations.

Assumption 78 (Number of Observations)
For the $n \times K$ data matrix $\boldsymbol{X}$, we have that $n > K$.

We are now interested in the question of when is the matrix $\boldsymbol{X}^T\boldsymbol{X}$ invertible (this will become clear when we perform OLS). To do so, we need to build up some more definitions.

Definition 79 (Linear Independence)
Let $\boldsymbol{X}$ be a $n \times K$ matrix. The **column rank** of $\boldsymbol{X}$ is the number of **linearly independent columns**. The **row rank** of $\boldsymbol{X}$ is the number of **linearly independent rows**.

We show an important result regarding the rank of a matrix.

Theorem 80 (Column Rank Equals Row Rank)
The row rank is equal to the column rank.

Hence, so far, we know that as $\boldsymbol{X}$ is a $n \times K$ matrix, we have that:

$$rank(\boldsymbol{X}) \leq \min\{n, K\}$$

However, we generally assume that we have more observations than regressors in linear regression $K < n$. Therefore, this implies that the rank of $\boldsymbol{X}$ is:

$$rank(\boldsymbol{X}) \leq K.$$

Therefore, if $rank(\boldsymbol{X}) = K$, then $\boldsymbol{X}$ has **full rank**. In the case that $rank(\boldsymbol{X}) < K$, this has the interpretation that the model has **perfect multicollinearity**.

Proposition 81 (Rank of Design Matrix)
A useful result for the design matrix is:

$$rank(\boldsymbol{X}^T\boldsymbol{X}) = rank(\boldsymbol{X})$$

where $\boldsymbol{X}$ is a $n \times K$ matrix and $\boldsymbol{X}^T\boldsymbol{X}$ is a $K \times K$ matrix.

However, it may be hard to compute the rank so we can compute the **determinant** instead. We have a result that relates the rank of a matrix and the determinant of a matrix.

Theorem 82 (Rank and Determinant of a Matrix)
A square matrix $\boldsymbol{A}$ has full rank if and only if $\det(\boldsymbol{A}) \neq 0$.

Therefore, we can pull the three concepts of invertibility, rank, and determinant of a matrix $\boldsymbol{A}$ as follows:

1. $\boldsymbol{A}$ is invertible

2. $\boldsymbol{A}$ is full rank

3. $det(\boldsymbol{A}) \neq 0$

This applies to our matrix $\boldsymbol{X}^T\boldsymbol{X}$.

4.1 Ordinary Least Squares

Let $\boldsymbol{u}$ be the $n \times 1$ **residual vector**. An important thing to note is the order of the vector multiplication.

If we write $\boldsymbol{u}^T\boldsymbol{u}$, this gives us the **residual sum of squares** as:

$$\boldsymbol{u}^T\boldsymbol{u} = \begin{bmatrix} u_1 & \dots & u_n \end{bmatrix} \begin{bmatrix} u_1 \\ \dots \\ u_n \end{bmatrix} = \sum_{i=1}^{n} u_i^2.$$

Alternatively, if we write $\boldsymbol{u}\boldsymbol{u}^T$, we get an **outer-product matrix**:

$$\boldsymbol{u}\boldsymbol{u}^T = \begin{bmatrix} u_1 \\ \dots \\ u_n \end{bmatrix} \begin{bmatrix} u_1 & \dots & u_n \end{bmatrix} = \begin{bmatrix} u_1^2 & \dots & u_1u_n \\ . & \dots & . \\ u_nu_1 & \dots & u_n^2 \end{bmatrix}.$$

We prove the most important result of this course.

Theorem 83 (Ordinary Least Squares Estimator)
Assume that we have n observations and K explanatory variables. Suppose that we had the model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{33}$$

whereby $\boldsymbol{y}, \boldsymbol{u}$ are $n \times 1$ vectors, $\boldsymbol{X}$ is a $n \times K$ matrix, and $\boldsymbol{\beta}$ is a $K \times 1$ vector. Assume that the design matrix $\boldsymbol{X}$ is **full rank**. Then, the **ordinary least squares estimator** is given by:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{34}$$

Proof. We have that the model is:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

The goal of ordinary least squares is to minimise the residuals sum of squares by choosing a vector $\hat{\boldsymbol{\beta}}$ that does so. The residuals sum of squares can be written as:

$$\begin{aligned} \boldsymbol{u}^T\boldsymbol{u} &= [\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}]^T[\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}] \\ &= [\boldsymbol{y}^T - \boldsymbol{\beta}^T\boldsymbol{X}^T][\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}] \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &=_{(1)} \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

whereby $=_{(1)}$ uses the fact that $\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}$ is a scalar so we have that $\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y})^T = \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta}$. To proceed further, we need some important rules from matrix calculus.

Theorem 84 (Matrix Calculus Rules)

We have the following rules of matrix calculus:

$$\frac{\partial \boldsymbol{a}^T\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \boldsymbol{a}$$
$$\frac{\partial \boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = 2\boldsymbol{A}\boldsymbol{\beta}$$

for the vectors $\boldsymbol{a}, \boldsymbol{\beta}$ and the symmetric matrix $\boldsymbol{A}$.

To minimise the residuals sum of squares, we can take the first order conditions:

$$\frac{\partial \boldsymbol{u}^T(\hat{\boldsymbol{\beta}})\boldsymbol{u}(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} = \boldsymbol{0}$$

which then becomes:

$$\begin{aligned}\frac{\partial}{\partial \hat{\boldsymbol{\beta}}}\left[\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}\right] &= \frac{\partial \boldsymbol{y}^T\boldsymbol{y}}{\partial \hat{\boldsymbol{\beta}}} - 2\frac{\partial \boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}} + \frac{\partial \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}} \\ &= \boldsymbol{0} - 2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}\end{aligned}$$

We therefore have that our first order condition is:

$$\begin{aligned}-2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} &= 0 \\ \leftrightarrow \quad \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

By assumption, we assumed that $\boldsymbol{X}$ is full rank, which implies that $(\boldsymbol{X}^T\boldsymbol{X})$ is full rank and hence invertible, ie. $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists. Therefore, we have that:

$$\hat{\boldsymbol{\beta}}_{OLS} = [\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

□

We now look at what we have computed from the OLS estimation. We have computed that an estimate of our parameters is given by:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

We can use these estimated parameters to fit OLS to the data. For ease of notation, we shall rewrite $\hat{\boldsymbol{\beta}}_OLS$ as just $\hat{\boldsymbol{\beta}}$.

Definition 85 (Fitted Values and Residuals)

The **fitted values** are given by:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{35}$$

The **fitted residuals** are given by:

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}}. \tag{36}$$

Geometrically, we are projecting our outcome variable $\boldsymbol{y}$ onto the plane spanned by the explanatory variables $\boldsymbol{X}$. The projected point is the fitted value $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The difference between the original outcome vector $\boldsymbol{y}$ and this projection is the fitted residual $\hat{\boldsymbol{u}}$. We show this in Fig. 1.

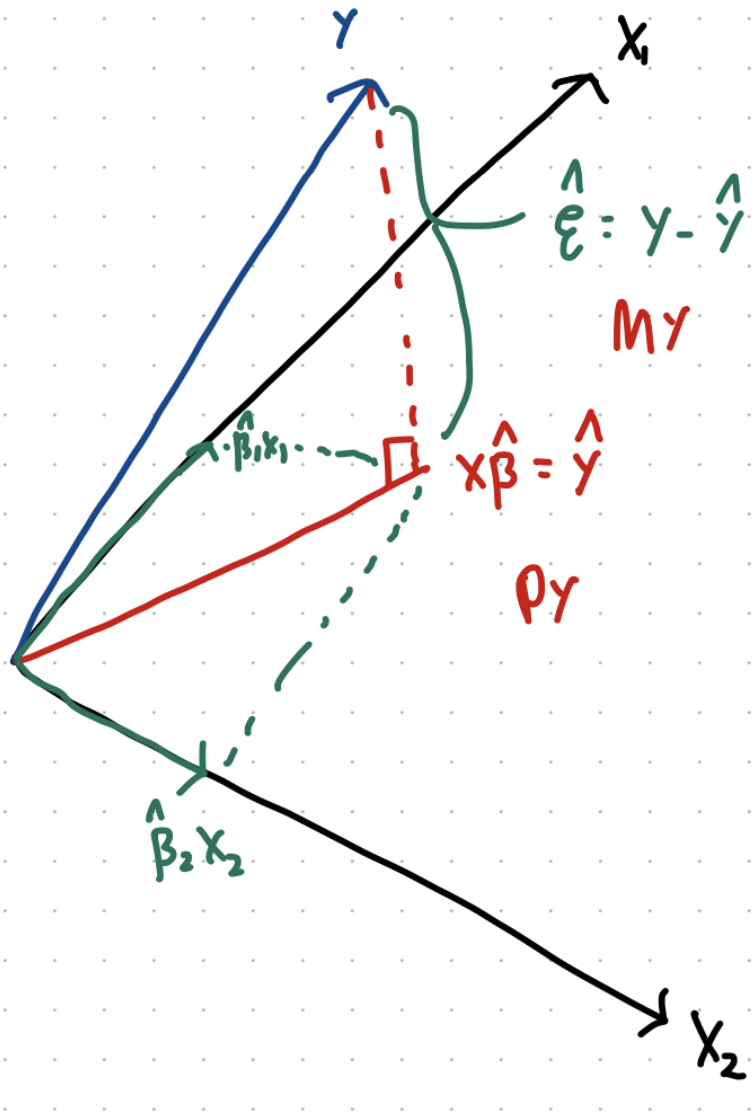

Figure 1: OLS is the projection of the observed variable onto the space spanned by the explanatory variables. This uses the example of having 2 explanatory variables.

Let us plug in our expression for the OLS estimator into our fitted value $\hat{y}$:

$$\hat{y} = X\hat{\beta} = X[(X^T X)^{-1} X^T y] = \underbrace{X(X^T X)^{-1} X^T}_{P} y$$

The matrix underlined by P has an important name.

Definition 86 (Projection Matrix)
The **projection matrix** P is given by:

$$P = X[X^T X]^{-1} X^T. \tag{37}$$

Remark 87. *This projection matrix is a matrix such that whenever it is applied to another vector/matrix B, it will project that matrix B onto the subspace that is spanned by X.*

Hence, we can write the fitted value of $\boldsymbol{y}$ as the projection of $\boldsymbol{y}$ onto the subspace spanned by $\boldsymbol{X}$:

$$\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y} \tag{38}$$

Now, if we write out fitted residuals, we get another important matrix:

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}[\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} = [I_n - \boldsymbol{P}]\boldsymbol{y}$$

From this, we can also define another important matrix that is thought of as the complement to the projection matrix.

Definition 88 (Annihilator Matrix)
The **annihilator matrix** is given by:

$$\boldsymbol{M} = I_n - \boldsymbol{P} \tag{39}$$

Remark 89. *The way to think of the annihilator matrix is that it produces the vector that is orthogonal to the space that is spanned by* $\boldsymbol{X}$.

Therefore, we can write the fitted residuals to be the annihilator matrix applied to the observed variables $\boldsymbol{y}$.

$$\hat{\boldsymbol{u}} = \boldsymbol{M}\boldsymbol{y} \tag{40}$$

The following result shows where the annihilator matrix gets its name from.

Proposition 90 (Annihilator matrix is Orthogonal to Design Matrix)
The annihilator matrix is orthogonal $\boldsymbol{M}$ to the design matrix $\boldsymbol{X}$:

$$\boldsymbol{M} \cdot \boldsymbol{X} = \boldsymbol{0}. \tag{41}$$

Proof.

$$\begin{aligned}
\boldsymbol{M} \cdot \boldsymbol{X} &= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{X} \\
&= \boldsymbol{X} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} \\
&= \boldsymbol{X} - \boldsymbol{X} \\
&= \boldsymbol{0}.
\end{aligned}$$

□

Proposition 91 (The Fitted Residuals are Orthogonal to Design Matrix)
The Fitted Residuals are Orthogonal to Design Matrix:

$$\boldsymbol{X}^T\hat{\boldsymbol{u}} = \boldsymbol{0} \tag{42}$$

Proof.

$$\begin{aligned}
\boldsymbol{X}^T\hat{\boldsymbol{u}} &= \boldsymbol{X}^T[\boldsymbol{M}\boldsymbol{y}] \\
&= \boldsymbol{X}^T[I_n - \boldsymbol{P}]\boldsymbol{y} \\
&= \boldsymbol{X}^T[I_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{y} \\
&= [\boldsymbol{X}^T - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{y} \\
&= [\boldsymbol{X}^T - \boldsymbol{I}_k\boldsymbol{X}^T]\boldsymbol{y} \\
&= [\boldsymbol{0}]\boldsymbol{y} = \boldsymbol{0}
\end{aligned}$$

□

We now show an important result.

Theorem 92 (Symmetry and Idempotency of $\boldsymbol{P}$ and $\boldsymbol{M}$)

The projection matrix $\boldsymbol{P}$ and the annihilator matrix $\boldsymbol{M}$ are:

1. Idempotent $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$ and $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$;
2. Symmetric $\boldsymbol{P}^T = \boldsymbol{P}$ and $\boldsymbol{M}^T = \boldsymbol{M}$.

We can also show another useful result.

Proposition 93 (Trace of the Annihilator Matrix)

The trace of the annihilator matrix $\boldsymbol{M}$ is given by:

$$tr(\boldsymbol{M}) = n - K \tag{43}$$

where n is the number of observations and K is the number of regressors.

Proof. To show this proposition, we need an important result regarding the trace operator:

Proposition 94 (Cyclic Properties of the Trace Operator)

Let A, B, C be matrices such that ABC is a **square matrix**. Then, the trace operator has the cyclic property:

$$tr(ABC) = tr(CAB) = tr(BCA).$$

Now we can use this cyclic property of the trace operator to help us.

$$\begin{aligned}
tr(\boldsymbol{M}) &= tr[\boldsymbol{I}_N - \boldsymbol{P}] \\
&= tr[\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\
&= tr[\boldsymbol{I}_N] - tr[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\
&=_{(1)} tr[\boldsymbol{I}_N] - tr[\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \\
&=_{(2)} tr[\boldsymbol{I}_N] - tr[\boldsymbol{I}_K] \\
&= n - K
\end{aligned}$$

where $=_{(1)}$ uses the cyclic property of the trace operator and $=_{(2)}$ uses the fact that $\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is a $K \times K$ matrix. □

To summarise, we can think of OLS as partitioning the outcome variable $\boldsymbol{y}$ into two components based on the projection and the residual from the projection:

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y}.$$

4.2 Partition Regression

Recall the formula for the OLS estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{44}$$

Suppose that we do not want to compute all the coefficients in $\hat{\boldsymbol{\beta}}$ but only a subset. To this end, let us partition the data matrix $\boldsymbol{X}$ into two sets of variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ whereby $\boldsymbol{X}_1$ is a $n \times K_1$ matrix and $\boldsymbol{X}_2$ is a $n \times K_2$ matrix. Without loss of generality, suppose we are interested in estimating the K_1 coefficients. Then, we have that:

$$\boldsymbol{y} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{bmatrix} + \epsilon = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \epsilon \tag{45}$$

Here, we are interested in estimating only $\boldsymbol{\beta}_1$. We know that the first order condition for OLS is given by:

$$\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

Therefore, substituting our partitioned matrices into the first order condition gives us:

$$\begin{bmatrix}\boldsymbol{X}_1^T\boldsymbol{y} \\ \boldsymbol{X}_2^T\boldsymbol{y}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1^T\boldsymbol{X}_1 & \boldsymbol{X}_1^T\boldsymbol{X}_2 \\ \boldsymbol{X}_2^T\boldsymbol{X}_1 & \boldsymbol{X}_2^T\boldsymbol{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{bmatrix}$$

We can then solve for our OLS parameters:

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1^T\boldsymbol{X}_1 & \boldsymbol{X}_1^T\boldsymbol{X}_2 \\ \boldsymbol{X}_2^T\boldsymbol{X}_1 & \boldsymbol{X}_2^T\boldsymbol{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{X}_1^T\boldsymbol{y} \\ \boldsymbol{X}_2^T\boldsymbol{y}\end{bmatrix}. \tag{46}$$

We are interested in figuring out what $\hat{\boldsymbol{\beta}}_1$ is. The first block of normal equations gives:

$$\boldsymbol{X}_1^T\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_1^T\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_1^T\boldsymbol{y},$$

so

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T(\boldsymbol{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2).$$

Now, suppose that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are **orthogonal** to each other so that:

$$\boldsymbol{X}_1^T\boldsymbol{X}_2 = \boldsymbol{0}.$$

Then, we have that our estimator for $\boldsymbol{\beta}_1$ becomes:

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y}$$

This has a very nice interpretation that if we want to get the estimators on the K_1 coefficients, we only need to regress $\boldsymbol{y}$ on the K_1 explanatory variables in $\boldsymbol{X}_1$ and we are done!

However, what if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are not orthogonal? Then we need to rely on the Frisch-Waugh-Lovell theorem in the next section.

4.3 Frisch-Waugh-Lovell Theorem

So far, we have arrived at the expression for the OLS parameters:

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1^T\boldsymbol{X}_1 & \boldsymbol{X}_1^T\boldsymbol{X}_2\\ \boldsymbol{X}_2^T\boldsymbol{X}_1 & \boldsymbol{X}_2^T\boldsymbol{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{X}_1^T\boldsymbol{y}\\ \boldsymbol{X}_2^T\boldsymbol{y}\end{bmatrix}. \tag{47}$$

Suppose that we want to derive an expression for $\hat{\boldsymbol{\beta}}_2$. We can do this in a two step procedure.

Step 1: Regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$

The first step is to regress $\boldsymbol{X}_2$ onto $\boldsymbol{X}_1$ and retrieve the residuals $\boldsymbol{X}_{2\cdot1}$:

$$\boldsymbol{X}_2 = \boldsymbol{X}_1\boldsymbol{b} + \boldsymbol{u} \tag{48}$$

where $\boldsymbol{b}$ is a $K_1 \times K_2$ matrix of parameters. This gives us the OLS estimator:

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2$$

where we project $\boldsymbol{X}_2$ onto the subspace spanned by $\boldsymbol{X}_1$. We can then retrieve the residuals $\boldsymbol{X}_{2\cdot1}$ as the true value $\boldsymbol{X}_2$ less the OLS projection:

$$\boldsymbol{X}_{2\cdot1} = \boldsymbol{X}_2 - \boldsymbol{X}_1\hat{\boldsymbol{b}}$$

What we have done by regression $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$ and retrieving the residuals is that we retrieved the variation in $\boldsymbol{X}_2$ that is **not explained** by the variation in $\boldsymbol{X}_1$. This can be thought of as purging the effects of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$.

Step 2: Regress $\boldsymbol{y}$ on $\boldsymbol{X}_{2\cdot1}$

We now regress $\boldsymbol{y}$ on the residuals $\boldsymbol{X}_{2\cdot1}$ and this gets us our desired estimator $\hat{\boldsymbol{\beta}}_2$. We now show why this would work. First, we can write out:

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}\\ &=_{(1)} \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} + \boldsymbol{X}_1\hat{\boldsymbol{b}}\boldsymbol{\beta}_2 - \boldsymbol{X}_1\hat{\boldsymbol{b}}\boldsymbol{\beta}_2\\ &= \boldsymbol{X}_1(\boldsymbol{\beta}_1 + \hat{\boldsymbol{b}}\boldsymbol{\beta}_2) + (\boldsymbol{X}_2 - \boldsymbol{X}_1\hat{\boldsymbol{b}})\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}\\ &=_{(2)} \boldsymbol{X}_1\boldsymbol{\delta}_1 + \boldsymbol{X}_{2\cdot1}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}\end{aligned}$$

where $=_{(1)}$ involved $\pm\boldsymbol{X}_1\hat{\boldsymbol{b}}\boldsymbol{\beta}_2$ and $=_{(2)}$ involved defining $\boldsymbol{\delta}_1 \equiv \boldsymbol{\beta}_1 + \hat{\boldsymbol{b}}\boldsymbol{\beta}_2$ and recognising that $\boldsymbol{X}_{2\cdot1} = \boldsymbol{X}_2 - \boldsymbol{X}_1\hat{\boldsymbol{b}}$ are the residuals from regression $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$.

Now, we again use our first order conditions to see that:

$$\begin{bmatrix}\hat{\boldsymbol{\delta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1^T\boldsymbol{X}_1 & \boldsymbol{X}_1^T\boldsymbol{X}_{2\cdot1}\\ \boldsymbol{X}_{2\cdot1}^T\boldsymbol{X}_1 & \boldsymbol{X}_{2\cdot1}^T\boldsymbol{X}_{2\cdot1}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{X}_1^T\boldsymbol{Y}\\ \boldsymbol{X}_{2\cdot1}^T\boldsymbol{Y}\end{bmatrix}$$

From this, we can state the expression for our desired estimator $\hat{\boldsymbol{\beta}}_2$.

Theorem 95 (Coefficients of Partition Regression)

The coefficients of the partition regression is given by:

$$\begin{bmatrix} \hat{\boldsymbol{\delta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{y} \\ (\boldsymbol{X}_{2\cdot 1}^T \boldsymbol{X}_{2\cdot 1})^{-1} \boldsymbol{X}_{2\cdot 1}^T \boldsymbol{y} \end{bmatrix} \tag{49}$$

Proof. Recall that we had:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1^T Y \\ X_2^T Y \end{bmatrix}$$

We can replace:

$$\begin{bmatrix} \hat{\boldsymbol{\delta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1^T \boldsymbol{X}_1 & \boldsymbol{X}_1^T \boldsymbol{X}_{2\cdot 1} \\ \boldsymbol{X}_{2\cdot 1}^T \boldsymbol{X}_1 & \boldsymbol{X}_{2\cdot 1}^T \boldsymbol{X}_{2\cdot 1} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1^T Y \\ \boldsymbol{X}_{2\cdot 1}^T Y \end{bmatrix}$$

Now, we want to show that the off-diagonal elements are zero.

Proposition 96 (Nuisance Vector and Residuals from First Stage Are Orthogonal)

$\boldsymbol{X}_1$ and $\boldsymbol{X}_{2\cdot 1}$ are orthogonal:

$$\boldsymbol{X}_1^T \cdot \boldsymbol{X}_{2\cdot 1} = \boldsymbol{0} \tag{50}$$

Proof.

$$\begin{aligned} \boldsymbol{X}_1^T \cdot \boldsymbol{X}_{2\cdot 1} &= \boldsymbol{X}_1^T \cdot (\boldsymbol{X}_2 - \boldsymbol{X}_1 \hat{\boldsymbol{b}}) \\ &= \boldsymbol{X}_1^T \cdot (\boldsymbol{X}_2 - \boldsymbol{X}_1 (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2) \\ &= \boldsymbol{X}_1^T \boldsymbol{X}_2 - \boldsymbol{X}_1^T \boldsymbol{X}_1 (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2 \\ &= \boldsymbol{X}_1^T \boldsymbol{X}_2 - \boldsymbol{X}_1^T \boldsymbol{X}_2 \\ &= \boldsymbol{0} \end{aligned}$$

□

Therefore, we have just shown that $\boldsymbol{X}_1^T \cdot \boldsymbol{X}_{2\cdot 1} = \boldsymbol{0}$ and so we have:

$$\begin{bmatrix} \hat{\boldsymbol{\delta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1^T \boldsymbol{X}_1 & 0 \\ 0 & \boldsymbol{X}_{2\cdot 1}^T \boldsymbol{X}_{2\cdot 1} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1^T Y \\ \boldsymbol{X}_{2\cdot 1}^T Y \end{bmatrix}$$

which gives us our desired result. □

Therefore, we can simply read off the second line in the theorem we have just proved to see our expression for our desired estimator as being:

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_{2\cdot 1}^T \boldsymbol{X}_{2\cdot 1})^{-1} \boldsymbol{X}_{2\cdot 1}^T \boldsymbol{Y}$$

This two-step procedure of regressing the variable of interest on the nuisance variables and then regressing the outcome on the residuals is known as the **Frisch-Waugh-Lovell theorem**.

4.4 Residual Variance Decomposition

Now, we want to derive another useful relationship.

Theorem 97 (Partial Sum of Squares Decomposition)
Partition $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$. Then:

$$\text{Restricted-model RSS} = \text{Incremental ESS from } \boldsymbol{X}_2 + \text{Unrestricted-model RSS}.$$

If $\boldsymbol{X}_1$ contains only an intercept, the restricted-model RSS is the usual centred total sum of squares.

Proof. First, start off with the fact that the fitted residuals are:

$$\hat{\boldsymbol{u}} = \boldsymbol{M}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y}$$

Then, the residual sum of squares is:

$$\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}} = \big((\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}\big)^T\big((\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}\big) = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P})^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$$

where we used the symmetry and idempotent of $\boldsymbol{M}$.

Now, partition the equation $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\delta}_1 + \boldsymbol{X}_{2\cdot1}\boldsymbol{\beta}_2 + \boldsymbol{u}$ whereby $\boldsymbol{X}_1^T\boldsymbol{X}_{2\cdot1} = \boldsymbol{0}$. Re-arrange this to get:

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\delta}}_1 - \boldsymbol{X}_{2\cdot1}\hat{\boldsymbol{\beta}}_2$$

Now, plug in our expression for $\hat{\boldsymbol{\delta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ we derived before to get:

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \underbrace{\boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T}_{\equiv\boldsymbol{P}_1}\boldsymbol{y} - \underbrace{\boldsymbol{X}_{2\cdot1}(\boldsymbol{X}_{2\cdot1}^T\boldsymbol{X}_{2\cdot1})^{-1}\boldsymbol{X}_{2\cdot1}^T}_{\equiv\boldsymbol{P}_{2\cdot1}}\boldsymbol{y}$$

and this becomes

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{P}_1\boldsymbol{y} - \boldsymbol{P}_{2\cdot1}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P}_1 - \boldsymbol{P}_{2\cdot1})\boldsymbol{y}$$

Then, this means that our residual sum of squares becomes:

$$\begin{aligned}
\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}} &= \big((\boldsymbol{I} - \boldsymbol{P}_1 - \boldsymbol{P}_{2\cdot1})\boldsymbol{y}\big)^T\big((\boldsymbol{I} - \boldsymbol{P}_1 - \boldsymbol{P}_{2\cdot1})\boldsymbol{y}\big) \\
&= \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_1 - \boldsymbol{P}_{2\cdot1})^T(\boldsymbol{I} - \boldsymbol{P}_1 - \boldsymbol{P}_{2\cdot1})\boldsymbol{y} \\
&= \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_1 - \boldsymbol{P}_{2\cdot1})\boldsymbol{y} \\
&= \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{P}_{2\cdot1}\boldsymbol{y}
\end{aligned}$$

So we then have that:

$$\underbrace{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}_{\text{unrestricted RSS}} = \underbrace{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}}_{\text{restricted RSS}} - \underbrace{\boldsymbol{y}^T\boldsymbol{P}_{2\cdot1}\boldsymbol{y}}_{\text{incremental ESS}}$$

Therefore, restricted RSS equals incremental ESS plus unrestricted RSS:

$$\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{P}_{2\cdot1}\boldsymbol{y} + \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_1 - \boldsymbol{P}_{2\cdot1})\boldsymbol{y}$$

□

5 Asymptotic Theory

5.1 Convergence in Distribution and Probability

We now state two important convergence concepts we will use frequently.

Definition 98 (Convergence in Distribution)
A sequence of random variables $X_1, X_2, \ldots$ **converges in distribution** to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{51}$$

at all points x where $F_X(x)$ is continuous, where F_{X_n} is the distribution function of X_n and F_X is the distribution function of X.

Remark 99. *The distribution to which the sequence X_n converges to is referred to as the* **limit distribution**.

Definition 100 (Convergence in probability)
A sequence of random variables $X_1, X_2, \ldots$ **converges in probability** to a random variable X if for every $\epsilon > 0$

$$\lim_{n\to\infty} P\big[|X_n - X| \geq \epsilon\big] = 0$$

Proposition 101 (Convergence in probability implies convergence in distribution)
Convergence in probability implies convergence in distribution. That is, if

$$X_n \xrightarrow{P} X$$

then

$$X_n \xrightarrow{D} X$$

Example 102
Let $X_1, X_2, \ldots$ be a sequence of random variables and suppose that X has a standard normal distribution. If

$$X_n \xrightarrow{P} X$$

By our proposition, this means that

$$X_n \xrightarrow{D} X$$

Therefore, the limiting random variable has the standard normal probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Proposition 103
The sequence of random variables $X_1, X_2, \ldots$ converges in probability to a constant μ if and only if the sequence converges in distribution to the constant μ.

Definition 104 (Consistent Estimator)

An estimator $\hat{\theta}_n$ is a consistent estimator of the parameter θ if:

$$\hat{\theta}_n \xrightarrow{P} \theta$$

Consistency is an asymptotic property. This is in contrast to unbiasedness, which is a finite-sample property. **Consistency does not imply unbiasedness and vice versa**. Consistency means that the distribution of $\hat{\theta}_n$ concentrates arbitrarily closely around θ as $n \to \infty$; convergence of the variance to zero is a useful sufficient condition when the relevant moments exist, but it is not part of the definition.

Example 105

Let $\hat{\theta}$ be an unbiased and consistent estimator of θ. Define the estimator:

$$\tilde{\theta} = \hat{\theta} + \frac{1}{n}$$

Clearly, this new estimator is biased:

$$\mathbb{E}[\tilde{\theta}] = \mathbb{E}[\hat{\theta} + \frac{1}{n}] = \theta + \frac{1}{n} \neq \theta.$$

However, it is consistent:

$$p \lim_{n\to\infty} \tilde{\theta} = p \lim_{n\to\infty} \hat{\theta} + p \lim_{n\to\infty} \frac{1}{n} = \theta + 0 = \theta.$$

5.1.1 Weak Law of Large Numbers

There are two versions of the weak law of large numbers we will need.

Theorem 106 (Weak Law of Large Numbers (1))

Let $X_1, X_2, \ldots$ be **independent** random variables with the conditions:

1. $\mathbb{E}[X_i] = \mu < \infty$
2. $Var[X_i] = \sigma^2 < \infty$

Then:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \xrightarrow{P} \mathbb{E}[X_i] = \mu.$$

Alternatively, we can write that:

$$p \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}[X_i] = \mu.$$

Theorem 107 (Weak Law of Large Numbers (2))

Let $X_1, X_2, ...$ be **independent and identically distributed** random variables with the condition:

1. $\mathbb{E}[X_i] = \mu < \infty$

Then:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \xrightarrow{P} \mathbb{E}[X_i] = \mu.$$

Alternatively, we can write that:

$$p \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}[X_i] = \mu.$$

Here, notice that version (1) of the weak law of large numbers (WLLN) required only independent random variables but we then required that the variance is finite. Meanwhile, in version (2), we required the random variables to also be identically distributed but then we only required that the first moment is finite. Whichever one you use depends on what is given to you.

Example 108

(WLLN Applied to Bernoulli Random Variables). Let $X_1, ..., X_n$ be i.i.d **Bernoulli random variables** with mean p. Then, we can apply the WLLN (2) to conclude that:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X_i] = p.$$

Separately, linearity of expectations gives $\mathbb{E}[\sum_{i=1}^{n} X_i] = np$; indeed, the sum has a **Binomial distribution**.

We show a very useful application of the weak law of large numbers to show a useful property of an important estimator.

Proposition 109 (Population Variance Estimator is Consistent)

Let $X_1, X_2, ...$ be i.i.d random variables with mean μ and variance σ^2. Then, the estimator of the variance:

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{52}$$

is a **consistent estimator** of σ^2.

Proof. First, we take the expectation of the terms inside the brackets:

$$\mathbb{E}[X_i - \mu] = \mathbb{E}[X_i] - \mu = \mu - \mu = 0.$$

Now, we know that for any random variable Z, the variance of Z is given by:

$$Var[Z] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$$

but if we have that the first moment of Z is $\mathbb{E}[Z] = 0$, then

$$Var[Z] = \mathbb{E}[Z^2].$$

We have shown that this is indeed the case for $X_i - \mu$ so we therefore have that:

$$Var[(X_i - \mu)] = \mathbb{E}[(X_i - \mu)^2] - \mathbb{E}[(X_i - \mu)]^2 = \mathbb{E}[(X_i - \mu)^2] - 0 = \mathbb{E}[(X_i - \mu)^2]$$

We now apply the weak law of large numbers (2) to our estimator:

$$\begin{aligned}\hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \xrightarrow{P} \mathbb{E}[(X_i - \mu)^2] \\ &=_{(1)} Var[X_i - \mu] \\ &=_{(2)} Var[X_i] \\ &= \sigma^2\end{aligned}$$

whereby $=_{(1)}$ comes from what we have shown earlier and $=_{(2)}$ comes from the fact that μ is a constant that is added and hence can be dropped from the variance. Therefore, we have shown that:

$$\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2$$

or equivalently:

$$p\lim_{n\to\infty} \hat{\sigma}_n^2 = \sigma^2$$

Therefore, $\hat{\sigma}_n^2$ is a **consistent** estimator. □

Remark 110. *In practice, we do not know μ and need to estimate it with $\hat{\mu}$. Replacing μ with $\hat{\mu}$ therefore requires an additional argument, which we give below.*

5.1.2 Continuous Mapping Theorem

Theorem 111 (Continuous Mapping Theorem)
Suppose that $X_1, X_2, \ldots$ converges in probability to a random variable X

$$X_n \xrightarrow{P} X$$

Let $h(\cdot)$ be a continuous function. Then, the **continuous mapping theorem** states that we have that:

$$h(X_n) \xrightarrow{P} h(X)$$

This also holds for convergence in distribution $\xrightarrow{D}$.

Example 112
(Consistency of standard deviation estimator). The sample variance S_n^2 is a consistent estimator for the variance σ^2. Define the continuous function $h(x) = \sqrt{x}$. Then,

$$S_n = \sqrt{S_n^2} = h(S_n^2)$$

is a consistent estimator of the standard deviation σ.

The continuous mapping theorem gives us some useful results.

Theorem 113 (Implications of Continuous Mapping Theorem)

Suppose that:

$$X_n \xrightarrow{P} X \quad \text{and} \quad Y_n \xrightarrow{P} Y.$$

Then:

1. $X_n + Y_n \xrightarrow{P} X + Y$;
2. $X_n \cdot Y_n \xrightarrow{P} X \cdot Y$.

Theorem 114 (Slutsky's Theorem)

Let $\{X_n\}_{n=1}^{\infty}$ and $\{Y_n\}_{n=1}^{\infty}$ be sequences of random variables. Let α be a constant. Suppose that $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} \alpha$. Then

1. $Y_n X_n \xrightarrow{D} \alpha X$
2. $X_n + Y_n \xrightarrow{D} X + \alpha$

Example 115

Let $X_1, X_2, \ldots$ be standard normal random variables with a limit $X \sim \mathcal{N}(0,1)$. Let $g(x) = x^2$. Then

$$g(X_i) = X_i^2 \xrightarrow{D} \chi^2(1).$$

Example 116

(Normal approximation with estimated variance). Suppose that we have

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \xrightarrow{D} \mathcal{N}(0,1)$$

where σ is unknown. We can plug in our sample variance estimator S_n^2. We are interested in the distribution of:

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S_n}$$

We know that the sample variance estimator is a consistent estimator, so we have that:

$$\frac{\sigma}{S_n} \xrightarrow{P} \frac{\sigma}{\sigma} = 1.$$

Then, by multiplying and dividing by σ, and invoking **Slutsky's theorem**, we have that

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S_n} = \frac{\sigma}{S_n} \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \xrightarrow{D} \mathcal{N}(0,1).$$

Proposition 117 (Sample Variance Estimator is Consistent)

Let $X_1, X_2, \ldots$ be i.i.d random variables with mean μ and variance σ^2. Then, the estimator of the variance:

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{53}$$

is a **consistent estimator** of σ^2.

Proof. First, we can rewrite the sample variance estimator as:

$$\hat{\sigma}_n^2 = \underbrace{\frac{n}{n-1}}_{\text{(A)}} \left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i^2}_{\text{(B)}} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}_{\text{(C)}} \right)$$

We now analyse each of these terms.

For (A), we have that clearly:

$$\frac{n}{n-1} = \frac{1}{1-1/n} \to 1.$$

Now for (B), recall that $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. We have that:

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2$$

if $\mathbb{E}[X] = \mu$ and $Var[X] = \sigma^2$.

Now, we know that $X_1, ..., X_n$ are i.i.d and $g(x) = x^2$ is a continuous function. This implies that $g(X_i) = X_i^2$ is also i.i.d. Hence, we have a sequence of i.i.d random variables with finite first moment. We can apply the **weak law of large numbers** to (B) to get:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} \mathbb{E}[X_i^2] = \sigma^2 + \mu^2.$$

We now look at (C). We have that for the term inside the square, we can again apply the **weak law of large numbers**:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$

Then, we can apply the **continuous mapping theorem**

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \xrightarrow{P} (\mu)^2 = \mu^2.$$

We now bring everything together. We first notice that $f(a, b) = a - b$ is a continuous function, so we can apply the continuous mapping theorem to conclude:

$$\hat{\sigma}_n^2 = \underbrace{\frac{n}{n-1}}_{\text{(A)}} \left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i^2}_{\text{(B)}} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}_{\text{(C)}} \right) \xrightarrow{P} 1\left[(\sigma^2 + \mu^2) - \mu^2\right] = \sigma^2$$

Therefore, we have shown that:

$$p\lim_{n\to\infty} \hat{\sigma}_n^2 = \sigma^2$$

is a consistent estimator. □

5.1.3 Central Limit Theorem

The central limit theorem states that the standardised mean, rescaled by $\sqrt{n}$, converges in distribution to a standard normal. That is, if $X_1, ..., X_n$ are i.i.d random variables with finite mean $\mathbb{E}[X_i] = \mu$ and variance $Var[X_i] = \sigma^2 > 0$, then:

$$\left[\frac{\overline{X}_n - \mathbb{E}[\overline{X}_n]}{\sqrt{Var[\overline{X}_n]}}\right] \xrightarrow{D} \mathcal{N}(0, 1) \tag{54}$$

whereby $\overline{X}_n$ is the sample average. However, we know that:

$$\mathbb{E}[\overline{X}_n] = \mu$$

and

$$Var[\overline{X}_n] = \frac{\sigma^2}{n}$$

Therefore, we can rewrite the central limit theorem as:

$$\left[\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}}\right] \xrightarrow{D} \mathcal{N}(0,1)$$

This gives us the more familiar form of the central limit theorem.

Theorem 118 (Lindeberg–Lévy Central Limit Theorem)
Let $X_1, ..., X_n$ be i.i.d random variables with finite mean $\mathbb{E}[X_i] = \mu$ and variance $Var[X_i] = \sigma^2 > 0$. Then, the standardised sample average has the following limiting distribution:

$$\sqrt{n}\left[\frac{\overline{X}_n - \mu}{\sigma}\right] \xrightarrow{D} \mathcal{N}(0,1). \tag{55}$$

We also state the central limit theorem for random vectors.

Theorem 119 (Central Limit Theorem for Random Vectors)
Suppose that $\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n$ are independent and identically distributed random vectors with $\mathbb{E}[\boldsymbol{z}_i] = \boldsymbol{\mu}$ and finite covariance matrix $Var[\boldsymbol{z}_i] = \boldsymbol{\Sigma}$. Denote $\overline{\boldsymbol{z}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i$ as the vector of sample means for a sample of size n. Then, the central limit theorem states that:

$$\sqrt{n}[\overline{\boldsymbol{z}}_n - \boldsymbol{\mu}] \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{56}$$

Example 120

(Asymptotic Normality of t-statistic) Define the t-test statistic:

$$t_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\hat{\sigma}^2}}$$

In the numerator, we have the **central limit theorem**:

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

In the denominator, we know that the sample variance estimator is consistent:

$$\hat{\sigma}^2 \xrightarrow{P} \sigma^2.$$

By the **continuous mapping theorem**

$$\frac{1}{\sqrt{\hat{\sigma}^2}} \xrightarrow{P} \frac{1}{\sqrt{\sigma^2}}$$

Now, we can apply **Slutsky's theorem** to the t-test statistic to get:

$$t_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\hat{\sigma}^2}} \xrightarrow{D} \frac{1}{\sigma}\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, 1)$$

which only holds if $\sigma^2 > 0$ due to the continuous mapping theorem.

5.1.4 Delta Method

Sometimes we are interested in the distribution of a **function** of a random variable. The delta method tells us the limiting distribution of $g(X_n)$. For intuition, we take a first-order Taylor approximation of $g(X_n)$ and show that the remainder is asymptotically negligible.

Theorem 121 (Delta Method)

Let X_n be a sequence of random variables such that

$$\sqrt{n}\big(X_n - \mu\big) \xrightarrow{D} \mathcal{N}(0, \sigma^2) \tag{57}$$

Suppose that we have a continuously differentiable function $g : \mathbb{R} \to \mathbb{R}$. Then, the **delta method** states that:

$$\sqrt{n}\big(g(X_n) - g(\mu)\big) \xrightarrow{D} \mathcal{N}(0, \sigma^2\big[g'(\mu)\big]^2) \tag{58}$$

provided that $g'(\mu)$ exists.

Proof. (Sketch). Differentiability at μ gives the first-order expansion

$$g(X_n) - g(\mu) = g'(\mu)(X_n - \mu) + o_p(|X_n - \mu|).$$

Multiplying by $\sqrt{n}$ and using the assumed limiting distribution yields

$$\sqrt{n}\big(g(X_n) - g(\mu)\big) = g'(\mu)\sqrt{n}(X_n - \mu) + o_p(1).$$

Slutsky's theorem therefore gives

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{D} \mathcal{N}(0, \sigma^2[g'(\mu)]^2).$$

□

An alternative way to think about the delta method is to think that $g(X_n)$ is asymptotically normal:

$$g(X_n) \overset{a}{\sim} \mathcal{N}\left(g(\mu), \frac{\sigma^2[g'(\mu)]^2}{n}\right)$$

However, it is important to note that this is an approximation as it uses a Taylor series approximation!!!

Example 122

(Squared Sample Average). Suppose $X_1, ..., X_n$ are i.i.d with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. We are interested in the limiting distribution of $\overline{X}_n^2$. By the **central limit theorem**

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

Define $g(y) = y^2$ so that:

$$g(\overline{X}_n) = \overline{X}_n^2$$

and

$$g(\mu) = \mu^2 \Rightarrow g'(\mu) = 2\mu$$

This means that

$$(g'(\mu))^2 = 4\mu^2.$$

We can therefore apply the delta method formula to conclude that:

$$\sqrt{n}(\overline{X}_n^2 - \mu^2) \xrightarrow{D} \mathcal{N}(0, 4\mu^2\sigma^2).$$

However, at $\mu = 0$, we must treat this differently as we get a degenerate distribution $\mathcal{N}(0, 0)$.
First, we notice for $\mu = 0$,

$$\sqrt{n}(\overline{X}_n - 0) = \sqrt{n}(\overline{X}_n) \xrightarrow{D} \mathcal{N}(0, \sigma^2) = \sigma\mathcal{N}(0, 1)$$

by the central limit theorem. Therefore, we have that:

$$\left(\sqrt{n}\overline{X}_n\right)^2 = n\overline{X}_n^2 \xrightarrow{D} (\sigma\mathcal{N}(0, 1))^2 = \sigma^2\chi^2(1).$$

Example 123

(Average of Poisson Random Variables). Let $X_1, ..., X_n$ be i.i.d $Pois(\lambda)$. Define

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We know $\mathbb{E}[X_i] = Var[X_i] = \lambda$.

By **central limit theorem**

$$\sqrt{n}\left[\frac{\overline{X}_n - \lambda}{\sqrt{\lambda}}\right] \xrightarrow{D} \mathcal{N}(0,1)$$

or

$$\sqrt{n}(\overline{X}_n - \lambda) \xrightarrow{D} \mathcal{N}(0,\lambda).$$

Apply $g(x) = \sqrt{x}$. We have that:

$$g(\overline{X}_n) = \sqrt{\overline{X}_n} \quad g(\lambda) = \sqrt{\lambda}$$

Then we have

$$g'(\lambda) = \frac{\lambda^{-1/2}}{2}$$

and

$$(g'(\lambda))^2 = \frac{\lambda^{-1}}{4}$$

Then, by delta theorem:

$$\sqrt{n}(\sqrt{\overline{X}_n} - \sqrt{\lambda}) \xrightarrow{D} \frac{\lambda^{-1/2}}{2}\mathcal{N}(0,\lambda) = \mathcal{N}(0, \frac{1}{4\lambda}\lambda) = \mathcal{N}(0, \frac{1}{4}).$$

This is known as a variance stabilising transformation as the variance no longer depends on λ.

Remark 124. *Note that we are saying that for every possible parameter value θ, our estimator $\hat{\theta}$ needs to converge in probability to the true parameter θ.*

5.2 Stochastic Order

We now are interested in formalising the ideas we have talked about so far regarding convergence of random variables. What we will talk about here is similar to convergence for ordinary sequences, but for random variables. Therefore, probabilities will play an important role. First, we ignore probabilities.

Definition 125 (Big-Oh)

Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of constants. We say that $a_n = \mathcal{O}(1)$ if there exists a constant C such that

$$|a_n| \leq C$$

for all $n \geq 1$.

Remark 126. *This is saying that the sequence $\{a_n\}_{n=1}^{\infty}$ is bounded by the constant C.*

Generally, if we have two sequences of constants $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we want to be able to relate them to each other. The following definition shows us how:

Definition 127 (Big-Oh Part Two)

Let $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ be sequences of constants. Then, we can write that $a_n = \mathcal{O}(b_n)$ if

$$\frac{a_n}{b_n} = \mathcal{O}(1)$$

Remark 128. *This is equivalent to saying that the sequence of ratios is bounded*

$$\left|\frac{a_n}{b_n}\right| \leq C$$

for all $n \geq 1$. *We can see that* $\mathcal{O}(1)$ *is a special case of* $\mathcal{O}(b_n)$ *whereby* $b_n = 1$.

Definition 129 (Little-Oh)

Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of constants. Then, suppose that:

$$a_n \to 0$$

We can then write that $a_n = o(1)$.

Remark 130. *Note that this is referring to ordinary convergence of sequences going to 0.*

Example 131

Let $a_n = \frac{1}{n}$. Clearly

$$\frac{1}{n} \to 0$$

as $n \to \infty$. Then, we can write that $a_n = o(1)$.

Definition 132 (Little-Oh Part Two)

Let $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ be sequences of constants. Then, suppose that:

$$\frac{a_n}{b_n} \to 0$$

We can then write that $a_n = o(b_n)$.

Remark 133. *If you compare this definition to our first definition of* $o(1)$, $a_n = o(1)$ *is the case of* $a_n = o(b_n)$ *where* $b_n = 1$.

We now want to talk about what it means for a random variable to be **bounded**. Since it is a random variable, it does not make sense to talk about being bounded without involving probabilities. Therefore, we can show how to define what it means for a random variable being stochastically bounded.

Definition 134 (Bounded in probability)

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. We say that the sequence of random variables X_n is **bounded in probability** if for every $\epsilon > 0$, there exists constants δ and N such that

$$\mathbb{P}\Big[|X_n| > \delta\Big] < \epsilon$$

for all $n > N$. We write that

$$X_n = O_p(1) \tag{59}$$

Like what we did before, we can also extend our definition of being bounded in probability.

Definition 135 (Bounded in probability Part Two)

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and let $\{a_n\}_{n=1}^{\infty}$ be a sequence of constants. We say that the sequence of random variables is **bounded in probability** if for every $\epsilon > 0$, there exists constants δ and N such that

$$\mathbb{P}\Big[|\frac{X_n}{a_n}| > \delta\Big] < \epsilon$$

for all $n > N$. We write that

$$X_n = O_p(a_n) \tag{60}$$

Let us break down what this is saying. We are saying that if we take our sequence of random variables X_n, divide it by our sequence of constants a_n, then the probability that this ratio X_n/a_n is greater than δ is smaller than ϵ (which we defined to be a small number) for all random variables X_n where their index $n > N$. N here could be any number such as 100. If that was the case, then that means $X_{101}, X_{102}, X_{103}, ...$ will all satisfy this property:

$$\mathbb{P}\Big[|\frac{X_n}{a_n}| > \delta\Big] < \epsilon$$

for $n \geq 101$.

We can also define an equivalent to little o in terms of probability.

Definition 136 (Little op)

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. We say that X_n converges in probability to zero if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n| \geq \epsilon\Big) = 0 \tag{61}$$

That is,

$$X_n \xrightarrow{P} 0$$

We write that

$$X_n = o_p(1) \tag{62}$$

Definition 137 (Little op Part Two)

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and let $\{a_n\}_{n=1}^{\infty}$ be a sequence of nonzero constants. We say that X_n/a_n converges in probability to zero if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_n}{a_n}\right| \geq \epsilon\right) = 0 \tag{63}$$

That is,

$$\frac{X_n}{a_n} \xrightarrow{P} 0$$

We write that

$$X_n = o_p(a_n) \tag{64}$$

When we talk about big-oh and little-oh notation, we are not actually talking about strict equality i.e. $X = o_p(1)$. It is instead a way for us to write the limits or talk about the convergence of random variables.

Example 138

Suppose that we have a sequence of random variables $\{X_n\}$ such that it converges in probability to X:

$$X_n \xrightarrow{P} X$$

Then, we can rewrite our expression for X_n as

$$X_n = X + o_p(1)$$

which is another way of writing that the difference $X_n - X$ goes to zero in probability.

We now show some interesting properties of $\mathcal{O}_p$ and o_p.

Lemma 139

$$o_p(1) + o_p(1) = o_p(1)$$

Example 140

Suppose that $X_n = o_p(1)$ and $Y_n = o_p(1)$. This implies that $X_n \xrightarrow{P} 0$ and $Y_n \xrightarrow{P} 0$. Then, by the continuous mapping theorem

$$X_n + Y_n \xrightarrow{P} 0$$

Lemma 141

$$\mathcal{O}_p(1) + o_p(1) = \mathcal{O}_p(1)$$

Proof. If one term is bounded in probability whilst the other term is going to 0, then the sum of the two will also be bounded in probability. □

Lemma 142

$$\mathcal{O}_p(1)o_p(1) = o_p(1)$$

Proof. If one term is bounded in probability and the other term goes to 0 in probability, then their product will also go to 0 in probability. □

Lemma 143

Let A be a nonzero constant. Then

$$o_p(A) = A \times o_p(1)$$

Proof. Suppose that $X_n = o_p(A)$. This implies that $\frac{X_n}{A} = o_p(1)$. Then, we get that

$$X_n = Ao_p(1).$$

□

Theorem 144 (Convergence in distribution with big oh-pee)

If $X_n \xrightarrow{D} X$, then $X_n = \mathcal{O}_p(1)$.

Theorem 145 (Little oh-pee implies big oh-pee)

If $X_n = o_p(a_n)$, then $X_n = \mathcal{O}_p(a_n)$.

Remark 146. *Little oh-pee is a stronger property than big oh-pee.*

Proposition 147 (Linking deterministic and stochastic order)

Deterministic order implies stochastic order:

1. If $X_n = o(1)$ then $X_n = o_p(1)$
2. If $X_n = \mathcal{O}(1)$ then $X_n = \mathcal{O}_p(1)$

Example 148

Suppose that we have that $n^{-1} \to 0$ and $X_n \xrightarrow{D} X$. First, we have that $X_n = \mathcal{O}_p(1)$. We also have that $n^{-1} = o(1) = o_p(1)$. Therefore, we can conclude that:

$$n^{-1}X_n = o_p(1)\mathcal{O}_p(1) = o_p(1).$$

6 Maximum Likelihood Estimation

6.1 Introduction and Set Up

Suppose we had observed data $\boldsymbol{y} = (y_1, ..., y_n)$. We assume that this data has the **joint probability density**:

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = f(y_1, ..., y_n; \boldsymbol{\theta})$$

We assume that the model is **correctly specified**. This means that there exists a parameter $\boldsymbol{\theta}_0 \in \Theta$ such that the data is actually generated by the density function:

$$\boldsymbol{y} \sim f(\cdot; \boldsymbol{\theta}_0).$$

Here, the data is **fixed** whilst the parameter $\boldsymbol{\theta}$ varies.

Definition 149 (Likelihood Function)
The **likelihood** function for $\boldsymbol{y}$ is defined as:

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) \equiv f(\boldsymbol{y}; \boldsymbol{\theta}) \tag{65}$$

where $f(\boldsymbol{y}; \boldsymbol{\theta})$ is the joint density function of $\boldsymbol{y}$. The **log-likelihood** function is the logarithm of the likelihood function:

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) \equiv \log(\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})). \tag{66}$$

In general, it is easier to compute the likelihood function if the observations are i.i.d.

Proposition 150 (Likelihood Function Under I.I.D Data)
Suppose that the observations $y_1, ..., y_n$ are i.i.d. Then, the likelihood function for $\boldsymbol{y}$ is:

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}). \tag{67}$$

The log-likelihood function is given by:

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \log(\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})) = \sum_{i=1}^{n} \log(f(y_i; \boldsymbol{\theta})) \tag{68}$$

Remark 151. *Suppose that we fixed the parameter $\boldsymbol{\theta}$. Now, a very important thing to realise is that $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ is a* **random variable** *as it is just a function of the random variables $\boldsymbol{y}$ when we fixed $\boldsymbol{\theta}$.*

The likelihood function measures the relative support that the observed data $\boldsymbol{y}$ provide for different values of the parameter $\boldsymbol{\theta}$. Given this, we want to find the parameter value with the greatest likelihood. This leads us to the maximum likelihood estimator.

Definition 152 (Maximum Likelihood Estimator)

The **maximum likelihood estimator** is the estimator $\hat{\boldsymbol{\theta}}_{MLE}$ of the parameter $\boldsymbol{\theta}$ that maximises the (log) likelihood function given the data $\boldsymbol{y}$. That is,

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}; \boldsymbol{y})$$

Remark 153. *Notice that due to monotonicity, maximising the likelihood function or the log-likelihood function gives us the same result for the MLE.*

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{MLE}$ maximises the log-likelihood function such that for any other estimator $\tilde{\boldsymbol{\theta}}$,

$$\ell(\hat{\boldsymbol{\theta}}_{MLE}; \boldsymbol{y}) \geq \ell(\tilde{\boldsymbol{\theta}}; \boldsymbol{y})$$

for all possible parameters $\boldsymbol{\theta} \in \Theta$.

6.2 Score Function and Fisher Information Matrix

We have defined the log-likelihood function and the MLE as the estimate that maximises this function. Hence, MLE is the process of finding the maximum of the log-likelihood function. Now, the sharper the curvature of the log-likelihood function around the maximum point, the more accurate it will be to locate this maximum point. The more curved a log-likelihood function is, the more accurate our MLE will be. Hence, we are interested in analysing the derivatives of the log-likelihood function. These derivatives have important names in Statistics.

Definition 154 (Score Function)

The random variable that is the first order derivatives of the log-likelihood function is called the **Score function**:

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} \tag{69}$$

The Score function indicates the steepness of the log-likelihood function and thereby the sensitivity to infinitesimal changes to the parameter values.

Example 155

Suppose that $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. Then, the score function is given by:

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \ell}{\partial \mu}, \frac{\partial \ell}{\partial \sigma^2} \right)^T.$$

We can in fact characterise the MLE by the Score function.

Proposition 156 (Necessary Conditions for MLE)

Suppose that the log-likelihood function $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ is differentiable and concave and that the MLE is an interior point of Θ. Then, the necessary condition for the MLE is that the Score function is equal to zero:

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

Proof. This immediately follows from the maximisation program for the MLE:

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}; \boldsymbol{y}).$$

□

As the **Score function is a random variable**, we want to compute the expectation and the variance of the this random variable.

Proposition 157 (Expectation of Score Function)
Suppose that the support of $\boldsymbol{y}$ does not depend on $\boldsymbol{\theta}$ and that differentiation and integration can be interchanged. Then, the expectation of the score function is zero:

$$\mathbb{E}_{\boldsymbol{y}}\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0} \tag{70}$$

Proof. First, recall that by definition of a probability density function:

$$\int_{\mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} = 1$$

where $\mathcal{Y}$ is the support of the random variable $\boldsymbol{y}$. Now, differentiate both sides with respect to $\boldsymbol{\theta}$:

$$\int_{\mathcal{Y}} \frac{\partial f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} d\boldsymbol{y} = \mathbf{0}$$

Now, divide and multiply by $f(\boldsymbol{y}; \boldsymbol{\theta})$ on the left hand side:

$$\int_{\mathcal{Y}} \frac{\frac{\partial f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{f(\boldsymbol{y}; \boldsymbol{\theta})} f(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} = \mathbf{0} \tag{71}$$

Now notice that in Eq. (71), the first term in the integral is actually the derivative of the logarithm of the density function by the chain rule:

$$\frac{\frac{\partial f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{f(\boldsymbol{y}; \boldsymbol{\theta})} = \frac{\partial \ln(f(\boldsymbol{y}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} =_{(1)} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

whereby $=_{(1)}$ is the realisation that the derivative of the log density is the definition of the score function! We can then plug this into Eq. (71) to get:

$$\int_{\mathcal{Y}} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} = \mathbf{0}$$

However, this by definition is just the expectation of the score function with respect to the density of the random variable $\boldsymbol{y}$, so we can replace the left hand side with the expectation of the score function and we are done:

$$\mathbb{E}_{\boldsymbol{y}}\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}$$

□

Proposition 158 (Variance of Score Function)

Under the same regularity conditions as above, the **variance of the score vector** is the negative of the expectation of the Hessian of the log-likelihood function:

$$Var\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] \tag{72}$$

Proof. First, we use the fact that the Score function has an expectation of zero:

$$\mathbb{E}_{\boldsymbol{y}}\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \int_{\mathcal{Y}} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} = \mathbf{0}$$

Now, we differentiate with respect to $\boldsymbol{\theta}$ (and assume we can interchange differentiation and integration):

$$\begin{aligned}
&\int_{\mathcal{Y}} \frac{\partial}{\partial \boldsymbol{\theta}^T}\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(\boldsymbol{y}; \boldsymbol{\theta})\right] d\boldsymbol{y} = \mathbf{0} \\
&\qquad \leftrightarrow \quad \int_{\mathcal{Y}} \left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} f(\boldsymbol{y}; \boldsymbol{\theta}) + \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right] d\boldsymbol{y} = \mathbf{0} \\
&\qquad \leftrightarrow \quad \int_{\mathcal{Y}} \left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} + \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\frac{\partial f(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}}{f(\boldsymbol{y}; \boldsymbol{\theta})}\right] f(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} = \mathbf{0} \\
&\qquad \leftrightarrow \quad \int_{\mathcal{Y}} \left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} + \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right] f(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} = \mathbf{0} \\
&\qquad \leftrightarrow \quad \int_{\mathcal{Y}} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} f(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} + \int_{\mathcal{Y}} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} f(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} = \mathbf{0}
\end{aligned}$$

Both integrals are just expectations with respect to $\boldsymbol{y}$:

$$\mathbb{E}_y\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] + \mathbb{E}_y\left[\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right] = \mathbf{0}$$

So then we have that the second moment of the score function is given by the negative expectation of the Hessian of the log-likelihood function:

$$\mathbb{E}_y\left[\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right] = -\mathbb{E}_y\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]$$

Now, we know that the score function has mean zero, this then implies for the variance of the score function:

$$Var\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbb{E}_y\left[\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right] - \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T = \mathbb{E}_y\left[\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right] - \mathbf{0}$$

We can then plug in our expression for the second moment to get our desired result:

$$Var\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = -\mathbb{E}_y\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]$$

□

The variance of the Score function actually has an important name in Statistics.

Definition 159 (Fisher Information Matrix)

The **Fisher Information matrix** is the variance of the Score function:

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] \tag{73}$$

Notice, that the **Fisher information matrix** is the negative of the expectation of the **Hessian** of the log-likelihood function $\ell(\boldsymbol{\theta})$. This is also equivalent to saying that it is the negative of the expectation of the **gradient** of the Score function (which itself is the derivative of the log-likelihood function):

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \frac{\partial}{\partial \boldsymbol{\theta}^T}\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$$

Now, this derivative of the Score function **without taking the expectation** also has an important name in Statistics.

Definition 160 (Observed Fisher Information Matrix)

The **Observed Fisher Information matrix** is the gradient of the Score function:

$$\boldsymbol{I}_o(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{74}$$

Remark 161. *Notice that the Fisher Information matrix and the Observed Fisher Information matrix are very similar objects. The only difference is that the Fisher Information matrix has an expectation operator whilst the Observed Fisher Information matrix does not. In practice, we would compute the Observed Fisher information matrix by computing the Hessian of the log-likelihood function. We would then take the expectation of the Observed Fisher information matrix to retrieve the Fisher Information matrix.*

Example 162

(Geometric Distribution). Suppose we had $\boldsymbol{X} = (X_1, ..., X_n)$ i.i.d random variables whereby $X_i \sim Geo(p)$ with the probability density function:

$$Pr(X_i = x_i) = (1-\pi)^{x_i}\pi$$

for $x_i = 0, 1, ...$

The **log-likelihood function** is given by:

$$\ell(\pi; \boldsymbol{X}) = \sum_{i=1}^{n} \log(Pr(X_i = x_i)) = \sum_{i=1}^{n} \Big[x_i \log(1-\pi) + \log(\pi) \Big] = n\big[\overline{x}\log(1-\pi) + \log(\pi)\big]$$

whereby $\overline{x}$ is the sample mean.

The **Score function** is given by:

$$\frac{d\ell(\pi; \boldsymbol{X})}{d\pi} = n\left(\frac{1}{\pi} - \frac{\overline{x}}{1-\pi}\right)$$

We can solve for the **maximum likelihood estimator** by setting the Score function to zero and solving for π to get:

$$n\left(\frac{1}{\hat{\pi}_{MLE}} - \frac{\overline{x}}{1-\hat{\pi}_{MLE}}\right) = 0$$

$$\Rightarrow \hat{\pi}_{MLE} = \frac{1}{1+\overline{x}}$$

The **Observed Fisher Information matrix** is the negative of the derivative of the Score function:

$$\boldsymbol{I}_o = -\frac{d}{d\pi}\left[\frac{d\ell(\pi; \boldsymbol{X})}{d\pi}\right] = -\frac{d}{d\pi}\left[n\left(\frac{1}{\pi} - \frac{\overline{x}}{1-\pi}\right)\right] = n\left[\frac{1}{\pi^2} + \frac{\overline{x}}{(1-\pi)^2}\right]$$

We now want to compute the **Fisher Information matrix**. To do so, we take the expectation of the Fisher Information matrix (remember that we need only x is the random variable here):

$$\boldsymbol{I} = \mathbb{E}\Big[\boldsymbol{I}_o\Big] = \mathbb{E}\left[n\left[\frac{1}{\pi^2} + \frac{\overline{x}}{(1-\pi)^2}\right]\right] = \frac{n}{\pi^2} + \frac{n\mathbb{E}[\overline{x}]}{(1-\pi)^2}$$

Now, the expectation of a Geometric random variable is $\mathbb{E}[X_i] = (1-\pi)/\pi$. This implies that the expectation of the sample average is:

$$\mathbb{E}[\overline{x}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] = \frac{1}{n}(n\frac{1-\pi}{\pi}) = \frac{1-\pi}{\pi}$$

Therefore, we plug this in to get our final expression for the Fisher information matrix:

$$\boldsymbol{I} = \frac{n}{\pi^2} + \frac{n\mathbb{E}[\overline{x}]}{(1-\pi)^2} = \frac{n}{\pi^2} + \frac{n\frac{1-\pi}{\pi}}{(1-\pi)^2} = \frac{n}{\pi^2(1-\pi)}.$$

We are usually interested in comparing the performance of estimators. In particular, we usually compare the efficiency between estimators. However, for technical reasons, we usually require to restrict our attention to the class of estimators that are **unbiased**. Within this class of estimators, we can actually derive what is the **lowest variance that any unbiased estimator** can achieve. This is known as the **Cramer-Rao lower bound** and it turns out that in fact, it is the inverse of the **Fisher Information matrix**.

Theorem 163 (Cramer–Rao Lower Bound)
Let $\tilde{\boldsymbol{\theta}}(\cdot)$ be an **unbiased estimator** and suppose that the usual regularity conditions hold. Then the covariance matrix of any unbiased estimator is bounded below, in the positive-semidefinite ordering, by the inverse of the Fisher Information matrix:

$$Var[\tilde{\boldsymbol{\theta}}] \succeq \boldsymbol{I}(\boldsymbol{\theta})^{-1}. \tag{75}$$

Therefore, if we have an unbiased estimator $\hat{\theta}$ and it has a variance that is equal to the inverse of the Fisher Information matrix, then this estimator has achieved the lowest possible variance any unbiased estimator can achieve and is hence efficient.

6.3 MLE for Regression

Suppose that we had the data $(y_i, \boldsymbol{x}_i)$ where y_i is the outcome and $\boldsymbol{x}_i$ are the covariates. Then, a key assumption is that we can factor the joint density as:

$$p(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) = q(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}) r(\boldsymbol{x}) \tag{76}$$

whereby the joint density can be written as a product of the conditional density and the marginal density. Notice that q depends on the parameter $\boldsymbol{\theta}$ but the marginal density r does not.

Under the assumption of i.i.d data, this means we can write the **conditional log-likelihood** as:

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(p(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})) = \sum_{i=1}^{n} \log(q(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})) + \sum_{i=1}^{n} \log(r(\boldsymbol{x}_i))$$

However, notice that when we are maximising over $\boldsymbol{\theta}$, the second term does not contain $\boldsymbol{\theta}$ and hence we can drop it. Therefore, the conditional log-likelihood is given by:

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(q(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})) \tag{77}$$

You can then maximise this to find your estimator.

6.3.1 Homoskedastic Gaussian Regression

We show how to derive the MLE for a regression model with Gaussian error terms. Suppose we had the linear regression model:

$$y_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2) \tag{78}$$

which is equivalent to writing:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma u_i$$

where $u_i \sim \mathcal{N}(0, 1)$ and is independent of $\boldsymbol{x}_i$. Here, we want to estimate both $\boldsymbol{\beta}$ and σ. Now, the joint density function $p(\cdot)$ is given by the conditional density $q(\cdot)$ times the marginal density $r(\cdot)$:

$$p(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\beta}, \sigma) = q(\boldsymbol{y}|\boldsymbol{X}; \boldsymbol{\beta}, \sigma) \cdot r(\boldsymbol{x}).$$

Now, when we take logs, we can see that the marginal density $r(x)$ does not depend on $\boldsymbol{\beta}$ or σ so we can ignore

it. Then, we can then write the conditional log-likelihood as:

$$\begin{aligned}\ell_n^c(\boldsymbol{\beta}, \sigma) &= \sum_{i=1}^{n} \log\{q(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}, \sigma)\} \\ &= \sum_{i=1}^{n} \log\left\{\frac{1}{\sigma}\phi\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\right\} \\ &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2,\end{aligned}$$

Now, if we wanted to find the estimator for $\boldsymbol{\beta}$, we can see that only the second term depends on $\boldsymbol{\beta}$. Therefore, we need to find a maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ that **minimises** the sum of squares:

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2.$$

However, this is just the definition of the OLS estimator! This gives us a useful result.

Theorem 164 (OLS and MLE Coincide)
The OLS estimator and the MLE estimator coincide for $\boldsymbol{\beta}$ in the linear regression model with Gaussian error terms.

Now, it can be shown that taking first order conditions, we can derive the MLE $\hat{\sigma}^2$ to be:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)^2 \tag{79}$$

This differs from the unbiased OLS variance estimator, which divides the residual sum of squares by $n - K$.

We now show a similar argument using a multivariate normal distribution.

Definition 165 (Multivariate Normal Distribution)
The $n \times 1$ random vector $\boldsymbol{Z}$ has the multivariate normal distribution $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ if its probability density function is:

$$f_{\boldsymbol{Z}}(\boldsymbol{Z} = \boldsymbol{a}) = (2\pi)^{-\frac{n}{2}}\frac{1}{\sqrt{\det(\Sigma)}}\exp\left(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{a} - \boldsymbol{\mu})\right). \tag{80}$$

Now, suppose we had the classical linear regression model with stochastic regressors and conditional normality on the error terms:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \quad \text{where } \boldsymbol{u}|\boldsymbol{X} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix.

This is equivalent to writing that:

$$\boldsymbol{y}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \tag{81}$$

Matching this to our notation for the multivariate normal, we can see that the mean and variance are:

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta} \quad \text{and} \quad \Sigma = \sigma^2\boldsymbol{I}.$$

Therefore, the conditional density of $\boldsymbol{y}$ given $\boldsymbol{X}$ is:

$$f(\boldsymbol{y}|\boldsymbol{X}) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \tag{82}$$

We can then derive the conditional likelihood function by taking the summation over the sample:

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp\left(\frac{-(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right). \tag{83}$$

We can then take the logarithm to get the **log-likelihood** function:

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{84}$$

Again, notice that only the final term depends on $\boldsymbol{\beta}$. Since the quadratic enters with a negative sign, to find the MLE for $\boldsymbol{\beta}$, we minimise:

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{85}$$

which is equivalent to minimising the sum of squared errors with respect to $\hat{\boldsymbol{\beta}}$. Therefore, again we find that the conditional maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{ML}$ in the classical linear regression model with normally distributed errors is identical to the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. Similarly, we can derive the MLE of the variance to be:

$$\hat{\sigma}^2_{MLE} = \frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{n} \tag{86}$$

6.4 Asymptotic Theory

The MLE has some really nice properties.

Theorem 166 (Consistency of MLE)
Under standard regularity and identification conditions, the MLE is a consistent estimator

$$\hat{\boldsymbol{\theta}}_{MLE} \xrightarrow{P} \boldsymbol{\theta}_0 \tag{87}$$

where $\boldsymbol{\theta}_0$ is the true parameter value.

Proof. First, for each value $\boldsymbol{\theta} \in \Theta$, we have that the average log-likelihood is given by:

$$\bar{\ell}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\log(p(\boldsymbol{x}_i; \boldsymbol{\theta}))$$

Now, notice that as $\boldsymbol{\theta}$ is fixed, this means that $\log(p(\boldsymbol{x}_i; \boldsymbol{\theta}))$ are i.i.d random variables. Therefore, we can apply the weak law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n}\log(p(\boldsymbol{x}_i; \boldsymbol{\theta})) \xrightarrow{P} \mathbb{E}_{\boldsymbol{\theta}_0}[\log(p(\boldsymbol{x}_i; \boldsymbol{\theta}))] = \int_{\mathcal{X}}\log(p(w; \boldsymbol{\theta}))p(w; \boldsymbol{\theta}_0)\, dw$$

where notice that the expectation is taken with respect to the true population value $\boldsymbol{\theta}_0$.

Heuristically, we can say that the average log-likelihood $\bar{\ell}_n(\boldsymbol{\theta}) \to \ell_0(\boldsymbol{\theta})$ which is the true log-likelihood.

Now, we want to argue heuristically:

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}\in\Theta}\bar{\ell}_n(\boldsymbol{\theta}) \xrightarrow{P} \arg\max_{\boldsymbol{\theta}\in\Theta}\ell_0(\boldsymbol{\theta}) = \boldsymbol{\theta}_0 \tag{88}$$

or

$$\hat{\boldsymbol{\theta}}_{MLE} \xrightarrow{P} \boldsymbol{\theta}_0$$

where $\boldsymbol{\theta}_0$ is the true value. We hand wave our argument and say that since $\bar{\ell}_n(\boldsymbol{\theta}) \to \ell_0(\boldsymbol{\theta})$ then as the functions:

$$\bar{\ell}_n(\hat{\boldsymbol{\theta}}_{MLE}) \to \ell_0(\boldsymbol{\theta}_0)$$

then the maximiser:

$$\hat{\boldsymbol{\theta}}_{MLE} \xrightarrow{P} \boldsymbol{\theta}_0.$$

Note that to rigorously make this argument, we need to show uniform convergence, not pointwise convergence.

□

We also have the distribution of the MLE.

Theorem 167 (MLE is Asymptotically Normal)
Under standard regularity and identification conditions, the MLE is **asymptotically normal**:

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{H}^{-1}\boldsymbol{S}\boldsymbol{H}^{-1}) \tag{89}$$

where $\boldsymbol{H} = -\mathbb{E}[\partial^2 \ell_i(\boldsymbol{\theta}_0)/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)]$, $\boldsymbol{S} = \mathbb{E}[\partial\ell_i(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\ \partial\ell_i(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}^T]$ is the variance of the individual score, and $\boldsymbol{\theta}_0$ is the true parameter value.

Fortunately, we can simplify this expression for the asymptotic distribution of the MLE.

Theorem 168 (Efficiency of MLE)
Suppose that the model is correctly specified and the standard regularity conditions hold. Then, the information matrix equality gives $\boldsymbol{H} = \boldsymbol{S}$ and the MLE attains the Cramer–Rao lower bound:

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{H}^{-1}) \tag{90}$$

6.5 Hypothesis Testing

We now look at three different types of tests.

6.5.1 Likelihood Ratio Test

The intuition for the likelihood ratio test is as follows. Suppose we have an unrestricted model where the parameter space is Θ. From this, we can compute the maximum likelihood of the model under this space of parameters:

$$\max_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}).$$

Now, suppose we impose a restriction R which shrinks the parameter space to $\Theta_R \subset \Theta$. Now, as this is a smaller set, the maximum likelihood over this space should be smaller:

$$\max_{\boldsymbol{\theta}\in\Theta_R} \mathcal{L}(\boldsymbol{\theta}).$$

The question is how big of a drop in the likelihood this restriction creates. The bigger the drop, the more likely we are to reject the null hypothesis. We can also show that the resulting test statistic is asymptotically χ^2.

Definition 169 (Likelihood Ratio Test)

Let Θ be the unrestricted parameter space. Let Θ_R be the restricted parameter space under the hypothesis H_R. Then, the likelihood-ratio statistic is defined as:

$$LR = -2\log\left[\frac{\max_{\boldsymbol{\theta}\in\Theta_R}\mathcal{L}(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta}\mathcal{L}(\boldsymbol{\theta})}\right] = -2\Big\{\ell(\hat{\boldsymbol{\theta}}_R) - \ell(\hat{\boldsymbol{\theta}})\Big\} \xrightarrow{D} \chi^2(p) \tag{91}$$

under the null hypothesis, where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_R$ are the unrestricted and restricted maximum likelihood estimators and $p = \dim\Theta - \dim\Theta_R$.

6.5.2 Wald Test

The Wald test asks the question on how big is the distance between the unrestricted estimator $\hat{\boldsymbol{\theta}}$ and the restricted estimator $\hat{\boldsymbol{\theta}}_R$ in the parameter space $\boldsymbol{\theta} \in \Theta = \mathbb{R}^k$ under the null hypothesis:

$$H_0 : \boldsymbol{R}^T\boldsymbol{\theta} = \boldsymbol{r}$$

for some known restriction $\boldsymbol{R} \in \mathbb{R}^{k\times p}$ and $\boldsymbol{r} \in \mathbb{R}^p$.

Definition 170 (Wald Test Statistic)

The Wald test statistic is given by:

$$W = (\boldsymbol{R}^T\hat{\boldsymbol{\theta}} - \boldsymbol{r})^T\Big[\boldsymbol{R}^T\big[\boldsymbol{I}(\hat{\boldsymbol{\theta}})\big]^{-1}\boldsymbol{R}\Big]^{-1}(\boldsymbol{R}^T\hat{\boldsymbol{\theta}} - \boldsymbol{r}) \xrightarrow{D} \chi^2(p) \tag{92}$$

where $\boldsymbol{I}$ is the Fisher information matrix.

6.5.3 Lagrange Multiplier Test

First, recall that the score function is zero when evaluated the maximum likelihood estimator. In this case, it is zero when evaluated at the **unrestricted** maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. The Lagrange Multiplier test looks at the value of the score at the restricted estimator $\hat{\boldsymbol{\theta}}_R$, knowing that this won't be zero and compares it to the score function evaluated at the unrestricted MLE. The test statistic associated is also asymptotically χ^2.

Definition 171 (LM Test)

The Lagrange Multiplier test statistic is given by:

$$LM = \big[\dot{\ell}(\hat{\boldsymbol{\theta}}_R)\big]^T\Big[\boldsymbol{I}(\hat{\boldsymbol{\theta}}_R)\Big]^{-1}\big[\dot{\ell}(\hat{\boldsymbol{\theta}}_R)\big] \xrightarrow{D} \chi^2(p) \tag{93}$$

where $\boldsymbol{I}$ is the Fisher information matrix and $\dot{\ell}(\boldsymbol{\theta})$ is the Score function.

7 The General Regression Model

7.1 Sample and Population Models

This section recaps the things we have covered so far and re-establish notation for the following sections. We are interested in single equation linear models of the form:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + u_i \tag{94}$$

whereby for each observation i, we have a scalar outcome variable y_i, an error term u_i, and a set of K explanatory variables $x_{1i}, x_{2i}, ..., x_{Ki}$. We then have a sample of n observations indexed by i = 1,2,...,n. The linearity of this model refers to the fact that the parameters $\beta_1, ..., \beta_k$ and the error term u_i enter the model additively.

We now denote $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ to be the $K \times 1$ column vectors

$$\boldsymbol{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ . \\ . \\ . \\ x_{ki} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}$$

We can therefore rewrite Eq. (94) as:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i \quad \text{for i = 1,...,n.} \tag{95}$$

In Eq. (95), this refers to the linear model for a **sample**. If we drop the subscripts, we get the **population model**:

$$y = \boldsymbol{x}^T \boldsymbol{\beta} + u \tag{96}$$

Definition 172 (Population Linear Model)

The matrix representation of the **population** version of the linear model is given by:

$$y = \boldsymbol{x}^T \boldsymbol{\beta} + u \tag{97}$$

Definition 173 (Sample Linear Model)

The matrix representation of the **sample** version of the linear model is given by:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i \quad \text{for i = 1,...,n.} \tag{98}$$

Finally, we can stack the observations in Eq. (95) to have the vectors and matrices:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ . \\ . \\ . \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & . & . & . & x_{K1} \\ x_{12} & x_{22} & . & . & . & x_{K2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{1n} & x_{2n} & . & . & . & x_{Kn} \end{bmatrix} \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ . \\ u_n \end{bmatrix}$$

Here, $\boldsymbol{y}, \boldsymbol{u}$ are both $n \times 1$ vectors whilst $\boldsymbol{X}$ is a $n \times K$ matrix whereby each row represents all K observed regressors associated to a particular observation.

Definition 174 (Matrix Representation of Sample Linear Model)
The matrix representation of the sample version of the linear model is given by:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{99}$$

We are interested in estimating the parameter vector $\boldsymbol{\beta}$ from the data on y_i and $x_{1i}, ..., x_{Ki}$ for a sample of $i = 1, ..., n$ observations. The **Ordinary Least Squares** (OLS) estimator finds the parameter vector $\boldsymbol{\beta}$ which minimizes the sum of the squared errors $u_i(\beta) = y_i - x_i^T\beta$:

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} u_i^2(\beta) = \arg\min_{\boldsymbol{\beta}} \boldsymbol{u}(\boldsymbol{\beta})^T \boldsymbol{u}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{100}$$

Definition 175 (Ordinary Least Squares Estimator)
The ordinary least squares estimator is the solution to the problem:

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{101}$$

We can then take first order conditions of the ordinary least squares estimator problem and arrive at the system of K **normal equations** for the K unknown elements of $\boldsymbol{\beta}$:

$$\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \tag{102}$$

If we know that the inverse $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exist, then the parameter vector $\boldsymbol{\beta}$ which minimises the sum of squared residuals is given by:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{103}$$

This can be written in summation form:

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i \right) \tag{104}$$

whereby we recall that $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i$ whereby $\boldsymbol{x}_i$ is a $K \times 1$ vector.

7.2 General Regression Model

We are interested in **predicting** the value taken by y_i using only the information taken by the random variables in $\boldsymbol{x}_i$. We can write our predictions using our regressors in the form: $h(\boldsymbol{x}_i)$ whereby $h(\boldsymbol{x}_i)$ is a function that maps from the K-dimensional vector $\boldsymbol{x}_i$ into the scalars. To measure how good our prediction, we can use the measure of mean squared error:

$$\mathbb{E}\left[(y_i - h(\boldsymbol{x}_i))^2\right].$$

The question is, what function $h(\cdot)$ can minimise the mean squared error of our prediction? The following theorem gives us the answer.

Theorem 176 (Unique Minimiser of Mean Squared Error)
The function $h^*(\boldsymbol{x}_i)$ that minimises the mean squared error:

$$MSE(h) = \mathbb{E}\left[\{y_i - h(\boldsymbol{x}_i)\}^2\right] \tag{105}$$

is **unique** and given by the **conditional expectation function**

$$h^*(\boldsymbol{x}_i) = \mathbb{E}[y_i|\boldsymbol{x}_i]. \tag{106}$$

This is why the conditional expectation function is really important for us. We now generalise our framework by introducing the general regression model. We then see that we can have two submodels arising from this: the linear regression model and the linear projection model.

Definition 177 (General Regression Model)
The general regression model with additive errors can be written as:

$$\boldsymbol{y} = \mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] + \boldsymbol{u} \tag{107}$$

which is the additive decomposition of the random vector $\boldsymbol{y}$ into its conditional expectation given $\boldsymbol{X}$ and a vector of additive deviations $\boldsymbol{u}$ from this conditional expectation.

We get an interesting result regarding the conditional expectation.

Theorem 178 (Orthogonality Property of the Conditional Expectation Function)
The conditional expectation $h^*(\boldsymbol{x}_i) = \mathbb{E}[y_i|\boldsymbol{x}_i]$ has the following orthogonal property that for every other bounded function $g(\boldsymbol{x}_i)$:

$$\mathbb{E}\left[\left\{y_i - h^*(\boldsymbol{x}_i)\right\}g(\boldsymbol{x}_i)\right] = 0 \tag{108}$$

We get the following properties for the conditional expectation which follows from the orthogonality property of the conditional expectation function.

Corollary 179

We have the following properties for the conditional expectation function.

- If x_i is a constant, then $\mathbb{E}[y_i|x_i] = \mathbb{E}[y_i]$;
- If y_i and x_i are independent, then $\mathbb{E}[y_i|x_i] = \mathbb{E}[y_i]$;
- If z_i is independent of (y_i, x_i) then $\mathbb{E}[y_i|x_i, z_i] = \mathbb{E}[y_i|x_i]$;

We can now define our first submodel.

Definition 180 (Linear Regression Model)

A linear regression model is obtained if the conditional expectation of $\boldsymbol{y}$ given $\boldsymbol{X}$ is a linear function of $\boldsymbol{X}$:

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}. \tag{109}$$

This implies that the general regression model:

$$\boldsymbol{y} = \mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] + \boldsymbol{u}$$

becomes

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{110}$$

The OLS estimator is really good in this linear regression model framework as it is able to estimate $\boldsymbol{\beta}$ under certain conditions. However, it is **not** the case that the conditional expectation is a linear function $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}$. In other words, we **do not necessarily** have that the true data generating process is:

$$\boldsymbol{y} = \mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] + \boldsymbol{u} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}.$$

Whilst this may not be the true data generating process for our population data, we may still be interested in the linear projection of the population outcome variable $\boldsymbol{y}$ on the population regressors $\boldsymbol{X}$. The conditional expectation is the best unrestricted predictor under mean squared error, whereas the linear projection is the best predictor within the class of linear functions of $\boldsymbol{X}$.

Definition 181 (Linear Projection Model)

The linear projection model is when we **do not** assume that the conditional expectation is a linear function, but we are still interested on the **linear projection** of $\boldsymbol{y}$ on $\boldsymbol{X}$, which is denoted by

$$\mathbb{E}^*[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}. \tag{111}$$

To summarise, we say that in our population model $\boldsymbol{y}$ and $\boldsymbol{X}$, we want to uncover the true relationship between the two. First, it is important to note that we can **always** project $\boldsymbol{y}$ onto $\boldsymbol{X}$ (this is just linear algebra). If we then assume that it is indeed the case that the relationship between the two variables is indeed that one is being projected onto the other, then we are under the linear regression model framework. This assumption of the outcome $\boldsymbol{y}$ being a projection onto the regressor $\boldsymbol{X}$ comes from the assumption that the conditional expectation is a linear function $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}$. If we do not make the assumption that the true relationship between the two variables is that $\boldsymbol{y}$ is a projection onto $\boldsymbol{X}$, we can still analyse this projection even though it is not the true relationship. This is the linear projection model.

7.3 Classical Linear Regression Models

We are now interested in establishing properties of the OLS estimators. However, in order to do so, we need to make more assumptions. We can first make assumptions that gives us properties that hold under finite samples. Alternatively, we can make assumptions that gives us properties that hold asymptotically.

Assumption 182 (Linear Conditional Expectation)
The conditional expectation of $\boldsymbol{y}$ given $\boldsymbol{X}$ is linear:

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta} \tag{112}$$

$$\mathbb{E}[\boldsymbol{u}|\boldsymbol{X}] = 0 \tag{113}$$

Remark 183. *The assumption*

$$\mathbb{E}[\boldsymbol{u}|\boldsymbol{X}] = 0$$

is also known as **strict exogeneity**.

Assumption 184 (Conditional Homoskedasticity)
The conditional variance of $\boldsymbol{y}$ given $\boldsymbol{X}$ is:

$$Var[\boldsymbol{y}|\boldsymbol{X}] = \sigma^2\boldsymbol{I} \tag{114}$$

$$Var[\boldsymbol{u}|\boldsymbol{X}] = \sigma^2\boldsymbol{I} \tag{115}$$

Assumption 185 (Full Rank)
The matrix $\boldsymbol{X}$ has full rank:

$$\text{rank}(\boldsymbol{X}) = K \tag{116}$$

with probability one.

Remark 186. *The matrix $\boldsymbol{X}$ is a $n \times K$ matrix whereby we have that $K < n$. We know that the rank of a matrix is at most either the minimum of the column or row rank of a matrix.*

Assumption 187 (Independence of Observations)
We assume that observations $(y_i, \boldsymbol{x}_i)$ are independent over $i = 1, 2, ..., n$, so that

$$\mathbb{E}[y_i|\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n] = \mathbb{E}[y_i|\boldsymbol{x}_i] = \boldsymbol{x}_i^T\boldsymbol{\beta} \tag{117}$$

or equivalently

$$\mathbb{E}[u_i|\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n] = \mathbb{E}[u_i|\boldsymbol{x}_i] = 0. \tag{118}$$

Independence also gives us:

Corollary 188

Independence of observations implies that:

$$Cov(y_i, y_j | \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) = 0 \tag{119}$$

or equivalently:

$$Cov(u_i, u_j | \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) = 0 \tag{120}$$

Corollary 189

Independence of observations together with the conditional-mean assumption implies that the error term u_i for observation i is uncorrelated with the explanatory variable $\boldsymbol{x}_j$ for all $i, j = 1, 2, ..., n$.

7.4 Finite Properties in Regression Models

We now take the four assumptions stated in the previous subsection and show finite sample properties of the OLS estimator.

Proposition 190 (Conditional Unbiasedness of OLS Estimator)
The OLS estimator has the conditional expectation

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = \boldsymbol{\beta}. \tag{121}$$

Proof.

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}|\boldsymbol{X}] \\ &=_{(1)} (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] \\ &=_{(2)} (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}\end{aligned}$$

whereby $=_{(1)}$ uses the fact that $\boldsymbol{X}$ is known when conditioned on $\boldsymbol{X}$ and $=_{(2)}$ uses the linear conditional expectation assumption $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}$. □

Theorem 191 (Unconditional Unbiasedness of OLS Estimator)
The OLS estimator is unbiased

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_{OLS}] = \boldsymbol{\beta}. \tag{122}$$

Proof. We can apply the law of iterated expectations

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_{OLS}] =_{(1)} \mathbb{E}[\mathbb{E}[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}]] = \mathbb{E}[\boldsymbol{\beta}] = \boldsymbol{\beta}$$

whereby we used the conditional expectation of the OLS estimator in $=_{(1)}$. □

Theorem 192 (Conditional Variance of OLS Estimator)

$$Var[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{123}$$

Proof.

$$\begin{aligned}Var[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] &= Var[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}|\boldsymbol{X}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Var[\boldsymbol{y}|\boldsymbol{X}]((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T \\ &=_{(1)} (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Var[\boldsymbol{y}|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &=_{(2)} (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned}$$

whereby we used the symmetry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ in $=_{(1)}$ and the assumption of homoskedasticity in $=_{(2)}$. □

Theorem 193 (Unconditional Variance of OLS Estimator)
When $\boldsymbol{X}$ is stochastic, the unconditional variance of the OLS estimator is given by:

$$Var[\hat{\boldsymbol{\beta}}_{OLS}] = \sigma^2 \mathbb{E}\big[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\big]. \tag{124}$$

Theorem 194 (Conditional Gauss-Markov Theorem)
The conditional Gauss-Markov theorem is given by

$$Var[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] \leq Var[\tilde{\boldsymbol{\beta}}|\boldsymbol{X}] \tag{125}$$

for any other unbiased estimator $\tilde{\boldsymbol{\beta}}$ that is a linear function of the random vector $\boldsymbol{y}$ conditional on $\boldsymbol{X}$.

In practice, we don't know what σ^2 is but we would like to estimate it so that we can compute the conditional variance $Var[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. We can estimate it with the following estimator.

Proposition 195 (OLS Estimator of Variance)
The OLS estimator of the variance is given by:

$$\hat{\sigma}^2_{OLS} = \frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{n-K} = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-K} \tag{126}$$

whereby $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS}$ is the vector of OLS residuals and K is the number of regressors.

Proposition 196 (OLS Estimator of Variance is Unbiased)
The OLS estimator for the variance σ^2 is unbiased:

$$\mathbb{E}[\hat{\sigma}^2_{OLS}] = \sigma^2. \tag{127}$$

Proof. First, we note that the fitted residual can be expressed as:

$$\hat{\boldsymbol{u}} = \boldsymbol{M}\boldsymbol{y}$$

where $\boldsymbol{M}$ is the annihilator matrix. Proceeding further, we have that:

$$\begin{aligned}
\hat{\boldsymbol{u}} &= \boldsymbol{M}\boldsymbol{y} \\
&= \boldsymbol{M}[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}] \\
&= \boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{u} \\
&=_{(1)} \boldsymbol{0} + \boldsymbol{M}\boldsymbol{u}
\end{aligned}$$

whereby $=_{(1)}$ follows from the fact that $\boldsymbol{M}\boldsymbol{X} = 0$. Therefore, we have that

$$\hat{\boldsymbol{u}} = \boldsymbol{M}\boldsymbol{u}$$

We can therefore write the residual sum of squares as:

$$\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}} = \boldsymbol{u}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{u} = \boldsymbol{u}^T\boldsymbol{M}\boldsymbol{u},$$

where the last equality uses symmetry and idempotency of $\boldsymbol{M}$. Now, we look at the conditional expectation:

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}|\boldsymbol{X}] &= \mathbb{E}[\boldsymbol{u}^T\boldsymbol{M}\boldsymbol{u}|\boldsymbol{X}] \\
&=_{(1)} \mathbb{E}[\text{tr}(\boldsymbol{M}\boldsymbol{u}\boldsymbol{u}^T)|\boldsymbol{X}] \\
&=_{(2)} \text{tr}\left(\boldsymbol{M}\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^T|\boldsymbol{X}]\right) \\
&= \text{tr}\left(\boldsymbol{M}\sigma^2\boldsymbol{I}\right) \\
&= \sigma^2\,\text{tr}\left(\boldsymbol{M}\right) \\
&=_{(4)} \sigma^2(n-K)
\end{aligned}$$

whereby $=_{(1)}$ uses the trace representation of a quadratic form, $=_{(2)}$ uses the fact that $\boldsymbol{M}$ is known when conditioned on $\boldsymbol{X}$, and $=_{(4)}$ uses the fact that the trace of $\boldsymbol{M}$ is $(n-K)$.

We have therefore shown that:

$$\mathbb{E}[\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}|\boldsymbol{X}] = (n-K)\sigma^2.$$

Therefore, it immediately follows that a conditionally unbiased estimator of σ^2 is:

$$\mathbb{E}\left[\frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{(n-K)}|\boldsymbol{X}\right] = \sigma^2$$

We can then use the law of iterated expectations to show that this is in fact unconditionally unbiased:

$$\mathbb{E}\left[\frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{(n-K)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{(n-K)}|\boldsymbol{X}\right]\right] = \mathbb{E}\left[\sigma^2\right] = \sigma^2.$$

Therefore, the OLS estimator

$$\hat{\sigma}^2_{OLS} = \frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{(n-K)}$$

is an unbiased estimator of the variance σ^2. □

7.5 Interpretation of Coefficients in Regression Models

Suppose we had 2 regressors so the **multiple linear regression model**:

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{u} \tag{128}$$

How can we interpret the coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$? We can interpret it as the result of running two separate regressions which we shall outline below.

First, suppose we had the **simple linear regression**:

$$\boldsymbol{y} = \boldsymbol{X}_1\gamma + r_y. \tag{129}$$

We then have that the OLS estimator $\hat{\gamma}$ would be:

$$\hat{\gamma} = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y}. \tag{130}$$

Now, also suppose we ran the regression:

$$\boldsymbol{X}_2 = \boldsymbol{X}_1\delta + r_2. \tag{131}$$

We would have that the OLS estimator δ would be:

$$\hat{\delta}_{OLS} = (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2. \tag{132}$$

Go back to Eq. (130) and plug in the multiple linear regression model:

$$\begin{aligned}
\hat{\gamma}_{OLS} &= (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{y} \\
&= (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \left[\boldsymbol{X}_1 \hat{\beta}_1 + \boldsymbol{X}_2 \hat{\beta}_2 + \hat{\boldsymbol{u}}\right] \\
&= (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_1 \hat{\beta}_1 + (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2 \hat{\beta}_2 + (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \hat{\boldsymbol{u}} \\
&=_{(1)} \hat{\beta}_1 + \hat{\delta}_{OLS} \hat{\beta}_2 + (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \hat{\boldsymbol{u}} \\
&= \hat{\beta}_1 + \hat{\delta}_{OLS} \hat{\beta}_2 + 0 \\
&= \hat{\beta}_1 + \hat{\delta}_{OLS} \hat{\beta}_2
\end{aligned}$$

where $=_{(1)}$ comes from Eq. (132)

We therefore have that:

$$\hat{\gamma}_{OLS} = \hat{\beta}_1 + \hat{\delta}_{OLS} \hat{\beta}_2 \tag{133}$$

We can see that the OLS estimator in the multiple linear regression does not equal the OLS estimator in the simple linear regression $\hat{\beta}_1 \neq \hat{\gamma}_{OLS}$ unless either:

1. $\hat{\beta}_2 = 0$, so that $\boldsymbol{X}_2$ has no estimated partial relationship with $\boldsymbol{y}$;
2. $\hat{\delta}_{OLS} = 0$, so that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal in the sample;

We can summarise it as this: the estimated coefficient in the multiple regression model considers the relationship between $\boldsymbol{y}$ and $\boldsymbol{X}_1$, controlling for the effect of $\boldsymbol{X}_2$.

7.6 Inference in Regression Models

We are now interested in making statistical statements about our OLS point estimates $\hat{\boldsymbol{\beta}}$. However, in order to do so, we need further assumptions. These assumptions can take one of two forms:

1. Make the assumption that $\boldsymbol{u}|\boldsymbol{X}$ has a normal distribution and this will allow us to derive exact, **finite** sample distribution properties of $\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}$;
2. Make **no** assumption on $\boldsymbol{u}|\boldsymbol{X}$ but use **asymptotics** to derive distribution properties of $\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}$.

In this section, we will look at the properties under finite samples by making the assumption of normality. We will look at asymptotics in the next section.

7.6.1 Classical Linear Regression Model with Normally Distributed Error Terms

We state the assumption of normality on the error term.

Assumption 197 (Classical Linear Regression Model with Normally Distributed Error Terms)
The classical linear regression model with normally distributed error terms can be stated as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{134}$$

$$\boldsymbol{u}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{135}$$

$$\boldsymbol{X} \text{ has full rank with probability one.} \tag{136}$$

Alternatively, we can write:

$$\boldsymbol{y}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \tag{137}$$

$$\boldsymbol{X} \text{ has full rank with probability one.} \tag{138}$$

We now get a useful result from this assumption of normality on the error term.

Theorem 198 (Conditional Normality of the OLS Estimator)
The conditional distribution of the OLS estimator is:

$$\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \tag{139}$$

Proof. We know we can write the OLS estimator as:

$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u}$$

We know that $\boldsymbol{u}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ by assumption.

We know that if $\boldsymbol{t} \sim \mathcal{N}(\mu, \Sigma)$ and $\boldsymbol{z} = A + B\boldsymbol{t}$ for fixed matrices A, B, then $\boldsymbol{z} \sim \mathcal{N}(A + B\mu, B\Sigma B^T)$. Conditional on $\boldsymbol{X}$, the relevant matrices are fixed, so we can apply this to our OLS estimator expression:

$$\begin{aligned}
\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u} &\overset{D}{=} \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \\
&= \mathcal{N}(\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \cdot \boldsymbol{0}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \\
&= \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \\
&= \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})
\end{aligned}$$

□

Looking at

$$\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

we now want to construct test statistics.

Proposition 199 (Marginal Distribution of OLS Estimator)
Suppose that

$$\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

then we have that:

$$\hat{\beta}_k|\boldsymbol{X} \sim \mathcal{N}(\beta_k, v_{kk}) \quad \text{for k } = 1, ..., K \tag{140}$$

where $\hat{\beta}_k$ is the k^{th} element of $\hat{\boldsymbol{\beta}}_{OLS}$ and $v_{kk} = Var[\hat{\beta}_k|\boldsymbol{X}]$ is the k^{th} diagonal element of $Var[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Proof. Immediately follows by the marginal distribution property of multivariate normal distributions! □

We have derived a conditional distribution result regarding the OLS estimator. Can we strengthen this to an unconditional result? No! However, we can if we standardise our OLS estimator. Therefore, we will now show why we standardise our OLS estimators when conducting hypothesis testing.

Theorem 200 (Unconditional Distribution of Standardised OLS Estimator)
The **unconditional distribution** of the standardised OLS estimator is given by:

$$z_k = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}\right) \sim \mathcal{N}(0, 1) \tag{141}$$

Proof. First, note that

$$z_k = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}\right) = \left(\frac{-\beta_k}{\sqrt{v_{kk}}}\right) + \left(\frac{1}{\sqrt{v_{kk}}}\right)\hat{\beta}_k$$

We know that $\hat{\beta}_k|\boldsymbol{X} \sim \mathcal{N}(\beta_k, v_{kk})$ so we then have

$$\begin{aligned}\left(\frac{-\beta_k}{\sqrt{v_{kk}}}\right) + \left(\frac{1}{\sqrt{v_{kk}}}\right)\hat{\beta}_k &\overset{D}{=} \left(\frac{-\beta_k}{\sqrt{v_{kk}}}\right) + \left(\frac{1}{\sqrt{v_{kk}}}\right)\mathcal{N}(\beta_k, v_{kk}) \\ &= \mathcal{N}(\frac{-\beta_k}{\sqrt{v_{kk}}} + \frac{\beta_k}{\sqrt{v_{kk}}}, \frac{v_{kk}}{v_{kk}}) \\ &= \mathcal{N}(0, 1).\end{aligned}$$

We therefore have that the conditional distribution is:

$$z_k|\boldsymbol{X} \sim \mathcal{N}(0, 1) \tag{142}$$

However, notice that the conditional distribution does not contain $\boldsymbol{X}$ at all so therefore, we can actually conclude that the **unconditional distribution** is:

$$z_k \sim \mathcal{N}(0, 1)$$

□

Remark 201. *Note that we can derive an* **unconditional distribution** *result*

$$z_k \sim \mathcal{N}(0, 1)$$

since the conditional distribution did **not** *depend on* $\boldsymbol{X}$*:*

$$z_k|\boldsymbol{X} \sim \mathcal{N}(0, 1).$$

However, we cannot do the same for the conditional distribution of the OLS estimator:

$$\hat{\beta}_k|\boldsymbol{X} \sim \mathcal{N}(\beta_k, v_{kk}).$$

This is because the conditional distribution depends on $\boldsymbol{X}$ via the v_{kk} term.

In practice, we do not know what v_{kk} is. We therefore need to estimate it. We can use the unbiased OLS estimator:

$$\hat{\sigma}^2_{OLS} = \frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{n-K}$$

to estimate σ^2 and then estimate $Var[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}]$ using:

$$\widehat{Var}[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = \hat{\sigma}^2_{OLS}(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

We denote $\hat{v}_{kk}$ to be the k^{th} element on the main diagonal of $\widehat{Var}[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}]$.

Definition 202 (t-Statistic)
The **t-statistic** is given by:

$$t_k = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}}\right) = \left(\frac{\hat{\beta}_k - \beta_k}{se_k}\right) \tag{143}$$

whereby $se_k = \sqrt{\hat{v}_{kk}}$ is called the **standard error** of the estimated parameter $\hat{\beta}_k$.

A big issue is now that we previously knew the distribution

$$z_k = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}\right) \sim \mathcal{N}(0,1)$$

but since we replaced $\sqrt{v_{kk}}$ with se_k, this changes the finite sample distribution of the statistic. We now need to find it.

Theorem 203 (Finite Sample Unconditional Distribution of t-statistic)
The (finite sample) unconditional distribution of the t-test statistic is given by:

$$t_k = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}}\right) = \left(\frac{\hat{\beta}_k - \beta_k}{se_k}\right) \sim t(n-K) \tag{144}$$

where $t(n-K)$ is the t-distribution with $n-K$ degrees of freedom.

Proof. (The proof for this is quite long but very informative). First, we can rewrite

$$\begin{aligned} t_k &= \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}}\right) \\ &= \left(\frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}}\right)\left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}\right) \\ &= \left(\frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}}\right) z_k \end{aligned}$$

where we have already shown that $z_k \sim \mathcal{N}(0,1)$. We are interested in the expression

$$\left(\frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}}\right).$$

First, recall that

$$Var[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

If we let $\boldsymbol{Q} = (\boldsymbol{X}^T\boldsymbol{X})$ and q_{kk} to be the k-th diagonal element of Q^{-1}, we have that

$$\sqrt{v_{kk}} = \sigma\sqrt{q_{kk}} \quad \text{and} \quad \sqrt{\hat{v}_{kk}} = \hat{\sigma}_{OLS}\sqrt{q_{kk}}$$

We therefore have that

$$\left(\frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}}\right) = \frac{\sigma\sqrt{q_{kk}}}{\hat{\sigma}_{OLS}\sqrt{q_{kk}}} = \frac{\sigma}{\hat{\sigma}_{OLS}}. \tag{145}$$

We will keep this in mind later. We now look at the inverse as it is easier to derive the distribution of the inverse:

$$\frac{\hat{\sigma}^2_{OLS}}{\sigma^2} = \left(\frac{1}{\sigma^2}\right)\left(\frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{n-K}\right) = \left(\frac{1}{n-K}\right)\left(\frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{\sigma^2}\right) = \frac{w_0}{n-K} \tag{146}$$

where we defined $w_0 \coloneqq \hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}/\sigma^2$. We make the following claim:

Claim 204.

$$w_0|\boldsymbol{X} \overset{D}{=} \left(\frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{\sigma^2}\right)|\boldsymbol{X} \sim \chi^2(n-K).$$

Proof. By assumption of normality, we had that:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \quad \text{where } \boldsymbol{u}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

Define $\boldsymbol{z}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ so we can rewrite this as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}$$

Now, recall that the fitted residual is the annihilator matrix $\boldsymbol{M}$ applied to the observed variable:

$$\begin{aligned}\hat{\boldsymbol{u}} = \boldsymbol{M}\boldsymbol{y} &= \boldsymbol{M}\left[\boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}\right] \\ &=_{(1)} \boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}\sigma\boldsymbol{z} \\ &= \boldsymbol{M}\sigma\boldsymbol{z}\end{aligned}$$

where $=_{(1)}$ uses the fact that $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$. We therefore have that the residual sum of squares:

$$\begin{aligned}\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}} = (\boldsymbol{M}\sigma\boldsymbol{z})^T\boldsymbol{M}\sigma\boldsymbol{z} &= \sigma\boldsymbol{z}^T\boldsymbol{M}^T\boldsymbol{M}\sigma\boldsymbol{z} \\ &=_{(1)} \sigma\boldsymbol{z}^T\boldsymbol{M}\sigma\boldsymbol{z} \\ &=_{(2)} \sigma^2\boldsymbol{z}^T\boldsymbol{M}\boldsymbol{z}\end{aligned}$$

whereby $=_{(1)}$ uses the fact that $\boldsymbol{M}$ is idempotent and $=_{(2)}$ uses the fact that σ is a scalar. We therefore have that:

$$w_0 = \left(\frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{\sigma^2}\right) = \left(\frac{\sigma^2\boldsymbol{z}^T\boldsymbol{M}\boldsymbol{z}}{\sigma^2}\right) = \boldsymbol{z}^T\boldsymbol{M}\boldsymbol{z}$$

whereby $\boldsymbol{z}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\text{rank}(\boldsymbol{M}) = n - K$ by the properties of annihilator matrices. It is a known property of quadratic forms that since $\boldsymbol{z}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{M}$ is a symmetric and idempotent matrix, we have that the scalar:

$$\boldsymbol{z}^T\boldsymbol{M}\boldsymbol{z}|\boldsymbol{X} \sim \chi^2(\text{rank}(\boldsymbol{M})) = \chi^2(n-K).$$

We have shown therefore shown our intermediate claim that

$$w_0|\boldsymbol{X} \overset{D}{=} \left(\frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{\sigma^2}\right)|\boldsymbol{X} \sim \chi^2(n-K).$$

□

We now go back to our earlier expression for the t-test statistic

$$\begin{aligned} t_k &= \left(\frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}}\right) z_k \\ &=_{(1)} \left(\frac{\sigma}{\hat{\sigma}_{OLS}}\right) z_k \\ &= \frac{z_k}{\hat{\sigma}_{OLS}/\sigma} \\ &= \frac{z_k}{\sqrt{\hat{\sigma}^2_{OLS}/\sigma^2}} \\ &=_{(2)} \frac{z_k}{\sqrt{w_0/(n-K)}} \end{aligned}$$

whereby $=_{(1)}$ uses Eq. (145) and $=_{(2)}$ uses Eq. (146). To recap, we now have the expression for our test statistic as:

$$t_k = \frac{z_k}{\sqrt{w_0/(n-K)}} \tag{147}$$

whereby we know that

$$z_k|\boldsymbol{X} \sim \mathcal{N}(0,1) \quad \text{and} \quad w_0|\boldsymbol{X} \sim \chi^2(n-K).$$

We have a normal distribution divided by a χ^2-distribution. To show that this is a t-distribution, we need to show that the numerator and the denominator are independent. We shall show this in our next claim.

Claim 205. *z_k and w_0 are independent after conditioning on $\boldsymbol{X}$.*

Proof. Let us take a look at our two expressions again:

$$z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}} \quad \text{and} \quad w_0 = \left(\frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{\sigma^2}\right)$$

We see that z_k is a function of $\hat{\beta}_{OLS}$ and w_0 is a function of $\hat{u}$. We will instead show that any function of $\hat{\beta}_{OLS}$ and any function of $\hat{u}$ are independent, which will imply the claim we are making.

First, write the OLS expression:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[\boldsymbol{X}\boldsymbol{\beta} + \sigma z] = \boldsymbol{\beta} + \sigma(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T z$$

where $\boldsymbol{z}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. We can rewrite this as:

$$\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta} = \sigma(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T z.$$

Now, consider the following random vector (these may seem like they come out of nowhere but will prove to be useful):

$$\begin{aligned} d_1 &= \frac{\hat{\boldsymbol{u}}}{\sigma} = \frac{\sigma \boldsymbol{M} z}{\sigma} = \boldsymbol{M} z \\ d_2 &= \frac{\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}}{\sigma} = \frac{\sigma(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T z}{\sigma} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T z \end{aligned}$$

Both $\boldsymbol{M}$ and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ are known when conditioned on $\boldsymbol{X}$. Therefore, we have that:

$$d_1|\boldsymbol{X} \sim \mathcal{N}(\cdot,\cdot) \quad d_2|\boldsymbol{X} \sim \mathcal{N}(\cdot,\cdot)$$

as $\boldsymbol{z}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. Furthermore, we have that $\boldsymbol{d}|\boldsymbol{X}$ is normally distributed whereby $\boldsymbol{d} = (d_1, d_2)^T$. We can then show that d_1 and d_2 are uncorrelated by first showing that:

$$\begin{aligned}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{0}.\end{aligned}$$

It follows that

$$Cov[d_2, d_1|\boldsymbol{X}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}Var[\boldsymbol{z}|\boldsymbol{X}] = 0.$$

This means that d_1, d_2 are uncorrelated conditioned on $\boldsymbol{X}$. This then implies that d_1, d_2 are independent since this is a multivariate normal distribution. We can then rewrite:

$$d_1 = \frac{\hat{\boldsymbol{u}}}{\sigma} \leftrightarrow \hat{\boldsymbol{u}} = \sigma d_1$$

$$d_2 = \frac{\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}}{\sigma} \leftrightarrow \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \sigma d_2$$

whereby $\boldsymbol{\beta}, \sigma$ are non-stochastic. This implies that conditioned on $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{u}}$ are therefore independent. Hence, we can conclude that conditional on $\boldsymbol{X}$, any function of $\hat{\boldsymbol{u}}$ and any function of $\hat{\boldsymbol{\beta}}_{OLS}$ are independent. □

Therefore,

$$t_k = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}}\right) = \frac{z_k}{\sqrt{w_0/(n-K)}}$$

is the ratio of a standard normal random variable and the square root of an independent χ^2 random variable divided by its degrees of freedom. By Proposition 55, we can therefore conclude that the conditional distribution of the t-statistic is:

$$t_k|\boldsymbol{X} \sim t(n-K) \tag{148}$$

As the conditional distribution of t_k does **not** depend on $\boldsymbol{X}$, this means that this holds for all $\boldsymbol{X}$ and we can in fact state that the **unconditional distribution** is:

$$t_k \sim t(n-K). \tag{149}$$

□

Proposition 206 (Asymptotic Normality of t-statistic)
Denote $\nu = n - K$ and let the finite-sample t-test statistic satisfy $t_k \sim t(\nu)$. Then

$$t_k \xrightarrow{D} \mathcal{N}(0,1) \tag{150}$$

as $n \to \infty$.

Proof. Write the t-statistic as

$$t_k = \frac{z}{\sqrt{w/\nu}}$$

whereby $z \sim \mathcal{N}(0,1)$ and $w \sim \chi^2(\nu)$.

Looking at the denominator, we know that a χ^2 random variable can be thought of as:

$$w = \sum_{i=1}^{\nu} z_i^2$$

where $z_i \sim \mathcal{N}(0,1)$ for $i = 1, 2, ..., \nu$ are independent standard normal random variables. First, we note that

$$\mathbb{E}[z_i] = 0$$

and

$$Var[z_i] = \mathbb{E}[z_i^2] - \mathbb{E}[z_i]^2 \quad \leftrightarrow \quad 1 = \mathbb{E}[z_i^2] - 0^2$$
$$\Rightarrow \mathbb{E}[z_i^2] = 1$$

Therefore, computing the first moment of w, we have:

$$\begin{aligned}\mathbb{E}[w] &= \mathbb{E}[\sum_{i=1}^{\nu} z_i^2] \\ &= \nu \mathbb{E}[z_i^2] \\ &=_{(1)} \nu \times 1 \\ &= \nu.\end{aligned}$$

where $=_{(1)}$ uses the fact that $\mathbb{E}[z_i^2] = 1$.

Then, we have that in the denominator:

$$\mathbb{E}\left[\frac{w}{\nu}\right] = \frac{1}{\nu}\mathbb{E}[w] = 1.$$

We have verified that z_i are independent and has finite first and second moments. We can therefore apply the **weak law of large numbers** to the denominator:

$$\frac{w}{\nu} = \frac{1}{\nu} w = \frac{1}{\nu} \sum_{i=1}^{\nu} z_i^2 \xrightarrow{P} \mathbb{E}[z_i^2] = 1.$$

By the **continuous mapping theorem**, we therefore have that:

$$\sqrt{\frac{w}{\nu}} \xrightarrow{P} \sqrt{1} = 1.$$

Going back to our t-statistic t_k, we can therefore apply the **Slutsky's theorem** and get our desired result:

$$t_k = \frac{z}{\sqrt{w/\nu}} \xrightarrow{D} z \sim \mathcal{N}(0,1).$$

This is the desired convergence-in-distribution result.

□

7.6.2 Hypothesis Testing

We have shown that:

$$t_k = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}}\right) = \left(\frac{\hat{\beta}_k - \beta_k}{se_k}\right) \sim t(n-K).$$

Suppose we are interested in the hypothesis test that

$$H_0 : \beta_k = 0.$$

Then, the **t-ratio** is

$$t_k = \left(\frac{\hat{\beta}_k - 0}{se_k}\right) = \left(\frac{\hat{\beta}_k}{se_k}\right) \sim t(n-K)$$

under $H_0 : \beta_k = 0$, which is simply the ratio of the OLS estimate of the coefficient of $\boldsymbol{x}_k$ to its standard error.

We now discuss hypothesis testing when talking about linear combination of parameters. Recall that we have K parameters in $\boldsymbol{\beta}$. Suppose we want to test $p \geq 1$ linear restrictions. We would then need a $p \times K$ matrix $\boldsymbol{H}$ whereby each row of $\boldsymbol{H}$ contains the coefficients of one of the linear restrictions. We can then store these linear restrictions in a $p \times 1$ vector $\boldsymbol{\theta}$. We therefore have that:

$$\boldsymbol{\theta} = \boldsymbol{H}\boldsymbol{\beta} \tag{151}$$

Example 207

Suppose we have a Cobb-Douglas production function of the form:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

where y_i is the log of value added for observation i, x_{2i} is the log of labour input for observation i and x_{3i} is the log of capital input for observations i. We may be interested in testing the constant returns to scale restriction:

$$H_0 : \beta_2 + \beta_3 = 1.$$

This is only one restriction. Therefore, we need a 1×3 matrix $\boldsymbol{H}$.

We would have that the $\boldsymbol{H}$ matrix would then be:

$$\boldsymbol{H} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$$

which will give us

$$\boldsymbol{\theta} = \boldsymbol{H}\boldsymbol{\beta} = \begin{bmatrix} \beta_2 + \beta_3 \end{bmatrix}.$$

We now derive a distribution result for a general test statistic.

Theorem 208 (Distribution of F-statistic)

Consider $p \geq 1$ linear restrictions. We have the $p \times 1$ vector $\boldsymbol{\theta} = H\boldsymbol{\beta}$ for some non-stochastic $p \times K$ matrix H with $rank(H) = p$ (full row rank). Estimate $\boldsymbol{\theta}$ by using the OLS $\hat{\boldsymbol{\theta}}_{OLS} = H\hat{\boldsymbol{\beta}}_{OLS}$. The (scalar) F-test statistic has the form:

$$v = \left(\frac{1}{p\hat{\sigma}^2_{OLS}}\right)\left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}\right)^T D \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}\right) \tag{152}$$

$$= \left(\frac{1}{p}\right)\left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}\right)^T \left[\widehat{Var}[\hat{\boldsymbol{\theta}}_{OLS}|\boldsymbol{X}]\right]^{-1} \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}\right) \sim F(p, n-K) \tag{153}$$

whereby $D^{-1} = H(\boldsymbol{X}^T\boldsymbol{X})^{-1}H^T$.

Proof. (This is another lengthy proof but also very informative.)

We have $\hat{\boldsymbol{\theta}}_{OLS} = H\hat{\boldsymbol{\beta}}_{OLS}$. Under the assumptions of the classical linear regression model, we have that:

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_{OLS}|\boldsymbol{X}] = H\mathbb{E}[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = H\boldsymbol{\beta} = \boldsymbol{\theta}$$

which shows that $\hat{\boldsymbol{\theta}}_{OLS}$ is an unbiased estimator of $\boldsymbol{\theta}$.

For the variance, we have that

$$\begin{aligned} Var[\hat{\boldsymbol{\theta}}_{OLS}|\boldsymbol{X}] &= HVar[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}]H^T \\ &= H[\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}]H^T \\ &= \sigma^2 H(\boldsymbol{X}^T\boldsymbol{X})^{-1}H^T \\ &= \sigma^2 D^{-1} \end{aligned}$$

whereby we define $D^{-}1 = H(\boldsymbol{X}^T\boldsymbol{X})^{-1}H^T$. However, we don't know what σ^2 is so we can estimate it with $\hat{\sigma}^2_{OLS}$ to get:

$$\widehat{Var}[\hat{\boldsymbol{\theta}}_{OLS}|\boldsymbol{X}] = \hat{\sigma}^2_{OLS}D^{-1}. \tag{154}$$

With normally distributed error terms, we can also conclude that:

$$\hat{\boldsymbol{\theta}}_{OLS}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 D^{-1}) \tag{155}$$

We make the first claim.

Claim 209. *The scalar random variable*

$$w|\boldsymbol{X} = (\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})^T[\sigma^2 D^{-1}]^{-1}(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})|\boldsymbol{X} \sim \chi^2(p)$$

Proof. We know our result about quadratic forms of normal random vectors that if we have a random variable $\boldsymbol{X}$ which has a multivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ whereby $\boldsymbol{X}$ has dimension d and the covariance matrix $\boldsymbol{\Sigma}$ has full rank d then, we have that the following quadratic form of $\boldsymbol{X}$ has the distribution:

$$[\boldsymbol{X} - \boldsymbol{\mu}]^T\boldsymbol{\Sigma}^{-1}[\boldsymbol{X} - \boldsymbol{\mu}] \sim \chi^2(d)$$

which is a χ^2 distribution with d degrees of freedom.

We already know that $\hat{\boldsymbol{\theta}}_{OLS}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 D^{-1})$ and hence, we can apply our result immediately to the scalar:

$$w = (\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})^T[\sigma^2 D^{-1}]^{-1}(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})$$

□

We have previously shown that w is a conditional χ^2 random variable:

$$w|\boldsymbol{X} = \left(\frac{1}{\sigma^2}\right)(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})^T D(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})|\boldsymbol{X}$$

We can estimate the variance to get:

$$\hat{w} = \left(\frac{1}{\hat{\sigma}^2_{OLS}}\right)(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})^T D(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})$$

We then redefine to get the variable:

$$v = \frac{\hat{w}}{p} = \left(\frac{1}{p\hat{\sigma}^2_{OLS}}\right)(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})^T D(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})$$

We then write:

$$\begin{aligned} v &= \frac{\hat{w}}{p} \\ &=_{(1)} \left(\frac{\sigma^2}{\hat{\sigma}^2_{OLS}}\right)\left(\frac{w}{p}\right) \\ &= \left(\frac{1}{p\hat{\sigma}^2_{OLS}}\right)(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})^T D(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}) \end{aligned}$$

where $=_{(1)}$ arises from re-expressing w as $\hat{w}$. We have shown before that

$$\frac{\hat{\sigma}^2_{OLS}}{\sigma^2} = \frac{w_0}{n-K}$$

whereby $w_0 = \frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{\sigma^2}$ and that

$$w_0|\boldsymbol{X} \sim \chi^2(n-K).$$

Therefore,

$$\begin{aligned} v &= \left(\frac{\sigma^2}{\hat{\sigma}^2_{OLS}}\right)\left(\frac{w}{p}\right) \\ &= \frac{(w/p)}{(\hat{\sigma}^2_{OLS}/\sigma^2)} \\ &= \frac{(w/p)}{(w_0/(n-K))} \end{aligned}$$

is the ratio of two chi-squared random variables, each divided by their respective degrees of freedom $w|\boldsymbol{X} \sim \chi^2(p)$ and $w_0|\boldsymbol{X} \sim \chi^2(n-K)$.

Moreover, we have that

$$w = \left(\frac{1}{\sigma^2}\right)(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})^T D(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta})$$

is a function of $\hat{\boldsymbol{\beta}}_{OLS}$ whilst

$$w_0 = \frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{\sigma^2}$$

is a function of $\hat{\boldsymbol{u}}$. We have previously shown that conditioned on $\boldsymbol{X}$, any function of $\hat{\boldsymbol{u}}$ and any function of $\hat{\boldsymbol{\beta}}_{OLS}$ are independent.

Therefore,

$$v = \frac{(w/p)}{(w_0/(n-K))}$$

is the ratio of two independent chi-squared random variables, each divided by their respective degrees of freedom.

We can conclude that the conditional distribution is:

$$v|\boldsymbol{X} \sim F(p, n-K) \tag{156}$$

Since this conditional distribution of $v|\boldsymbol{X}$ is the same distribution (F distribution with p and n - K degrees of freedom) for any realisation of $\boldsymbol{X}$, we can therefore conclude that the unconditional distribution is the same:

$$v = \frac{(w/p)}{(w_0/(n-K))} \sim F(p, n-K) \tag{157}$$

□

If we want to test a hypothesis about p linear combinations of the K $\boldsymbol{\beta}$ parameters, we can formulate the null

hypothesis:

$$H_0 : H\boldsymbol{\beta} = \boldsymbol{\theta}^0$$

where H is a non-stochastic $p \times K$ matrix with full rank $rank(H) = p$.

The alternative hypothesis is therefore:

$$H_1 : H\boldsymbol{\beta} \neq \boldsymbol{\theta}^0$$

We use the estimator $\hat{\boldsymbol{\theta}}_{OLS} = H\hat{\boldsymbol{\beta}}_{OLS}$ to estimate our parameter of interest $\boldsymbol{\theta}$. We can then construct the scalar test statistic:

$$v = (\frac{1}{p})(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0)^T \left[\widehat{Var}[\hat{\boldsymbol{\theta}}_{OLS}|\boldsymbol{X}]\right]^{-1} (\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0)$$

whereby we have shown that in the classical linear regression model with normally distributed errors:

$$\widehat{Var}[\hat{\boldsymbol{\theta}}_{OLS}|\boldsymbol{X}] = \widehat{Var}[H\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = H\widehat{Var}[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}]H^T = H\hat{\sigma}^2_{OLS}(\boldsymbol{X}^T\boldsymbol{X})^{-1}H^T.$$

If the null hypothesis is true, we have shown that the test statistic has a F-distribution:

$$v \sim F(p, n-K)$$

We ask whether the value of the test statistic v is an unlikely draw from this null distribution. We reject $H_0 : H\boldsymbol{\beta} = \boldsymbol{\theta}^0$ if the probability of drawing a test statistic with a value as high as v from an F-distribution with (p, n - K) degrees of freedom is below our chosen level of significance, conventionally $\alpha = 5\%$. As the F-distribution has support only over the positive real numbers, we need to do a 1-sided upper tail test and see if the test statistic v lies above the upper 95% percentile of the F(p, n - K) distribution.

Example 210

Suppose that $K = 3$ so we have that:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

If we want to test that $\beta_2 = 0$ and $\beta_3 = 0$, then our null hypothesis would be:

$$H_0 : H\boldsymbol{\beta} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Using our OLS estimates, we have that:

$$\hat{\boldsymbol{\theta}}_{OLS} = H\hat{\boldsymbol{\beta}}_{OLS} = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

The variance is given by:

$$\begin{aligned} \widehat{Var}[\hat{\boldsymbol{\theta}}_{OLS}|\boldsymbol{X}] &= H\widehat{Var}[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}]H^T \\ &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{v}_{11} & \hat{v}_{12} & \hat{v}_{13} \\ \hat{v}_{21} & \hat{v}_{22} & \hat{v}_{23} \\ \hat{v}_{31} & \hat{v}_{32} & \hat{v}_{33} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \hat{v}_{22} & \hat{v}_{23} \\ \hat{v}_{23} & \hat{v}_{33} \end{bmatrix} \end{aligned}$$

We can then construct the test statistic to test $H_0 : \boldsymbol{\theta}^0 = \mathbf{0}$:

$$\begin{aligned} v &= \left(\frac{1}{p}\right)(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0)^T \left[\widehat{Var}[\hat{\boldsymbol{\theta}}_{OLS}|\boldsymbol{X}]\right]^{-1} (\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0) \\ &= \left(\frac{1}{2}\right) \begin{bmatrix} \hat{\beta}_2 - 0 \\ \hat{\beta}_3 - 0 \end{bmatrix}^T \begin{bmatrix} \hat{v}_{22} & \hat{v}_{23} \\ \hat{v}_{23} & \hat{v}_{33} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\beta}_2 - 0 \\ \hat{\beta}_3 - 0 \end{bmatrix} \\ &= \left(\frac{1}{2}\right) \begin{bmatrix} \hat{\beta}_2 & \hat{\beta}_3 \end{bmatrix} \begin{bmatrix} \hat{v}_{22} & \hat{v}_{23} \\ \hat{v}_{23} & \hat{v}_{33} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} \sim F(2, n-K) \end{aligned}$$

We therefore reject $H_0 : \boldsymbol{\theta} = \mathbf{0}$ at the 5% level of significance if the probability of drawing this value of v from the F(2, n - K) distribution is less than 0.05 i.e. if v lies above the upper 95% percentile of the F(2, n - K) distribution. For example, if $n - K = 30$, the upper 95% percentile of the F(2, 30) distribution is 3.32 and values of the test statistic $v > 3.32$ would reject the null hypothesis at the 5% level.

7.6.3 Using Linear Transformations to Simplify Hypothesis Tests

We can reparameterise models so that we can do t-tests instead of F-tests. For example, suppose we had the model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

and we wish to do a two-sided test on the null:

$$H_0 : \beta_2 + \beta_3 = 1.$$

We can subtract x_{3i} from both sides to get:

$$y_i - x_{3i} = \beta_1 + \beta_2 x_{2i} + (\beta_3 - 1)x_{3i} + u_i$$

We can then add and subtract $\beta_2 x_{3i}$ on the right hand side to get:

$$y_i - x_{3i} = \beta_1 + \beta_2(x_{2i} - x_{3i}) + (\beta_2 + \beta_3 - 1)x_{3i} + u_i$$

We can therefore run the regression of $(y_i - x_{3i})$ on the intercept, $(x_{2i} - x_{3i})$ and x_{3i} to get:

$$y_i - x_{3i} = \hat{\beta}_1 + \hat{\beta}_2(x_{2i} - x_{3i}) + \hat{\delta}x_{3i} + u_i$$

whereby $\hat{\delta}$ is an estimator of $(\beta_2 + \beta_3 - 1)$. Therefore, we can see that the simple t-test on:

$$H_0 : \hat{\delta} = 0 \quad \leftrightarrow \quad H_0 : \beta_2 + \beta_3 = 1$$

are equivalent!

Therefore, the general principle of this is to add and subtract terms so that one of your coefficients on your variables is the your original null hypothesis of interest.

7.6.4 Goodness of Fit

Recall that we can always decompose the observed variable y as the fitted value plus the fitted residual.

$$y = \hat{y} + \hat{u}$$

Proposition 211 (Fitted Values are Orthogonal to Fitted Residuals)
The fitted values are orthogonal to the fitted residuals from OLS:

$$\hat{y}^T \hat{u} = 0. \tag{158}$$

Proof.

$$\begin{aligned} \hat{y}^T \hat{u} &= (X\hat{\beta}_{OLS})^T \hat{u} \\ &= \hat{\beta}_{OLS}^T X^T \hat{u} \\ &=_{(1)} \hat{\beta}_{OLS}^T \cdot 0 = 0 \end{aligned}$$

whereby $=_{(1)}$ comes from the first order conditions used to obtain $\hat{\beta}_{OLS}$ whereby $X^T \hat{u} = 0$. □

Theorem 212 (Decomposition of Total Sum of Squares)

The total sum of squares (TSS) can be decomposed as the sum of the explained sum of squares (ESS) and the residual sum of squares (RSS) **provided** that the average of the sample residuals $\overline{\hat{u}}$ are 0:

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^{n}\hat{u}_i^2}_{\text{RSS}} \tag{159}$$

Proof. First, we can write:

$$\begin{aligned} y^T y &= (\hat{y} + \hat{u})^T(\hat{y} + \hat{u}) \\ &= \hat{y}^T\hat{y} + \hat{y}^T\hat{u} + \hat{u}^T\hat{y} + \hat{u}^T\hat{u} \\ &= \hat{y}^T\hat{y} + 0 + 0 + \hat{u}^T\hat{u} \\ &= \hat{y}^T\hat{y} + \hat{u}^T\hat{u} \end{aligned}$$

which is equivalent to:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n}\hat{y}_i^2 + \sum_{i=1}^{n}\hat{u}_i^2$$

We can subtract $n\overline{y}^2$ from both sides:

$$\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = \sum_{i=1}^{n}\hat{y}_i^2 - n\overline{y}^2 + \sum_{i=1}^{n}\hat{u}_i^2$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \underbrace{\sum_{i=1}^{n}\hat{y}_i^2 - n\overline{y}^2}_{(*)} + \sum_{i=1}^{n}\hat{u}_i^2$$

We are now interested in the (*) term. First, since we have that $y_i = \hat{y}_i + \hat{u}_i$, this implies:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\hat{y}_i + \sum_{i=1}^{n}\hat{u}_i$$

and dividing through by n, we get:

$$\overline{y} = \overline{\hat{y}} + \overline{\hat{u}}$$

By assumption, we said that the average of the fitted residuals is 0:

$$\overline{\hat{u}} = 0$$

which means:

$$\overline{y} = \overline{\hat{y}}.$$

We can then plug this into (*) to get:

$$\sum_{i=1}^{n}\hat{y}_i^2 - n\overline{y}^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2.$$

Therefore, we have arrived at our desired expression:

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{\text{TSS}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}_{\text{ESS}}+\underbrace{\sum_{i=1}^{n}\hat{u}_i^2}_{\text{RSS}}$$

□

Remark 213. *We made the seemingly random assumption that the average of the sample residuals is 0:* $\overline{\hat{u}}=0$. *This in fact always hold if* **the model includes an intercept term**.

We can take our expression:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}\hat{u}_i^2$$

and divide by $\sum_{i=1}^{n}(y_i-\overline{y})^2$ on both sides to get:

$$1=\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}+\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

We can then re-arrange this to get:

$$\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=1-\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

Definition 214 (R-squared)

R-squared is given by:

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=1-\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \tag{160}$$

or equivalently:

$$R^2=\frac{\text{ESS}}{\text{TSS}}=1-\frac{\text{RSS}}{\text{TSS}} \tag{161}$$

whereby $0\leq R^2\leq 1$ provided that $\overline{\hat{u}}=0$.

R^2 measures the proportion of the total variation in the dependent variable that is accounted for by the model. In fact, R^2 measures the goodness of fit of a model with additional explanatory variables **relative** to the fit of a model that includes only an intercept term. Hence, we should not use this centred R^2 as a measure of fit for models that do not include an intercept term. With an intercept, R^2 is invariant to affine rescaling of the dependent variable and to reparameterisations of the regressors that preserve the column space of X. However, R^2 should **not** be used to compare the fit of models that have **different dependent variables**, such as levels and logarithms.

7.6.5 Using R^2 to test exclusion restrictions

An alternative way to test p exclusion restrictions is to compute the test statistic:

$$v=\left(\frac{n-K}{p}\right)\left(\frac{R_U^2-R_R^2}{1-R_U^2}\right)\sim F(p,n-K) \tag{162}$$

whereby R_U^2 is the R^2 from the unrestricted model and R_R^2 is the R^2 from the restricted model. Here, the null hypothesis is rejected if the exclusion of these p explanatory variables results in a sufficiently large fall in the R^2 goodness of fit measure, or in a sufficiently large deterioration in the fit of the model.

This test can be used to test whether do **all** $K-1$ slope coefficients in the model are equal to 0. Here, the restricted model $y_i = \beta_1 + u_i$ has $R^2_R = 0$. Therefore, the test statistic simplifies to:

$$v = \left(\frac{n-K}{K-1}\right)\left(\frac{R^2}{1-R^2}\right) \sim F(K-1, n-K) \tag{163}$$

where $R^2 = R^2_U$ denotes the R^2 in the unrestricted model. This is the **F-test** for the model.

7.7 Asymptotic Properties in Regression Models

So far we have looked at regression in a finite sample setting. Under such a setting, we had to impose strong assumptions of normality on error terms. We can in fact relax this assumption of normality by instead considering asymptotics. We now want to look at properties of the OLS estimators that hold in large samples. That is, fixing the number of explanatory variables K fixed, we let the sample size n go to infinity. This approach of analysing properties actually hold under **weaker assumptions** than under finite sample results.

7.7.1 Consistency of OLS Estimator

We are interested in showing that the OLS estimator is consistent. We state the assumptions required for this.

Assumption 215 (Assumptions for Consistency)

The assumptions needed for consistency of the OLS estimator are:

1. $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$;
2. The data on $(y_i, \boldsymbol{x}_i^T)$ are independent and identically distributed for $i = 1, ..., n$;
3. $\mathbb{E}[\boldsymbol{x}_i u_i] = 0$ for all $i = 1, ..., n$;
4. The $K \times K$ matrix $M_{XX} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T]$ exists and is non-singular.

Theorem 216 (Consistency of OLS Estimator)

Under the assumptions of consistency, the OLS estimator is **consistent**

$$p \lim_{n\to\infty} \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} \tag{164}$$

or equivalently:

$$\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{P} \boldsymbol{\beta}. \tag{165}$$

Proof. First, write the OLS estimator as:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{OLS} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u} \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u}
\end{aligned}$$

We then multiply and divide by n:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{OLS} &= \boldsymbol{\beta} + n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\frac{1}{n}\right)\boldsymbol{X}^T\boldsymbol{u} \\
&= \boldsymbol{\beta} + \underbrace{\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}}_{(*)}\underbrace{\left(\frac{\boldsymbol{X}^T\boldsymbol{u}}{n}\right)}_{(**)}
\end{aligned}$$

We first look at (*). By assumption (2), the observations $\boldsymbol{x}_i$ are i.i.d. Furthermore, by assumption (4), the second moment $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T]$ exists. Therefore, by the **weak law of large numbers**, we have that:

$$\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \xrightarrow{P} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T] =_{(1)} M_{XX}$$

whereby $=_{(1)}$ comes from the definition of M_{XX} by definition. We can then apply the **continuous mapping theorem** to conclude that:

$$p \lim_{n \to \infty} \left(\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \right)^{-1} = (M_{XX})^{-1} =_{(1)} M_{XX}^{-1}$$

whereby in $=_{(1)}$, by assumption (4), we assumed that M_{XX} is non-singular so its inverse exists.

We now look at (**). By assumption (2), we know that $(\boldsymbol{x}_i, u_i)$ is i.i.d. Furthermore, $\mathbb{E}[\boldsymbol{x}_i u_i] = \boldsymbol{0}$, so the first moment exists. We can therefore apply the **weak law of large numbers** to get:

$$\frac{\boldsymbol{X}^T \boldsymbol{u}}{n} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i u_i \xrightarrow{P} \mathbb{E}[\boldsymbol{x}_i u_i] =_{(1)} \boldsymbol{0}$$

whereby $=_{(1)}$ arises from the orthogonal assumption (3) of $\boldsymbol{X}$ and $\boldsymbol{u}$.

Therefore, we can apply the **continuous mapping theorem** to the product of (*) and (**) that we started with, whereby we now have that the **product of the limits is the limit of the products**:

$$p \lim_{n \to \infty} \left[\left(\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \right)^{-1} \left(\frac{\boldsymbol{X}^T \boldsymbol{u}}{n} \right) \right] = \left[p \lim_{n \to \infty} \left(\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \right)^{-1} \right] \left[p \lim_{n \to \infty} \left(\frac{\boldsymbol{X}^T \boldsymbol{u}}{n} \right) \right] = \left[M_{XX}^{-1} \right] \left[\boldsymbol{0} \right] = \boldsymbol{0}$$

We can therefore conclude that:

$$p \lim_{n \to \infty} \hat{\boldsymbol{\beta}}_{OLS} = p \lim_{n \to \infty} \boldsymbol{\beta} + p \lim_{n \to \infty} \left[\left(\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \right)^{-1} \left(\frac{\boldsymbol{X}^T \boldsymbol{u}}{n} \right) \right] = \boldsymbol{\beta} + \boldsymbol{0}$$

Therefore, we have shown consistency:

$$p \lim_{n \to \infty} \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta}.$$

□

In our proof, we assumed that the data $(y_i, \boldsymbol{x}_i^T)$ were independent and identically distributed. This is stronger than necessary: consistency can also be established for independent but non-identically distributed observations when suitable laws of large numbers apply to $n^{-1} \sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T$ and $n^{-1} \sum_i \boldsymbol{x}_i u_i$.

Theorem 217 (OLS Consistency under Weaker Assumptions)
Suppose that the data $(y_i, \boldsymbol{x}_i^T)$ are independent over $i = 1, 2, ..., n$ but **not** identically distributed. If $n^{-1} \sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T$ converges in probability to a finite non-singular matrix and $n^{-1} \sum_i \boldsymbol{x}_i u_i \xrightarrow{P} \boldsymbol{0}$ under suitable moment conditions, then the OLS estimator is still **consistent**

$$p \lim_{n \to \infty} \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} \tag{166}$$

or equivalently:

$$\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{P} \boldsymbol{\beta}. \tag{167}$$

Proof. See section 4.4 of Cameron and Trivedi (2005). □

7.7.2 Linear Projection Model Revisited

So far, the key assumption we required was the **orthogonality assumption**

$$\mathbb{E}[\boldsymbol{x}_i u_i] = 0.$$

In practice, we generally consider linear models whereby:

1. We have an intercept term so that $x_{1i} = 1$ for all $i = 1, 2, ..., n$. This then implies that the first element of the $K \times 1$ vector $\mathbb{E}[x_{1i} u_i] = 0$ implies that

$$\mathbb{E}[x_{1i} u_i] = \mathbb{E}[(1) u_i] = \mathbb{E}[u_i] = 0.$$

2. Alternatively, we can assume that $\mathbb{E}[y_i] = 0$ and $\mathbb{E}[\boldsymbol{x}_i] = \boldsymbol{0}$. This then implies that:

$$\mathbb{E}[u_i] = 0.$$

Since in both cases, we have that $\mathbb{E}[u_i] = 0$ and the orthogonality assumption $\mathbb{E}[\boldsymbol{x}_i u_i] = 0$, this then implies that:

$$Cov[\boldsymbol{x}_i, u_i] = \mathbb{E}[\boldsymbol{x}_i u_i] - \mathbb{E}[\boldsymbol{x}_i]\mathbb{E}[u_i] = 0$$

and hence:

$$Corr[\boldsymbol{x}_i, u_i] = 0.$$

This allows us to define the **linear projection model** based off this orthogonality assumption.

Definition 218 (Linear Projection Model)
We make the assumptions that:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i$$

and

$$\mathbb{E}[\boldsymbol{x}_i u_i] = \boldsymbol{0}.$$

The **linear projection** model is then defined as:

$$\mathbb{E}^*[y_i | \boldsymbol{x}_i] = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

Remark 219. *The linear projection model examines the linear relationship between y_i and the explanatory variable $\boldsymbol{x}_i$ with the property that the additive error term u_i is* **uncorrelated in the population equation** *with each of the included explanatory variables in $\boldsymbol{x}_i$.*

In our traditional linear regression model setting, we made the assumption of **linear conditional expectation**:

$$\mathbb{E}[y_i | \boldsymbol{x}_i] = \boldsymbol{x}_i^T \boldsymbol{\beta} \quad \leftrightarrow \quad \mathbb{E}[u_i | \boldsymbol{x}_i] = 0.$$

Definition 220 (Linear Regression Model)

We make the assumptions that:

$$\mathbb{E}[y_i|\boldsymbol{x}_i] = \boldsymbol{x}_i^T\boldsymbol{\beta} \quad \leftrightarrow \quad \mathbb{E}[u_i|\boldsymbol{x}_i] = 0.$$

The **linear regression** model is then defined as:

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i$$

This in fact implies the orthogonality assumption in the linear projection model.

Proposition 221 (Linear Conditional Expectation Implies Orthogonality)

The linear conditional expectation assumption implies the orthogonality assumption:

$$\mathbb{E}[y_i|\boldsymbol{x}_i] = \boldsymbol{x}_i^T\boldsymbol{\beta} \quad \Rightarrow \quad \mathbb{E}[\boldsymbol{x}_i u_i] = 0$$

Proof. Apply the law of iterated expectations:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_i u_i] &= \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_{u|\boldsymbol{x}}[\boldsymbol{x}_i u_i|\boldsymbol{x}_i]] \\
&= \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}_i \mathbb{E}_{u|\boldsymbol{x}}[u_i|\boldsymbol{x}_i]] \\
&=_{(1)} \mathbb{E}_{\boldsymbol{x}}[0] \\
&= 0
\end{aligned}$$

whereby $=_{(1)}$ comes from the linear conditional expectation assumption $\mathbb{E}[u_i|\boldsymbol{x}_i] = 0$. □

Theorem 222 (Linear Regression Model Implies Linear Projection Model)

The linear regression model is a *model subset* of the linear projection model.

It is not the case that the converse holds i.e. $\mathbb{E}[\boldsymbol{x}_i u_i] = 0$ **does not** imply that $\mathbb{E}[u_i|\boldsymbol{x}_i] = 0$. It is a weaker assumption as we are not specifying that the conditional expectation $\mathbb{E}[y_i|\boldsymbol{x}_i]$ be a linear function.

Furthermore, the orthogonality function for the linear regression model, whereby for any bounded function $g(x_i)$, we have that:

$$\mathbb{E}\Big[\{y_i - \mathbb{E}[y_i|\boldsymbol{x}_i]\}g(\boldsymbol{x}_i)\Big] = \mathbb{E}\Big[\{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\}g(\boldsymbol{x}_i)\Big] = 0$$

implies the orthogonality condition in the linear projection model (whereby $g(\boldsymbol{x}_i) = \boldsymbol{x}_i$):

$$\mathbb{E}\Big[\{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\}\boldsymbol{x}_i\Big] = \mathbb{E}\Big[u_i\boldsymbol{x}_i\Big] = 0.$$

However, the converse **does not hold**.

7.7.3 Asymptotic Normality

We now want to show the limiting distribution of the OLS estimator where we do not have the assumption of normality of the error terms. Doing so will allow us to construct confidence intervals and conduct asymptotically valid hypothesis testing.

Assumption 223 (Assumptions for Asymptotic Normality)

The assumptions needed for asymptotic normality of the OLS estimator are:

1. $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$;
2. The data on $(y_i, \boldsymbol{x}_i^T)$ are independent and identically distributed for $i = 1, ..., n$;
3. $\mathbb{E}[\boldsymbol{x}_i u_i] = 0$ for all $i = 1, ..., n$;
4. The $K \times K$ matrix $M_{XX} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T]$ exists and is non-singular;
5. The $K \times K$ matrix $M_{X\Omega X} = \mathbb{E}[u_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T]$ exists and is non-singular;

We now state the most important result of this section.

Theorem 224 (Asymptotic Normality of OLS Estimator)

Under the assumptions for asymptotic normality, we have that:

$$\sqrt{n}[\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}] \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}). \tag{168}$$

Proof. First, recall that the expression for the OLS estimator is:

$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}^T\boldsymbol{u}}{n}\right)$$

or alternatively:

$$\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta} = \left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}^T\boldsymbol{u}}{n}\right)$$

We then rescale it by $\sqrt{n}$ to get:

$$\sqrt{n}[\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}] = \underbrace{\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}}_{(*)}\underbrace{\left(\frac{\boldsymbol{X}^T\boldsymbol{u}}{\sqrt{n}}\right)}_{(**)}$$

We look at the limiting result for (*) and (**). First, for (*), we can apply the **weak law of large numbers**:

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \xrightarrow{P} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T] =_{(1)} M_{XX}$$

whereby M_{XX} holds by definition of M_{XX} in assumption (4). Furthermore, we know that by assumption (4), the inverse of M_{XX} exists, so by the **continuous mapping theorem**, we have that

$$\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1} \xrightarrow{P} \left(M_{XX}\right)^{-1} = M_{XX}^{-1}.$$

Now looking at (**), we can rewrite (**) as:

$$\frac{\boldsymbol{X}^T\boldsymbol{u}}{\sqrt{n}} = \sqrt{n}\left(\frac{\boldsymbol{X}^T\boldsymbol{u}}{n}\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i u_i\right)$$

where we now have a $K \times 1$ vector of sample means: $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i u_i$. This gives us a hint to apply the **central limit theorem** to this vector of sample means. First, we compute the mean and variance of this expression:

$$\mathbb{E}[\boldsymbol{x}_i u_i] =_{(1)} \boldsymbol{0}$$

whereby $=_{(1)}$ comes from assumption (3) of orthogonality. For the variance, we have:

$$\begin{aligned} Var[\boldsymbol{x}_i u_i] &= \mathbb{E}[(\boldsymbol{x}_i u_i)(\boldsymbol{x}_i u_i)^T] - \mathbb{E}[\boldsymbol{x}_i u_i]\mathbb{E}[\boldsymbol{x}_i u_i]^T \\ &= \mathbb{E}[u_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T] - \mathbf{0} \\ &=_{(1)} M_{X\Omega X} \end{aligned}$$

whereby $=_{(1)}$ comes from the definition of $M_{X\Omega X}$ in assumption (5). Finally, the $K \times 1$ random vectors $(\boldsymbol{x}_i u_i)$ are independent and identically distributed over $i = 1, 2, ..., n$ by assumption (2). We can therefore apply the central limit theorem to (**):

$$\frac{\boldsymbol{X}^T \boldsymbol{u}}{\sqrt{n}} = \sqrt{n}\left(\frac{\boldsymbol{X}^T \boldsymbol{u}}{n}\right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, M_{X\Omega X}).$$

We now go back to our main expression:

$$\sqrt{n}[\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}] = \underbrace{\left(\frac{\boldsymbol{X}^T \boldsymbol{X}}{n}\right)^{-1}}_{(*)} \underbrace{\left(\frac{\boldsymbol{X}^T \boldsymbol{u}}{\sqrt{n}}\right)}_{(**)}$$

We have shown that $(*) \xrightarrow{P} M_{XX}^{-1}$ and $(**) \xrightarrow{D} \mathcal{N}(\mathbf{0}, M_{X\Omega X})$. We apply **Slutsky's theorem** to conclude that:

$$\begin{aligned} \left(\frac{\boldsymbol{X}^T \boldsymbol{X}}{n}\right)^{-1} \left(\frac{\boldsymbol{X}^T \boldsymbol{u}}{\sqrt{n}}\right) \xrightarrow{D} M_{XX}^{-1} \cdot \mathcal{N}(\mathbf{0}, M_{X\Omega X}) &= \mathcal{N}(M_{XX}^{-1} \cdot \mathbf{0}, M_{XX}^{-1} M_{X\Omega X} \left(M_{XX}^{-1}\right)^T) \\ &=_{(1)} \mathcal{N}(\mathbf{0}, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}) \end{aligned}$$

whereby $=_{(1)}$ uses the fact that M_{XX}^{-1} is **symmetric**. Therefore, we can conclude our desired result:

$$\sqrt{n}[\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}] \xrightarrow{D} \mathcal{N}(\mathbf{0}, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}).$$

□

Remark 225. *We could have relaxed the assumption that $(y_i, \boldsymbol{x}_i^T)$ is i.i.d. to independence under suitable non-identically distributed central limit theorem conditions.*

This brings us to a broader and more general definition.

Definition 226 (Asymptotically Normal)
An estimator $\hat{\boldsymbol{\theta}}$ of the parameter vector $\boldsymbol{\theta}$ with the property that:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, V) \tag{169}$$

for some finite variance matrix V is said to be **asymptotically normal**.

We are saying that if appropriately scaled, the estimator has a non-degenerate limit distribution which is normal. Using this terminology, we say that the OLS estimator is **asymptotically normal**.

We now wish to explore further properties of asymptotic normality. We now look at a *special case* of the classical linear regression model with linear conditional expectation and conditional homoskedasticity.

Assumption 227 (Assumptions for Asymptotic Normality)

The assumptions needed for asymptotic normality of the OLS estimator are:

1. $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$;
2. The data on $(y_i, \boldsymbol{x}_i^T)$ are independent and identically distributed for $i = 1, ..., n$;
3. Linear conditional expectation $\mathbb{E}[u_i|\boldsymbol{x}_i] = 0$ for all $i = 1, ..., n$;
4. The $K \times K$ matrix $M_{XX} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T]$ exists and is non-singular;
5. The $K \times K$ matrix $M_{X\Omega X} = \mathbb{E}[u_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T]$ exists and is non-singular;
6. Conditional homoskedasticity $Var[u_i|\boldsymbol{x}_i] = \mathbb{E}[u_i^2|\boldsymbol{x}_i] = \sigma^2$.

We now have a different limiting variance matrix for our OLS estimator.

Theorem 228 (Asymptotic Normality of OLS Estimator under Homoskedasticity)

Under the strengthened assumptions for asymptotic normality with conditional homoskedasticity, we have that:

$$\sqrt{n}[\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}] \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \sigma^2 M_{XX}^{-1}). \tag{170}$$

Proof. Under assumption (6), we have that:

$$\begin{aligned}
M_{X\Omega X} &= \mathbb{E}[u_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T] \\
&=_{(1)} \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{u|\boldsymbol{X}}[u_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T | \boldsymbol{x}_i]\right] \\
&= \mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T \mathbb{E}_{u|\boldsymbol{X}}[u_i^2|\boldsymbol{x}_i]\right] \\
&=_{(2)} \mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T \sigma^2\right] \\
&= \sigma^2 \mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T\right] \\
&= \sigma^2 M_{XX}
\end{aligned}$$

where $=_{(1)}$ is the application of the law of iterated expectations and $=_{(2)}$ is from assumption (6) of conditional homoskedasticity.

We can therefore apply this to our original limiting distribution result:

$$\sqrt{n}[\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}] \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}) = \mathcal{N}(\boldsymbol{0}, M_{XX}^{-1}(\sigma^2 M_{XX}) M_{XX}^{-1}) = \mathcal{N}(\boldsymbol{0}, \sigma^2 M_{XX}^{-1})$$

□

We therefore have that nice result that:

$$\sqrt{n}[\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}] \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, V)$$

where the limiting variance $V = \sigma^2 M_{XX}^{-1}$. We can rewrite this expression for the OLS estimator to be asymptotically normal:

$$\hat{\boldsymbol{\beta}}_{OLS} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \frac{V}{n})$$

where $\overset{a}{\sim}$ denotes the asymptotic distribution. This is called the **asymptotic distribution** of $\hat{\boldsymbol{\beta}}_{OLS}$. However, this is not practical to use. To see why, look at the asymptotic variance:

$$V = \sigma^2 M_{XX}^{-1}.$$

We have shown that:

$$\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1} \xrightarrow{P} M_{XX}^{-1}$$

so that means we have a **consistent estimator** of the $K \times K$ matrix M_{XX}^{-1}. However, we do not know what σ^2 is. Therefore, we need an estimator for σ^2.

We can use either the OLS estimator:

$$\hat{\sigma}^2_{OLS} = \frac{1}{n-K}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{n-K} \tag{171}$$

or the MLE estimator:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}}{n} \tag{172}$$

to estimate the unknown variance σ^2. Both of these estimators are consistent. Therefore, we can apply **Slutsky's theorem** on the estimator:

$$\hat{V} = \hat{\sigma}^2_{OLS}\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1} \xrightarrow{P} V$$

which means we have a consistent estimator:

$$\hat{V} \xrightarrow{P} V$$

We are able to replace V with $\hat{V}$ in our asymptotic distribution without changing any of our asymptotic distribution results. We therefore have the approximation of the limiting distribution:

$$\hat{\boldsymbol{\beta}}_{OLS} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \frac{\hat{V}}{n})$$

which we can use to construct approximate confidence intervals and asymptotically valid test statistics. Note the limiting variance and the asymptotic distribution for $\hat{\boldsymbol{\beta}}_{OLS}$ coincides with the **exact finite** sample conditional distribution of $\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}$ we made when we assumed that the error terms were normally distributed. Hence, we know that in this scenario, our asymptotic approximation will be accurate.

7.7.4 Asymptotic Inference

We state the test statistic we can use for **simple hypothesis testing**.

Proposition 229 (Asymptotic T-test Statistic For Simple Hypothesis Testing)
We can use the following test statistic for **simple hypothesis testing** of the form $H_0 : \beta_k = \beta_k^0$:

$$t_k = \frac{\hat{\beta}_k - \beta_k^0}{se_k} \overset{a}{\sim} \mathcal{N}(0,1) \quad \text{under } H_0. \tag{173}$$

where se_k is the asymptotic standard error.

Proof. Recall our asymptotic distribution:

$$\hat{\boldsymbol{\beta}}_{OLS} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \frac{\hat{V}}{n})$$

Let $\hat{v}_{kk}$ denote the k^{th} element of the main diagonal of the $K \times K$ variance matrix estimator:

$$\frac{\hat{V}}{n} = \hat{\sigma}^2_{OLS}(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

We have that for each OLS estimator β_k:

$$\hat{\beta}_k \overset{a}{\sim} \mathcal{N}(\beta_k, \hat{v}_{kk})$$

We can standardise this to get:

$$t_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}} = \frac{\hat{\beta}_k - \beta_k}{se_k} \overset{a}{\sim} \mathcal{N}(0,1).$$

□

We can then construct **simple hypothesis tests** of the form $H_0 : \beta_k = \beta_k^0$ by comparing the test statistic

$$t_k = \frac{\hat{\beta}_k - \beta_k^0}{se_k}$$

computed under the null hypothesis and comparing it with the standard normal distribution $\mathcal{N}(0,1)$ at the desired significance level. We can also justify the asymptotic test statistic we just derived by looking at the limiting distribution of a finite sample t-test statistic. It may seem strange that we were able to treat t_k as being drawn from the standard normal distribution even though we estimated the unknown σ^2. We can in fact show that this is entirely justified.

Theorem 230 (Asymptotic Validity of T-test Statistic)

Define the z-statistic as:

$$z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}} \xrightarrow{D} \mathcal{N}(0,1) \tag{174}$$

We can also show that the t-statistic with an estimated v_{kk} is asymptotically equivalent:

$$t_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}} \xrightarrow{D} \mathcal{N}(0,1) \tag{175}$$

Proof. Under the asymptotic normality result for OLS:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2 M_{XX}^{-1}).$$

This automatically implies that the standardised statistic:

$$z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}} \xrightarrow{D} \mathcal{N}(0,1).$$

We now look at the t-statistic and rewrite it as:

$$t_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}} = \frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}}\left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}\right) = \frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}} z_k$$

We have already shown that $z_k \xrightarrow{D} \mathcal{N}(0,1)$. We look at the term $\frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}}$, which contains the true variance in the numerator and the estimated variance in the denominator. Since $\hat{\sigma}^2/\sigma^2 \xrightarrow{P} 1$, the corresponding variance estimator is ratio-consistent:

$$\frac{\hat{v}_{kk}}{v_{kk}} \xrightarrow{P} 1.$$

We can then apply the **continuous mapping theorem**:

$$\frac{\sqrt{\hat{V}_{kk}}}{\sqrt{V_{kk}}} \xrightarrow{P} 1.$$

We then apply **Slutsky's theorem**:

$$\frac{\sqrt{V_{kk}}}{\sqrt{\hat{V}_{kk}}} \xrightarrow{P} 1.$$

We conclude and show our result by again applying Slutsky's theorem:

$$t_k = \frac{\sqrt{V_{kk}}}{\sqrt{\hat{V}_{kk}}} z_k \xrightarrow{D} (1) \cdot \mathcal{N}(0,1) = \mathcal{N}(0,1).$$

□

We now want to test linear restrictions $\boldsymbol{\theta} = H\boldsymbol{\beta}$ whereby H is a non-stochastic $p \times K$ matrix with $rank(H) = p$ (full rank). Go back to our asymptotic distribution for the OLS estimator:

$$\hat{\boldsymbol{\beta}}_{OLS} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \frac{\hat{V}}{n})$$

The asymptotic distribution of the $p \times 1$ random vector $\hat{\boldsymbol{\theta}}_{OLS} = H\hat{\boldsymbol{\beta}}_{OLS}$ is:

$$\hat{\boldsymbol{\theta}}_{OLS} \overset{a}{\sim} \mathcal{N}(H\boldsymbol{\beta}, H\frac{\hat{V}}{n}H^T)$$

Proposition 231 (Consistency of Linear Restrictions)
The estimator of $\boldsymbol{\theta}$ constructed from the OLS estimator $\hat{\boldsymbol{\theta}}_{OLS} = H\hat{\boldsymbol{\beta}}_{OLS}$ is consistent:

$$p\lim_{n\to\infty} \hat{\boldsymbol{\theta}}_{OLS} = \boldsymbol{\theta}.$$

Proof. By consistency of $\hat{\boldsymbol{\beta}}_{OLS}$, we have the consistency result:

$$p\lim_{n\to\infty} \hat{\boldsymbol{\theta}}_{OLS} = p\lim_{n\to\infty} \left(H\hat{\boldsymbol{\beta}}_{OLS}\right) = H\left(p\lim_{n\to\infty} \hat{\boldsymbol{\beta}}_{OLS}\right) = H\boldsymbol{\beta} = \boldsymbol{\theta}.$$

□

We can then construct the scalar random variable w as our test statistic:

$$w = \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0\right)^T \left[H\frac{\hat{V}}{n}H^T\right]^{-1} \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0\right)$$

This is similar to what we did earlier when we were constructing an F-test in finite sample. We can then show the asymptotic property of this test statistic.

Proposition 232 (Asymptotic Distribution of Quadratic Test Statistic)
The asymptotic distribution of the quadratic form test statistic for testing p linear restrictions is given by:

$$w = \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0\right)^T \left[H\frac{\hat{V}}{n}H^T\right]^{-1} \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0\right) \overset{a}{\sim} \chi^2(p) \quad \text{under } H_0. \tag{176}$$

Proof. The proof of this distribution follows the proof for the asymptotic distribution of the F-test. □

Under the special case of the classical linear regression model, we have that:

$$\hat{V}/n = \hat{\sigma}^2_{OLS}(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Therefore, we have that under conditional homoskedasticity:

$$w = \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0\right)^T \left[\hat{\sigma}^2_{OLS} H(\boldsymbol{X}^T\boldsymbol{X})^{-1}H^T\right]^{-1} \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0\right) \overset{a}{\sim} \chi^2(p) \quad \text{under } H_0.$$

This has an important name.

Definition 233 ((Asymptotic) Wald Statistic)
The Wald statistic is given by:

$$w = \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0\right)^T \left[\hat{\sigma}^2_{OLS} H(\boldsymbol{X}^T\boldsymbol{X})^{-1}H^T\right]^{-1} \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0\right) \overset{a}{\sim} \chi^2(p) \quad \text{under } H_0. \tag{177}$$

Therefore, joint hypothesis tests of the form: $H_0 : H\boldsymbol{\beta} = \boldsymbol{\theta}^0$ can be conducted by computing the test statistic:

$$w = \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0\right)^T \left[\hat{\sigma}^2_{OLS} H(\boldsymbol{X}^T\boldsymbol{X})^{-1}H^T\right]^{-1} \left(\hat{\boldsymbol{\theta}}_{OLS} - \boldsymbol{\theta}^0\right)$$

which has a distribution that is well approximated in large finite samples by the $\chi^2(p)$ distribution if the null hypothesis is true. We can compare the value of the test statistic to the relevant critical values of the $\chi^2(p)$ distribution at the desired level of significance. This is known as the **Wald test**.

Alternatively, we can justify this χ^2 distribution by another asymptotic argument. We can take our (finite sample) F-test for linear restrictions and show that it converges to a χ^2 distribution.

Proposition 234 (Asymptotic Distribution of F distribution is chi-squared distribution)
Denote $v \sim F(m, n)$. Then

$$mv \xrightarrow{D} \chi^2(m) \tag{178}$$

as $n \to \infty$.

Proof. First, we write out the test statistic:

$$v = \frac{w_1/m}{w_2/n}$$

whereby $w_1 \sim \chi^2(m)$ and $w_2 \sim \chi^2(n)$. This then implies that:

$$mv = \frac{w_1}{w_2/n}.$$

We now look at the denominator. By definition of χ^2-distribution, we have that we can express w_2 as the sum of independent standard normals:

$$w_2 = \sum_{i=1}^{n} z_i^2$$

whereby $z_i \sim \mathcal{N}(0, 1)$ for i = 1,2,...,n are independent standard normal random variables. First, we note that

$$\mathbb{E}[z_i] = 0$$

and

$$Var[z_i] = \mathbb{E}[z_i^2] - \mathbb{E}[z_i]^2 \quad \leftrightarrow \quad 1 = \mathbb{E}[z_i^2] - 0^2$$

$$\Rightarrow \mathbb{E}[z_i^2] = 1$$

We can see therefore see that:

$$\begin{aligned}\mathbb{E}[w_2] &= \mathbb{E}[\sum_{i=1}^{n} z_i^2] \\ &= n\mathbb{E}[z_i^2] \\ &=_{(1)} n \times 1 \\ &= n.\end{aligned}$$

Hence,

$$\mathbb{E}\left[\frac{w_2}{n}\right] = \frac{n}{n} = 1.$$

We have verified that z_i are independent and has finite first and second moments. We can therefore apply the **weak law of large numbers** to the denominator:

$$\frac{w_2}{n} = \frac{1}{n} w_2 = \frac{1}{n} \sum_{i=1}^{n} z_i^2 \xrightarrow{P} \mathbb{E}[z_i^2] = 1.$$

We can therefore apply both the weak law of large numbers and **Slutsky's theorem** to our original test statistic to get our desired result:

$$mv = \frac{w_1}{w_2/n} \xrightarrow{D} w_1 \sim \chi^2(m).$$

□

Hence, in large samples, to test p linear restrictions, we can use the test statistic pv and critical values obtained from the $\chi^2(p)$ distribution.

8 Heteroskedasticity and Generalised Least Squares

8.1 Conditional Heteroskedasticity

We now drop the assumption of homoskedasticity whereby the conditional variance

$$Var[u_i|\boldsymbol{x}_i] = \sigma^2 \quad \text{for } i = 1, ..., n$$

is common to all observations.

Assumption 235 (Conditional Heteroskedasticity)
The conditional variance is allowed to be different for different observations:

$$Var[u_i|\boldsymbol{x}_i] = \sigma_i^2 \quad \text{for } i = 1, ..., n.$$

It is important to note that neither consistency nor asymptotic normality is lost when we allow for conditional heteroskedasticity, provided the other assumptions continue to hold. Instead, the variance of the scaled estimator will be different.

Assumption 236 (Assumptions for Asymptotic Normality)
The assumptions needed for asymptotic normality of the OLS estimator are:

1. $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$;
2. The data on $(y_i, \boldsymbol{x}_i^T)$ are independent and identically distributed for $i = 1, ..., n$;
3. $\mathbb{E}[\boldsymbol{x}_i u_i] = 0$ for all $i = 1, ..., n$;
4. The $K \times K$ matrix $M_{XX} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T]$ exists and is non-singular;
5. The $K \times K$ matrix $M_{X\Omega X} = \mathbb{E}[u_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T]$ exists and is non-singular;

Recall that under these assumptions, we have the asymptotic normality of the OLS estimator:

$$\sqrt{n}[\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}] \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}). \tag{179}$$

Alternatively, we can write this as:

$$\hat{\boldsymbol{\beta}}_{OLS} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \frac{\boldsymbol{V}}{n})$$

whereby the variance-covariance matrix is:

$$\frac{\boldsymbol{V}}{n} = \left(\frac{1}{n}\right) M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}$$

We now want to find an estimator $\hat{\boldsymbol{V}}$ for $\boldsymbol{V}$ under conditional heteroskedasticity. We state what this estimator is.

Theorem 237 (Heteroskedasticity-Consistent Estimator for Variance Matrix)
The following estimator $\hat{\boldsymbol{V}}$ is a **consistent** estimator for the variance-covariance matrix $\boldsymbol{V}$:

$$\hat{\boldsymbol{V}} = \left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T\right) \left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1} \xrightarrow{P} M_{XX}^{-1} M_{\boldsymbol{X}\Omega\boldsymbol{X}} M_{XX}^{-1} = \boldsymbol{V} \tag{180}$$

Proof. We have previously shown that the *bread* in our estimator:

$$\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1} \xrightarrow{P} M_{XX}^{-1}$$

gave us a consistent estimator of the $K \times K$ matrix M_{XX}^{-1}. We now want to find an estimator for the $K \times K$ matrix $M_{X\Omega X}$.

Theorem 238 (Consistent Estimator for Filling)
The following estimator is a consistent estimator for the filling:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T \xrightarrow{P} \mathbb{E}[u_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T] = M_{X\Omega X} \tag{181}$$

whereby $\hat{u}_i \xrightarrow{P} u_i$.

We can then use Slutsky's theorem to get our desired result:

$$\hat{\boldsymbol{V}} = \left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T\right)\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1} \xrightarrow{P} M_{XX}^{-1} M_{\boldsymbol{X}\Omega\boldsymbol{X}} M_{XX}^{-1} = \boldsymbol{V}$$

□

Therefore, as $\hat{\boldsymbol{V}} \xrightarrow{P} \boldsymbol{V}$, then we can replace $\boldsymbol{V}$ by $\hat{\boldsymbol{V}}$ without affecting the limiting distribution result:

$$\hat{\boldsymbol{\beta}}_{OLS} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \frac{\hat{\boldsymbol{V}}}{n})$$

whereby:

$$\frac{\hat{\boldsymbol{V}}}{n} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T\right)(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

is the heteroskedastic-consistent estimator whereby we can compute $\frac{\hat{\boldsymbol{V}}}{n}$ using the data on $\boldsymbol{X}$ and the OLS residuals $\hat{\boldsymbol{u}}$. We can then conduct hypothesis testing with the standard errors that arise from this estimator.

Definition 239 (White Standard Errors)
The White Standard Errors (heteroskedasticity-robust standard errors) are the square roots of the scalar elements on the main diagonal of $\hat{\boldsymbol{V}}/n$.

Note that conditional homoskedasticity $\mathbb{E}[u_i^2|\boldsymbol{x}] = \sigma^2$ is a *special case* of conditional heteroskedasticity and hence we can use the White standard errors even in the case of homoskedasticity.

We can also conduct tests for conditional heteroskedasticity based on the OLS residuals.

Proposition 240 (White Test)

The White test for conditional heteroskedasticity is found by regressing the squared OLS residuals $\hat{u}_i^2$ on a constant and on all the explanatory variables and their squares and their cross-products.

That is, first retrieve the residuals $\hat{u}_i$ from the regression:

$$y_i = \beta_1 + \beta_2 x_{2i} + ...\beta_K x_{Ki} + u_i$$

and then run the auxiliary regression:

$$\hat{u}_i^2 = \gamma_1 + \boldsymbol{z}_i^T \boldsymbol{\gamma} + v_i,$$

where $\boldsymbol{z}_i$ contains all non-redundant explanatory variables, their squares, and their cross-products. Then jointly test the restriction:

$$H_0 : \boldsymbol{\gamma} = \boldsymbol{0}.$$

If we fail to reject the null, then we have insufficient evidence of conditional heteroskedasticity; failure to reject does not prove homoskedasticity.

The intuition behind the test is to see whether is the conditional variance a function of the observed variables $\boldsymbol{x}_i$.

Under the assumptions stated above, the OLS estimator remains consistent and asymptotically normal under heteroskedasticity. White standard errors consistently estimate its asymptotic variance and therefore make asymptotic inference valid. However, this is a passive response to heteroskedasticity as we are not specifying its form and hence, the OLS estimator is generally **not** asymptotically efficient.

8.2 Generalised Least Squares

We can construct a more efficient estimator if we are willing to specify what is the form of the conditional heteroskedasticity. The process of generalised least squares does this. We need to impose some assumptions before we can proceed.

Assumption 241 (Generalised Least Squares Assumptions)

We assume that:

1. $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X\beta}$
2. $Var[\boldsymbol{y}|\boldsymbol{X}] = \Omega$ whereby $\Omega \neq \sigma^2 \boldsymbol{I}$ as a known positive definite $n \times n$ matrix
3. $\boldsymbol{X}$ has full rank.

Since we have that Ω is a positive definite and known matrix, we can find a non-stochastic $n \times n$ matrix $\boldsymbol{B}$ with properties such that:

$$\boldsymbol{B}^T \boldsymbol{B} = \Omega^{-1}$$

and

$$\boldsymbol{B\Omega B}^T = \boldsymbol{I}.$$

The trick is to now transform our model from one that contains heteroskedasticity to one with homoskedasticity.

To this end, let us define the transformed variables:

$$\boldsymbol{y}^* = \boldsymbol{B}\boldsymbol{y}$$
$$\boldsymbol{X}^* = \boldsymbol{B}\boldsymbol{X}.$$

Proposition 242 (Moments of Transformed Regression Model)
The moments of the transformed regression model are:

$$\mathbb{E}[\boldsymbol{y}^*|\boldsymbol{X}] = \boldsymbol{X}^*\boldsymbol{\beta}$$
$$Var[\boldsymbol{y}^*|\boldsymbol{X}] = \boldsymbol{I}$$
$$\boldsymbol{X}^* = \boldsymbol{B}\boldsymbol{X} \text{ has full rank.}$$

Proof. First:

$$\mathbb{E}[\boldsymbol{y}^*|\boldsymbol{X}] = \mathbb{E}[\boldsymbol{B}\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{B}\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{B}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^*\boldsymbol{\beta}.$$

Then:

$$Var[\boldsymbol{y}^*|\boldsymbol{X}] = Var[\boldsymbol{B}\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{B}Var[\boldsymbol{y}|\boldsymbol{X}]\boldsymbol{B}^T = \boldsymbol{B}\Omega\boldsymbol{B}^T = \boldsymbol{I}$$

by the defining property $\boldsymbol{B}\Omega\boldsymbol{B}^T = \boldsymbol{I}$. □

Remark 243. *Since the matrix $\boldsymbol{B}$ is known and non-singular, conditioning on $\boldsymbol{X}$ and conditioning on $\boldsymbol{X}^* = \boldsymbol{B}\boldsymbol{X}$ are equivalent.*

Hence, we now have a transformed model whereby we have removed the heteroskedasticity and ended up back at homoskedasticity.

Definition 244 (Classical Linear Regression Model with Conditional Homoskedasticity)
The transformed model:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{u}^* \quad \text{with } Var[\boldsymbol{u}^*|\boldsymbol{X}^*] = \boldsymbol{I} \tag{182}$$

is a classical linear regression model with conditional homoskedasticity.

This new transformed model gives us an important result whereby we now have an efficient estimator under this transformation.

Theorem 245 (Aitken's Theorem)
The transformed model:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{u}^* \quad \text{with } Var[\boldsymbol{u}^*|\boldsymbol{X}^*] = \boldsymbol{I} \tag{183}$$

has an OLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ that is efficient in the class of linear unbiased estimators:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Omega^{-1}\boldsymbol{y}. \tag{184}$$

Remark 246. *This result says that this estimator has the smallest variance out of all linear unbiased estimators.*

Proof. First, recall that $\boldsymbol{X}^* = \boldsymbol{B}\boldsymbol{X}$ and $\boldsymbol{y}^* = \boldsymbol{B}\boldsymbol{y}$. Now, use the formula of OLS:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GLS} &= ((\boldsymbol{X}^*)^T\boldsymbol{X}^*)^{-1}(\boldsymbol{X}^*)^T\boldsymbol{y}^* \\ &= (\boldsymbol{X}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{y} \\ &= (\boldsymbol{X}^T\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{y} \\ &= (\boldsymbol{X}^T\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Omega^{-1}\boldsymbol{y}\end{aligned}$$

where we used $\boldsymbol{B}^T\boldsymbol{B} = \Omega^{-1}$. □

We seemed to have picked a random matrix $\boldsymbol{B}$ for our transformation but in fact, we could have picked any matrix proportional to $\boldsymbol{B}$ and still obtained the same GLS estimator.

Proposition 247 (GLS Estimator Invariant Under Scaling Transformation)
Let $\tilde{\boldsymbol{B}} = a\boldsymbol{B}$ for some scalar a. Then, under the transformed variables:

$$\begin{aligned}\tilde{\boldsymbol{y}} &= \tilde{\boldsymbol{B}}\boldsymbol{y} \\ \tilde{\boldsymbol{X}} &= \tilde{\boldsymbol{B}}\boldsymbol{X}\end{aligned}$$

we have the same GLS estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Omega^{-1}\boldsymbol{y}. \tag{185}$$

Proof. We simply write out the OLS estimator in the transformed model:

$$\begin{aligned}((\tilde{\boldsymbol{X}})^T\tilde{\boldsymbol{X}})^{-1}(\tilde{\boldsymbol{X}})^T\tilde{\boldsymbol{y}} &= (a^2\boldsymbol{X}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{X})^{-1}a^2\boldsymbol{X}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{y} \\ &= (\boldsymbol{X}^T\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Omega^{-1}\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{GLS}\end{aligned}$$

□

Remark 248. *This invariance property in fact holds even when we replace $\boldsymbol{B}$ by $\boldsymbol{Q}\boldsymbol{B}$ where $\boldsymbol{Q}$ is a $n \times n$ matrix with the property that $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$.*

We now impose another assumption.

Assumption 249 (Normality)
We assume normality:

$$\boldsymbol{y}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \Omega). \tag{186}$$

We then get a useful result about the estimator under normality.

Proposition 250 (GLS and MLE)
Under normality of the error term, then the GLS estimator coincides with the conditional maximum likelihood estimator when Ω is known.

Remark 251. *This is an analogous result to the fact that OLS coincides with MLE when the error terms are normally distributed.*

We can then get a finite sample distribution of the GLS estimator.

Proposition 252 (Finite Sample Normality of GLS)
Under normality of the error term, we have that the finite sample distribution of $\hat{\boldsymbol{\beta}}_{GLS}$ is:

$$\hat{\boldsymbol{\beta}}_{GLS}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\beta}, (\boldsymbol{X}^T\Omega^{-1}\boldsymbol{X})^{-1}) \tag{187}$$

8.2.1 Feasible Generalized Least Squares

In practice, we cannot compute this GLS estimator as we do not know the conditional variance matrix Ω. Therefore, we need to estimate Ω when constructing our estimator.

Definition 253 (Feasible GLS Estimator)
The **feasible generalised least squares** estimator is given by:

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}^T\hat{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\Omega}^{-1}\boldsymbol{y} \tag{188}$$

where $\hat{\Omega}$ is an estimator of Ω.

In practice, we can estimate the feasible GLS by again transforming our model and running OLS. The transformation in this case is given by:

$$\boldsymbol{y}^* = \hat{\boldsymbol{B}}\boldsymbol{y}$$
$$\boldsymbol{X}^* = \hat{\boldsymbol{B}}\boldsymbol{X}$$

where we require that $\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}} = \hat{\Omega}^{-1}$ and $\hat{\boldsymbol{B}}\hat{\Omega}\hat{\boldsymbol{B}}^T = \boldsymbol{I}$.

We can derive the properties of the FGLS estimator. However, such properties depend on the properties of $\hat{\Omega}$ as well.

Proposition 254 (Asymptotic Normality of FGLS)
Suppose that $\Omega = \Omega(\phi)$ is correctly specified with a fixed-dimensional parameter ϕ, that $\hat{\phi}$ is consistent, and that the standard regularity conditions hold. Then, $\hat{\boldsymbol{\beta}}_{FGLS}$ has the asymptotic distribution:

$$\hat{\boldsymbol{\beta}}_{FGLS} \overset{a}{\sim} \mathcal{N}\left(\boldsymbol{\beta}, (\boldsymbol{X}^T\hat{\Omega}^{-1}\boldsymbol{X})^{-1}\right). \tag{189}$$

Furthermore, $\hat{\boldsymbol{\beta}}_{FGLS}$ is asymptotically efficient.

We can therefore see that consistency of $\hat{\Omega}$ is quite important in order for us to have nice properties for the FGLS. However, obtaining a consistent estimator $\hat{\Omega}$ is not straightforward. The reason is that the $n \times n$ symmetric matrix Ω has n(n+1)/2 distinct elements. Even if we restrict our attention that all off-diagonal elements are zero, we will still have n distinct elements in Ω, which we cannot estimate consistently from a sample of size n.

We have just seen the issue for estimating Ω even with many assumptions. Hence, in order for us to actually estimate Ω, we need to specify a parametric model for the conditional variance matrix $Var[\boldsymbol{y}|\boldsymbol{X}] = \Omega$ of the form:

$$\Omega = \Omega(\phi)$$

where the $n \times n$ matrix $\Omega(\phi)$ is a function of the vector ϕ which contains a finite number of parameters that **does not increase** with the sample size. Therefore, we in fact estimate ϕ instead and see if we can consistently estimate ϕ. This gets us around the earlier issue with having too many parameters to estimate that we mentioned earlier.

Therefore, if the specification of the conditional variance matrix $Var[\boldsymbol{y}|\boldsymbol{X}] = \Omega(\phi)$ is correct and we can find a **consistent estimator** $\hat{\phi}$ of the vector ϕ, then we can then use the consistent estimator:

$$\hat{\Omega} = \Omega(\hat{\phi})$$

to obtain the asymptotically efficient FGLS estimator.

8.2.2 Weighted Least Squares

We look at a simple example of computing the FGLS estimator. Let us specify the conditional variance to be:

$$Var[y_i|\boldsymbol{X}] = Var[u_i|\boldsymbol{X}] = \sigma_i^2$$

to be proportional to the squared values of one of the regressors

$$\sigma_i^2 = \sigma^2 x_{Ki}^2, \qquad x_{Ki} \neq 0.$$

We picked the squared value to ensure a non-negative variance, and assume $x_{Ki} \neq 0$ so that the variance is positive and the transformation below is well-defined.

We then define the transformations:

$$y_i^* = \frac{y_i}{x_{Ki}}$$
$$x_{ji}^* = \frac{x_{ji}}{x_{Ki}}$$

for each $j = 1, ..., K$. Note that under this transformation, the intercept $x_{1i} = 1$, we have that:

$$x_{1i}^* = \frac{x_{1i}}{x_{Ki}} = \frac{1}{x_{Ki}} = (x_{Ki})^{-1}$$

and we also have that:

$$x_{Ki}^* = \frac{x_{Ki}}{x_{Ki}} = 1.$$

Definition 255 (Transformed Model in Weighted Least Squares)
Define the transformations:

$$y_i^* = \frac{y_i}{x_{Ki}}$$
$$x_{ji}^* = \frac{x_{ji}}{x_{Ki}}$$

for each $j = 1, ..., K$. Then, the transformed model in weighted least squares is:

$$y_i^* = (x_i^*)^T \boldsymbol{\beta} + u_i^*. \tag{190}$$

Remark 256. *The intuition to what we are doing here is that we are downweighting the observations where the variance of u_i is relatively high.*

We now have a useful result from this!

Proposition 257 (Conditional Homoskedasticity)
Under the weighted least squares transformations, we have conditional homoskedasticity:

$$Var[y_i^*|\boldsymbol{X}] = \sigma^2 \tag{191}$$

for all $i = 1, ..., n$.

Proof. We can write this out:

$$\begin{aligned} Var[y_i^*|\boldsymbol{X}] &= Var[u_i^*|\boldsymbol{X}] \\ &= \frac{\sigma_i^2}{x_{Ki}^2} \\ &= \frac{\sigma^2 x_{Ki}^2}{x_{Ki}^2} \\ &= \sigma^2. \end{aligned}$$

□

We also give an example with a linear model with one exogenous regressor and an intercept:

$$y_i = \beta_1 + \beta_2 x_{2i} + u_i.$$

Suppose we impose the parametric form:

$$\mathbb{E}[u_i^2|\boldsymbol{x}_i] = \phi_1 + \phi_2 x_{2i}^2 = \sigma^2(\boldsymbol{x}_i).$$

This is equivalent to saying:

$$u_i^2 = \phi_1 + \phi_2 x_{2i}^2 + \epsilon_i.$$

We first retrieve the OLS residuals $\hat{u}_i$ from the original regression.
We can retrieve ϕ_1 and ϕ_2 by regressing:

$$\hat{u}_i^2 = \phi_1 + \phi_2 x_{2i}^2 + \epsilon_i.$$

Under correct specification and the usual regularity conditions, the OLS estimators $\hat{\phi}_1$ and $\hat{\phi}_2$ consistently estimate ϕ_1 and ϕ_2. We then form the fitted conditional variance and standard deviation:

$$\hat{\sigma}_i^2 = \hat{\phi}_1 + \hat{\phi}_2 x_{2i}^2, \qquad \hat{\sigma}_i = \sqrt{\hat{\sigma}_i^2},$$

provided the fitted variance is positive.

Here, we weight each observation by a factor which is proportional to $\frac{1}{\sigma_i}$. Hence, this is where the name weighted least squares estimator comes from. We give less weight to observations where the variance of u_i is relatively high and more weight to observations where the variance of u_i is relatively low. If our specification for the conditional variance $Var[\boldsymbol{y}|\boldsymbol{X}] = \Omega(\phi)$ is correct, and we estimate $\Omega(\phi)$ consistently, this weighting is the source of the efficiency gain compared to OLS.

However, the feasible GLS estimator is **not** generally the conditional maximum likelihood estimator even under normality of error terms. This is because FGLS uses a consistent estimator of ϕ to construct a consistent estimator

of Ω and then optimises with respect to $\boldsymbol{\beta}$ only. Meanwhile, the conditional MLE maximises the likelihood function with respect to $\boldsymbol{\beta}$ and ϕ jointly. These are two different estimators unless we have that:

$$\hat{\phi} = \hat{\phi}_{ML},$$

which is unlikely. As a result, we generally have that:

$$\hat{\boldsymbol{\beta}}_{FGLS} \neq \hat{\boldsymbol{\beta}}_{ML}$$

although they are asymptotically equivalent under correct specification and standard regularity conditions. With a correctly specified conditional mean and well-behaved weights, consistency need not depend on the parametric variance specification being correct, but the efficiency property **does** depend on that specification.

8.3 Clustering

We now look at relaxing the assumption that observations are independent over $i = 1, ..., n$.

Suppose that each of our n individual observations is allocated to one of $c = 1, ..., C$ clusters. Let n_c denote the number of observations in cluster c, and let y_{ic} be the dependent variable for observation $i = 1, \ldots, n_c$ in cluster c. We define the vector of explanatory variables $\boldsymbol{x}_{ic}$ and error term u_{ic} similarly.

Definition 258 (Clustered Linear Regression Model)
The clustered linear regression model is given by:

$$y_{ic} = \boldsymbol{x}_{ic}^T \boldsymbol{\beta} + u_{ic} \tag{192}$$

for $i = 1, ..., n_c$ and $c = 1, ..., C$.

We denote n_c as the number of observations in cluster c and we must have that:

$$\sum_{c=1}^{C} n_c = n.$$

We now define $\boldsymbol{y}_c$ to be the $n_c \times 1$ vector containing the observations on y_{ic} for the n_c observations in cluster c. We also define the matrix $\boldsymbol{X}_c$ which is the $n_c \times K$ matrix where each row contains the row vector $\boldsymbol{x}_{ic}^T$ for an observation in cluster c. Finally, we have the error term vector $\boldsymbol{u}_c$ as the $n_c \times 1$ vector. We can then define the system of C equations:

$$\boldsymbol{y}_c = \boldsymbol{X}_c \boldsymbol{\beta} + \boldsymbol{u}_c \quad \text{for c = 1,..., C.}$$

We can then stack these vectors again to get the $n \times 1$ vectors $\boldsymbol{y}$ and $\boldsymbol{u}$ in addition to the $n \times K$ matrix $\boldsymbol{X}$ to get the model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}.$$

We can now define the OLS estimator of this model.

Proposition 259 (Clustered OLS)

Let the clustered linear regression model be written as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{193}$$

Then the OLS estimator of $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \left(\sum_{c=1}^{C}\boldsymbol{X}_c^T\boldsymbol{X}_c\right)^{-1}\left(\sum_{c=1}^{C}\boldsymbol{X}_c^T\boldsymbol{y}_c\right). \tag{194}$$

The important result from aggregating the variables by clusters is that the observations $(y_{ic}, \boldsymbol{x}_{ic}^T)$ are now assumed to be **independent across clusters** but not within clusters. We will do something similar later on when we look at panel data. This allows the error terms u_{ic} and u_{jc} to be correlated for different observations in the same cluster c.

Asymptotic results are derived by letting the number of clusters $C \to \infty$. We state important assumptions that we need to help us derive such results.

Assumption 260 (Strict Exogeneity Assumption)

We assume strict exogeneity of the error terms:

$$\mathbb{E}[\boldsymbol{u}_c|\boldsymbol{X}] = \boldsymbol{0} \quad \text{for } c = 1, \ldots, C, \tag{195}$$

$$\mathbb{E}[\boldsymbol{u}_c\boldsymbol{u}_d^T|\boldsymbol{X}] = \boldsymbol{0} \quad \text{for } c \neq d, \tag{196}$$

while dependence among observations within the same cluster is unrestricted.

We have that under these assumptions:

$$\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^T|\boldsymbol{X}] = \Omega = \begin{bmatrix} \Omega_1 & 0 & \ldots & 0 \\ 0 & \Omega_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \Omega_C \end{bmatrix}$$

where each Ω_c is a symmetric and positive definite $n_c \times n_c$ matrix.

Proposition 261 (Asymptotic Normality)

Suppose clusters are independent, $C \to \infty$, suitable moment and rank conditions hold, and no cluster dominates the sample. Then we have asymptotic normality and consistency of the OLS estimator:

$$\sqrt{C}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, \boldsymbol{V}). \tag{197}$$

We can construct a consistent estimator of $\boldsymbol{V}$ by:

$$\hat{\boldsymbol{V}} = C(\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\sum_{c=1}^{C}\boldsymbol{X}_c^T\hat{\boldsymbol{u}}_c\hat{\boldsymbol{u}}_c^T\boldsymbol{X}_c\right)(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

where $\hat{\boldsymbol{u}}_c = \boldsymbol{y}_c - \boldsymbol{X}_c\hat{\boldsymbol{\beta}}_{OLS}$ is the vector of OLS residuals for the n_c observations in cluster c.

Definition 262 (Cluster-Robust Variance Matrix Estimator)

The variance of $\hat{\boldsymbol{\beta}}_{OLS}$ can be estimated using the **cluster-robust variance matrix** estimator:

$$\frac{\hat{\boldsymbol{V}}}{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\sum_{c=1}^{C}\boldsymbol{X}_c^T\hat{\boldsymbol{u}}_c\hat{\boldsymbol{u}}_c^T\boldsymbol{X}_c\right)(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{198}$$

This allows us to carry out asymptotically valid inference while allowing for dependence within clusters.

Proposition 263 (Cluster-Robust Variance Matrix Estimator and White's Estimator)

The cluster-robust variance matrix estimator simplifies to White's heteroskedasticity-robust variance matrix estimator in the special case with one observation in each cluster.

9 Endogeneity

9.1 Endogenous Explanatory Variable

In practice, there are many instances when the orthogonality condition $\mathbb{E}[\boldsymbol{x}_i u_i] = 0$ is violated. This is a big issue as the orthogonality condition was required in order to establish consistency. In particular, if one of the explanatory variables is correlated with the error term u_i, then $\text{Cov}(\boldsymbol{x}_i, u_i) \neq 0$. In a model with an intercept, so that $\mathbb{E}[u_i] = 0$, this violates both $\mathbb{E}[u_i|\boldsymbol{x}_i] = 0$ and $\mathbb{E}[\boldsymbol{x}_i u_i] = 0$, and hence the OLS estimator will generally be both **biased** and **inconsistent**.

There are two situations we will look at where this occurs.

1. One or more of the explanatory variables is measured with error. This is known as **measurement error** or **errors-in-variables**.
2. One or more of the relevant explanatory variables is omitted from the model. This is known as **omitted variable bias**.

9.2 Measurement Error

A common concern in applied econometrics is that relevant explanatory variables may be poorly measured. In particular, **attenuation bias** is a potential cause for concern.

Definition 264 (Attenuation Bias)
Attenuation bias is the biasing of the OLS estimator towards 0.

We can state the condition for when attenuation bias occurs.

Proposition 265 (Condition for Attenuation Bias)
Under classical additive measurement error in a single explanatory variable, the OLS slope estimator is attenuated towards zero asymptotically.

Remark 266. *This asymptotic bias does not disappear in large samples. An exact finite-sample direction requires additional distributional assumptions.*

Fortunately, it is the case that if the dependent variable $\boldsymbol{y}$ is measured with error, attenuation bias does **not** occur.

Proposition 267 (Mismeasured Dependent Variables does Not Cause Attenuation Bias)
Attenuation bias does **not** occur if the dependent variable $\boldsymbol{y}$ is measured with error, provided the measurement error in $\boldsymbol{y}$ is uncorrelated with the correctly measured $\boldsymbol{X}$.

We now mathematically show this. Assume we have a **single explanatory variable** and no intercept term in our model:

$$y_i^* = x_i^* \beta + u_i \quad \text{for i = 1,...,n} \tag{199}$$

whereby y_i^* and x_i^* denote the true values of these variables which we may not observe. To make our analysis easier, we impose the following assumptions.

Assumption 268 (Measurement Error Assumptions)

We assume that the following variables have mean zero:

$$\mathbb{E}[u_i] = 0$$
$$\mathbb{E}[x_i^*] = 0$$
$$\mathbb{E}[y_i^*] = 0$$

for $i = 1, ..., n$.

Furthermore, we assume that

$$\mathbb{E}[x_i^* u_i] = 0$$

for $i = 1, ..., n$.

Finally, we have **independent and identically distributed** data.

Remark 269. *A motivation for why this may be the case could be that these variables are expressed as deviations from their sample means.*

We now show our first result.

Proposition 270 (Mismeasured Dependent Variables does Not Cause Attenuation Bias)

Suppose that there is an additive, mean zero measurement error in the dependent variable only

$$y_i = y_i^* + v_i$$

where y_i is the observed value, y_i^* is the true value, and v_i is the measurement error with $\mathbb{E}[v_i] = 0$ for $i = 1, ..., n$.

Let the model be:

$$y_i^* = x_i^* \beta + u_i \quad \text{for i = 1,...,n}$$

where x_i^* is the true value. Then, there is **no attenuation bias** if the measurement error in y_i is uncorrelated with the correctly measured x_i^*:

$$\mathbb{E}[x_i^* v_i] = 0 \quad \text{for i = 1,...,n.}$$

Proof. Start off with the model:

$$y_i^* = x_i^* \beta + u_i$$

and substitute in the mismeasured dependent variable $y_i^* = y_i - v_i$ This results in

$$(y_i - v_i) = x_i^* \beta + u_i$$

or alternatively:

$$y_i = x_i^* \beta + (u_i + v_i)$$

For consistency to hold, we require that x_i^* is uncorrelated with the new error term $(u_i + v_i)$ which contains the true linear model error u_i and the measurement error v_i.

By our measurement error assumptions, we assumed that $\mathbb{E}[x_i^* u_i] = 0$. Furthermore, in the proposition, we also assumed that $\mathbb{E}[x_i^* v_i] = 0$.

Therefore, $\mathbb{E}[x_i^*(u_i + v_i)] = 0$ and therefore the OLS estimator for β is a consistent estimator. □

We now show that there are issues when the explanatory variable is mismeasured.

Proposition 271 (Mismeasured Explanatory Variables Cause Attenuation Bias)
Suppose that there is an additive, mean zero measurement error in the explanatory variable only

$$x_i = x_i^* + e_i$$

where x_i is the observed value, x_i^* is the true value, and e_i is the measurement error with $\mathbb{E}[e_i] = 0$ for $i = 1, ..., n$. Furthermore, suppose that $\mathbb{E}[x_i^* e_i] = 0$, $\mathbb{E}[u_i e_i] = 0$, and $0 < Var[e_i] < \infty$.

Let the model be:

$$y_i^* = x_i^* \beta + u_i \quad \text{for i = 1,...,n}$$

where y_i^* is the true value. Then, the OLS estimator is **biased towards zero asymptotically**.

Proof. Start off with the model:

$$y_i^* = x_i^* \beta + u_i$$

and substitute in the mismeasured explanatory variable $x_i^* = x_i - e_i$ This results in

$$y_i^* = (x_i - e_i)\beta + u_i$$

or alternatively:

$$y_i^* = x_i \beta + (u_i - e_i \beta)$$

Under the classical measurement-error assumptions, $Cov(x_i, e_i) = Var(e_i) > 0$ and

$$Cov(x_i, u_i - e_i \beta) = -\beta Var(e_i).$$

This implies a non-zero correlation between x_i and the error term in this model $(u_i - e_i \beta)$ whenever $\beta \neq 0$.

From this, we can show that the OLS estimator of β here is biased. First, we look at the model under a mismeasurement in the explanatory variable:

$$y_i^* = x_i \beta + (u_i - e_i \beta).$$

Suppose that the true value of $\beta > 0$. Then, this implies a **negative correlation** between the observed variable x_i and the error term $(u_i - e_i \beta)$. This causes a **downward bias** in the OLS estimator of β. As $\beta > 0$ was positive, this **downward bias** makes the regression slope flatter and biases the OLS estimator **towards zero**.

Suppose that the true value of $\beta < 0$. Then, this implies a **positive correlation** between the observed variable x_i and the error term $(u_i - e_i \beta)$. This causes a **upward bias** in the OLS estimator of β. As $\beta < 0$ was negative, this **upward bias** makes the regression slope flatter and biases the OLS estimator **towards zero**.

Therefore, under both scenarios, the OLS estimator of β will be **biased towards zero** and this is known as **attenuation bias**. □

We now want to show that a mismeasurement in the explanatory variable also causes the OLS estimator to be inconsistent. To do so, we need some further assumptions.

Assumption 272 (Classical Errors-in-variables Assumption)

For $i = 1, ..., n$, we assume that:

1. $\mathbb{E}[x_i^* e_i] = 0$ Measurement error is uncorrelated with the true value of x_i^*
2. $\mathbb{E}[u_i e_i] = 0$ Measurement error is uncorrelated with the true model error u_i
3. $Var[e_i] = \sigma_e^2$ Measurement error has finite variance
4. $Var[x_i^*] = \sigma_{x^*}^2$ Population variance of the true value x_i^* exists and is finite.

Proposition 273 (Mismeasured Explanatory Variables Causes Inconsistency of OLS)

Assume that the classical errors-in-variables assumption holds. Then, suppose that there is an additive, mean zero measurement error in the explanatory variable only

$$x_i = x_i^* + e_i$$

where x_i is the observed value, x_i^* is the true value, and e_i is the measurement error with $\mathbb{E}[e_i] = 0$ for $i = 1, ..., n$.

Let the model be:

$$y_i^* = x_i^* \beta + u_i \quad \text{for i = 1,...,n}$$

where y_i^* is the true value. Then, the OLS estimator is **inconsistent**.

Proof. Following our earlier derivation when showing the bias of the OLS estimator, we can rewrite the model as:

$$y_i^* = x_i \beta + (u_i - e_i \beta).$$

Then, by the formula of the OLS estimator, we have that:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y^* = \frac{\sum_{i=1}^n x_i y_i^*}{\sum_{i=1}^n x_i^2} = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i^*}{\frac{1}{n}\sum_{i=1}^n x_i^2}$$

since x_i and y_i^* are scalars. We now plug in $x_i = x_i^* + e_i$ and $y_i^* = x_i^* \beta + u_i$ into our OLS formula to get:

$$= \frac{\frac{1}{n}\sum_{i=1}^n \left(x_i^* + e_i\right)\left(x_i^* \beta + u_i\right)}{\frac{1}{n}\sum_{i=1}^n \left(x_i^* + e_i\right)^2}$$

We can then expand out the terms:

$$= \frac{\beta \frac{1}{n}\sum_{i=1}^n (x_i^*)^2 + \frac{1}{n}\sum_{i=1}^n x_i^* u_i + \beta \frac{1}{n}\sum_{i=1}^n e_i x_i^* + \frac{1}{n}\sum_{i=1}^n e_i u_i}{\frac{1}{n}\sum_{i=1}^n (x_i^*)^2 + 2\frac{1}{n}\sum_{i=1}^n (x_i^* e_i) + \frac{1}{n}\sum_{i=1}^n e_i^2}$$

Now, we take limits on both sides and apply the weak law of large numbers:

$$\begin{aligned}
p\lim_{n\to\infty}\hat{\beta}_{OLS} &= \frac{\beta p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i^*)^2 + p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}x_i^*u_i + \beta p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}e_ix_i^* + p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}e_iu_i}{p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i^*)^2 + 2p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i^*e_i) + p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}e_i^2} \\
&= \frac{\beta\mathbb{E}[(x_i^*)^2] + \mathbb{E}[x_i^*u_i] + \beta\mathbb{E}[e_ix_i^*] + \mathbb{E}[e_iu_i]}{\mathbb{E}(x_i^*)^2 + 2\mathbb{E}[x_i^*e_i] + \mathbb{E}[e_i^2]} \\
&=_{(1)} \frac{\beta\mathbb{E}[(x_i^*)^2] + 0 + 0 + 0}{\mathbb{E}(x_i^*)^2 + 0 + \mathbb{E}[e_i^2]} \\
&=_{(2)} \left(\frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_e}\right)\beta \\
&= \frac{\beta}{1 + (\sigma^2_e/\sigma^2_{x^*})} \neq \beta
\end{aligned}$$

if $\sigma_e^2 > 0$ whereby $=_{(1)}$ uses the classical errors-in-variables assumption and $=_{(2)}$ uses the fact that the first moment of the true variable x_i^* is 0.

Therefore, we have the result that:

$$p\lim_{n\to\infty}\hat{\beta}_{OLS} = \frac{\beta}{1 + (\sigma^2_e/\sigma^2_{x^*})} < \beta \quad \text{for } \beta > 0 \text{ and } \sigma_e^2 > 0$$

$$p\lim_{n\to\infty}\hat{\beta}_{OLS} = \frac{\beta}{1 + (\sigma^2_e/\sigma^2_{x^*})} > \beta \quad \text{for } \beta < 0 \text{ and } \sigma_e^2 > 0$$

where we can see that the OLS estimator of β is inconsistent, with a bias towards zero that does not diminish as the sample becomes large. □

For a given level $\sigma^2_{x^*}$, the severity of the **attenuation bias** increases with the variance of the measurement error σ_e^2. That is, the bias gets worse as there is more variance in the measurement error. Hence, the magnitude of the inconsistency depends inversely on the **signal-to-noise** ratio ($\sigma^2_{x^*}/\sigma^2_e$).

One thing to be careful of is that under heteroskedastic measurement error, the presence of measurement error may introduce an incorrect indication of non-linearity in the relationship between the true value y_i^* and the observed variable x_i.

9.2.1 Multiple regression with errors in variables

We can now generalise our scenario to a multiple regression with errors in variables. Suppose that we have K explanatory variables now whereby:

$$\begin{aligned}
y_i^* &= (\boldsymbol{x}_i^*)^T\boldsymbol{\beta} + u_i \\
\boldsymbol{x}_i^T &= (\boldsymbol{x}_i^*)^T + \boldsymbol{e}_i^T \quad \text{(Measurement error in explanatory variables)}
\end{aligned}$$

whereby $\boldsymbol{x}_i^T, (\boldsymbol{x}_i^*)^T, \boldsymbol{e}_i^T$ are $1 \times K$ vectors. We still have measurement error in the explanatory variables. We can then substitute the terms in to get:

$$y_i^* = \boldsymbol{x}_i^T\boldsymbol{\beta} + (u_i - \boldsymbol{e}_i^T\boldsymbol{\beta})$$

Under classical vector measurement error, we have

$$\mathbb{E}[\boldsymbol{x}_i(u_i - \boldsymbol{e}_i^T\boldsymbol{\beta})] = -\mathbb{E}[\boldsymbol{e}_i\boldsymbol{e}_i^T]\boldsymbol{\beta},$$

which is generally non-zero. In particular, it is non-zero if the measurement-error covariance matrix is positive definite and $\boldsymbol{\beta} \neq \boldsymbol{0}$. The OLS estimator of the $K \times 1$ parameter vector $\boldsymbol{\beta}$ will therefore generally be biased and inconsistent.

We have that if **only one** of the explanatory variables $\boldsymbol{x}_i$ is measured with error, we can show that:

1. The OLS estimator of the coefficient on that mismeasured variable is biased towards zero (**attenuation bias**)
2. The OLS estimator of the coefficients on the other explanatory variables are also biased but in **unknown directions**.

If several explanatory variables are measured with error, it is very difficult to sign the biases for any of the coefficients.

9.3 Omitted Variable

Another common concern in applied econometrics is that relevant explanatory variables may be omitted from the model. This can cause the OLS estimator to be **biased** if the omitted variable is relevant and correlated with the included regressor and the OLS is also inconsistent. Finally, the direction of the bias depends on the sign of the correlation between the included variable and the omitted variable and the sign of the coefficient on the omitted variable in the data generating process.

9.3.1 Single Linear Regression with Omitted Variables

First consider a simple model with one included variable x_{1i} and one omitted variable x_{2i}. The true process which generates the outcomes y_i is assumed to be:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + u_i \quad \text{for } i = 1, ..., n$$

satisfying

$$\mathbb{E}[u_i] = \mathbb{E}[x_{1i}] = \mathbb{E}[x_{2i}] = 0$$

and

$$\mathbb{E}[x_{1i}u_i] = \mathbb{E}[x_{2i}u_i] = 0$$

Suppose that the model excludes x_{2i}:

$$y_i = x_{1i}\beta_1 + (u_i + x_{2i}\beta_2) \quad \text{for } i = 1, ..., n$$

We can stack across the n observations to get:

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + (\boldsymbol{u} + \boldsymbol{X}_2\beta_2)$$

The OLS estimator of β_1 is then given by:

$$\hat{\beta}_1 = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y}$$

$$\begin{aligned}
\hat{\beta}_1 &= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y} \\
&= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T(\boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{u}) \\
&= \beta_1 + [(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2]\beta_2 + (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{u} \\
&= \beta_1 + \hat{\delta}\beta_2 + (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{u}
\end{aligned}$$

Here, we defined:

$$\hat{\delta} \equiv (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2$$

to be the coefficient δ in a linear projection of the omitted variable x_{2i} on the included variable x_{1i}:

$$x_{2i} = x_{1i}\delta + e_i \quad \text{for } i = 1, ..., n$$

Here, we have that $\hat{\delta}$ is actually a consistent estimator of δ so we have that:

$$p \lim_{n\to\infty} \hat{\delta} = \delta.$$

We also use the fact that:

$$\mathbb{E}[\boldsymbol{X}_1^T \boldsymbol{u}] = \boldsymbol{0}$$

to get rid of the third term by the weak law of large numbers.

We go back to our original expression and we take probability limits:

$$p \lim_{n\to\infty} \hat{\beta}_1 = \beta_1 + (p \lim_{n\to\infty} \hat{\delta})\beta_2 = \beta_1 + \delta\beta_2$$

Therefore, we have the important expression:

$$p \lim_{n\to\infty} \hat{\beta}_1 = \beta_1 + \delta\beta_2 \tag{200}$$

We can see two things. The OLS estimator of β_1 in the model that omits x_{2i} is **inconsistent** unless either:

1. $\delta = 0$ The omitted variable is orthogonal to the included variable;
2. $\beta_2 = 0$ The omitted variable is not a relevant explanatory variable in the true model.

We can see the direction of the bias of the OLS estimator for an omitted relevant explanatory variable ($\beta_2 \neq 0$). The sign of the asymptotic bias is the sign of $\delta\beta_2$. Thus, if x_{1i} and x_{2i} are positively correlated, then $\delta > 0$ and the bias is upward when $\beta_2 > 0$ but downward when $\beta_2 < 0$. The signs are reversed when x_{1i} and x_{2i} are negatively correlated, so that $\delta < 0$.

9.3.2 Multiple Regression with Omitted Variables

We now look at the regression case with multiple explanatory variables. Suppose we had K_1 included regressors and K_2 omitted variables. Then, our regression model is:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + (\boldsymbol{u} + \boldsymbol{X}_2\boldsymbol{\beta}_2)$$

where $\boldsymbol{X}_1$ is a $n \times K_1$ matrix, $\boldsymbol{\beta}_1$ is a $K_1 \times 1$ vector, $\boldsymbol{X}_2$ is a $n \times K_2$ matrix, and $\boldsymbol{\beta}_2$ is a $K_2 \times 1$ vector.

Following the same steps as before, we can obtain the expression for the OLS estimator as:

$$p \lim_{n\to\infty} \hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1 + (p \lim_{n\to\infty} (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2)\boldsymbol{\beta}_2 \tag{201}$$

whereby

$$\hat{\boldsymbol{\delta}} = (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2$$

is the $K_1 \times K_2$ estimator of the coefficient matrix in the linear projection:

$$\boldsymbol{X}_2 = \boldsymbol{X}_1\boldsymbol{\delta} + \boldsymbol{e}$$

It is harder to be confident about the direction of the biases now.

However, if there was only one omitted variable that is correlated with several of the included explanatory variables, then the bias in each estimated coefficients will depend on the partial correlation with the omitted variable and not the simple correlation between the included regressor and omitted variable. That is, under the single omitted variable case:

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_{K-1} x_{K-1,i} + (\beta_K x_{Ki} + u_i)$$

where $\mathbb{E}[x_{ki} u_i] = 0$ for $k = 1, ..., K$ and x_{Ki} is the omitted variable, we can show that the source of the bias for each of the OLS estimators is:

$$p \lim_{n\to\infty} \hat{\beta}_k = \beta_k + (p \lim_{n\to\infty} \hat{\delta}_k)\beta_K \quad \text{for } k = 1, ..., K-1$$

whereby $\hat{\delta}_k$ is from the linear projection:

$$X_{Ki} = \delta_1 + \delta_2 x_{2i} + ... + \delta_{K-1} x_{K-1,i} + v_i$$

whereby $\mathbb{E}[x_{ki} v_i] = 0$ for $k = 1, ..., K-1$.

If there are several omitted variables, it is hard to predict the direction of the biases. And hence, the OLS estimator will be a biased and inconsistent estimator of the true population parameters.

10 Instrumental Variables

We have seen what are the issues when we have correlation between the error term and at least one of the explanatory variables. Instrumental variables is a method of estimation that we can use to rectify this situation.

Before we proceed, we recap a few things. First, recall that in a **linear projection model**:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

where $\boldsymbol{X}$ is a $N \times K$ matrix (K explanatory variables) we have the **orthogonality condition**:

$$\mathbb{E}[\boldsymbol{X}^T \boldsymbol{u}] = \boldsymbol{0}$$

where this holds for a system of K equations (for each explanatory variable). Now, if one of our explanatory variables is an intercept term $\boldsymbol{X}_1 = \boldsymbol{1}_N$, then this orthogonality condition implies:

$$\mathbb{E}[\boldsymbol{X}_1^T \boldsymbol{u}] = \mathbb{E}[\boldsymbol{1}_N^T \boldsymbol{u}] = 0.$$

Therefore, under i.i.d. sampling, when the regression matrix $\boldsymbol{X}$ contains an intercept, the population projection error has mean zero:

$$\mathbb{E}[u_i] = 0.$$

If we do not have an intercept, however, the error term need not have mean zero:

$$\mathbb{E}[u_i] \text{ need not equal } 0.$$

This was an important result we need when we want to show finite sample properties of OLS. Therefore, if we **do not** include an intercept term, we then need to impose the assumption that:

$$\mathbb{E}[u_i] = 0.$$

So now going back, suppose we have a structural model:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i \tag{202}$$

for observations $i = 1, ..., n$ and $\boldsymbol{x}_i$ is a $K \times 1$ vector for our K explanatory variables. If we had correlation between the explanatory variable $\boldsymbol{x}_i$ and the error term u_i, we would have issues with OLS not being a consistent estimator. We now go on to explain what we can do to solve this.

10.1 Two Stage Least Squares

The general intuition for 2SLS is that we want to construct predicted values of $\boldsymbol{x}_i$ which are not correlated with the error terms. To construct these predicted values, we can project the explanatory variables $\boldsymbol{x}_i$ onto a set of **instrumental variables** which we hope are:

1. **Informative**: Related to the explanatory variable $\boldsymbol{x}_i$
2. **Valid**: Uncorrelated with the error terms u_i.

If we have these properties, we can replace $\boldsymbol{x}_i$ by these predicted values and therefore guarantee consistency of our OLS estimator $\hat{\boldsymbol{\beta}}$.

We will now go into 2SLS. We make the following assumptions to make life easier by restricting our attention to the case of a single endogenous variable and instrumental variable.

Assumption 274 (Single Explanatory And Instrumental Variable)
We have a single endogenous explanatory variable x_i. We have a single outside instrumental variable z_i. We assume that both have mean zero:

$$\mathbb{E}[x_i] = 0 \tag{203}$$

$$\mathbb{E}[z_i] = 0. \tag{204}$$

We also impose an assumption on the structural error term.

Assumption 275 (Zero Mean Error)
The structural error term has zero mean:

$$\mathbb{E}[u_i] = 0. \tag{205}$$

We shall explain why we have these assumptions later on. First, let us look at our structural model again:

$$y_i = x_i\beta + u_i$$

for $i = 1, ..., n$. By our assumption, x_i is endogenous which then implies that:

$$\mathbb{E}[x_i u_i] \neq 0.$$

This is an issue as this is the key orthogonality condition we need for the consistency of OLS estimators. We now describe the first stage of the two stage least squares estimator.

Definition 276 (First Stage Linear Projection)
Let x_i be our endogenous variable and z_i be our outside instrumental variable. Define our structural model as

$$y_i = x_i\beta + u_i$$

for $i = 1, ..., n$. Then, the first stage linear projection is:

$$x_i = z_i\pi + r_i \tag{206}$$

for $i = 1, ..., n$ where π is a scalar parameter coefficient.

Now, if we look at the first linear projection, we can see where our initial assumptions that $\mathbb{E}[x_i] = \mathbb{E}[z_i] = 0$ came in. If we regressed:

$$x_i = z_i\pi + r_i$$

which **does not** contain an intercept term, we are not guaranteed that the error term r_i has mean zero $\mathbb{E}[r_i] = 0$. Therefore, by assuming both $\mathbb{E}[x_i] = 0$ and $\mathbb{E}[z_i] = 0$, we do not need to include an intercept and we automatically get the result we need for $\mathbb{E}[r_i] = 0$.

We now define the two conditions needed for an instrumental variable: validity and informativeness.

Definition 277 (Valid Instrument)
An instrument z_i is **valid** if it is uncorrelated with the structural error term:

$$\mathbb{E}[z_i u_i] = 0. \tag{207}$$

Definition 278 (Informative Instrument)
An instrument z_i is **informative** if it is correlated with the endogenous variable x_i:

$$\mathbb{E}[z_i x_i] \neq 0. \tag{208}$$

Note that we can rewrite the condition for informativeness in a different manner. Our first stage linear projection was:

$$x_i = z_i \pi + r_i$$

which therefore implies that correlation between z_i and x_i, $\mathbb{E}[z_i x_i] \neq 0$, can also be expressed as:

$$\pi \neq 0.$$

Proposition 279 (Alternative Expression for Informative Instrument)
Suppose an instrument z_i is **informative** for the endogenous variable x_i

$$\mathbb{E}[z_i x_i] \neq 0.$$

if and only if

$$\pi \neq 0 \tag{209}$$

where π is the coefficient in the first stage linear projection:

$$x_i = z_i \pi + r_i$$

for $i = 1, ..., n$.

Proof. First, recall the first stage linear projection:

$$x_i = z_i \pi + r_i.$$

Now, multiply by z_i on both sides:

$$z_i x_i = z_i^2 \pi + z_i r_i.$$

Now take expectations of both sides:

$$\mathbb{E}[z_i x_i] = \pi \mathbb{E}[z_i^2] + \mathbb{E}[z_i r_i] = \pi \mathbb{E}[z_i^2] + 0$$

where we used the fact that $\mathbb{E}[z_i r_i] = 0$ by definition of a linear projection model. This means that when we run a regression, unless we assume endogeneity, the orthogonality conditions hold. Therefore, we have that:

$$\mathbb{E}[z_i x_i] = \pi \mathbb{E}[z_i^2].$$

By assumption, we said that:

$$\mathbb{E}[z_i x_i] \neq 0$$

as z_i is informative. Now, looking on the right-hand side as $\mathbb{E}[z_i] = 0$ by assumption, we have that:

$$Var[z_i] = \mathbb{E}[z_i^2].$$

Furthermore, as $\mathbb{E}[z_i x_i] \neq 0$, z_i **must** have a positive variance since a constant cannot be correlated with a random variable $\mathbb{E}[z_i x_i] \neq 0$, therefore we can conclude that:

$$Var[z_i] = \mathbb{E}[z_i^2] > 0.$$

Therefore, for the expression

$$\mathbb{E}[z_i x_i] = \pi \mathbb{E}[z_i^2]$$

to hold, we therefore require that $\pi \neq 0$ as well since we showed that all the other terms are not equal to zero. □

We can describe the 2SLS estimator. Starting off from our first stage linear projection:

$$x_i = z_i \pi + r_i$$

we can rewrite this in matrix form:

$$\boldsymbol{X} = \boldsymbol{Z}\pi + \boldsymbol{r}$$

where π is still a scalar and $\boldsymbol{Z}$ is a $n \times 1$ vector. Then, we have that the first stage estimator is given by our usual OLS estimator:

$$\hat{\pi} = (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{X}.$$

From this, we can now construct our predicted values of the endogenous variable $\boldsymbol{X} = (x_1, ..., x_n)^T$ where recall that we only have one explanatory variable. Therefore, if you recall that the predicted value in usual OLS is $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, then the predicted values from our first stage of the endogenous variable is:

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\pi} = \boldsymbol{Z}\Big[(\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{X}\Big] = \boldsymbol{Z}(\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{X}.$$

We can add and subtract in our predicted value for $\hat{x}_i$ for the endogenous x_i to get the **second stage of the model**:

$$\begin{aligned}
&y_i = x_i\beta + u_i \\
&\leftrightarrow y_i = x_i\beta + u_i + \hat{x}_i\beta - \hat{x}_i\beta \\
&\leftrightarrow y_i = \hat{x}_i\beta + \Big[u_i - \hat{x}_i\beta + x_i\beta\Big] \\
&\leftrightarrow y_i = \hat{x}_i\beta + \Big[u_i + \big(x_i - \hat{x}_i\big)\beta\Big]
\end{aligned}$$

Therefore we have the expression:

$$y_i = \hat{x}_i\beta + \Big[u_i + \big(x_i - \hat{x}_i\big)\beta\Big].$$

We can then write this in matrix form:

$$\boldsymbol{y} = \hat{\boldsymbol{X}}\beta + (\boldsymbol{u} + (\boldsymbol{X} - \hat{\boldsymbol{X}})\beta)$$

where all vectors are $n \times 1$. We can therefore run this regression to get our two-stage least squares (2SLS) estimator.

Definition 280 (First Representation of 2SLS Estimator)

Let x_i be a single endogenous variable and z_i be an instrumental variable. Let our second stage regression be:

$$\boldsymbol{y} = \hat{\boldsymbol{X}}\beta + (\boldsymbol{u} + (\boldsymbol{X} - \hat{\boldsymbol{X}})\beta). \tag{210}$$

Then, the 2SLS estimator is given by:

$$\hat{\beta}_{2SLS} = (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y}. \tag{211}$$

We can therefore summarise the process for two stage least squares estimation.

Proposition 281 (2SLS Estimation Procedure)

The first stage involves running the regression:

$$\boldsymbol{X} = \boldsymbol{Z}\pi + \boldsymbol{r} \tag{212}$$

and retrieving the first stage estimator:

$$\hat{\pi} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}. \tag{213}$$

Then, construct the predicted value:

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\pi} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}. \tag{214}$$

Finally, run the second stage regression:

$$\boldsymbol{y} = \hat{\boldsymbol{X}}\boldsymbol{\beta} + (\boldsymbol{u} + (\boldsymbol{X} - \hat{\boldsymbol{X}})\boldsymbol{\beta}) \tag{215}$$

and retrieve the 2SLS estimator:

$$\hat{\beta}_{2SLS} = (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y}. \tag{216}$$

Notice that we have another term $(\boldsymbol{X} - \hat{\boldsymbol{X}})\beta$ in the error term. This is in fact a source of finite sample bias.

Proposition 282 (2SLS Estimator is Biased)

The 2SLS estimator is a biased estimator:

$$\mathbb{E}[\hat{\beta}_{2SLS}] \neq \beta. \tag{217}$$

So far, our expression for the 2SLS estimator contains fitted values of $\hat{\boldsymbol{X}}$ and $\boldsymbol{y}$. We can rewrite our expression for the 2SLS estimator in terms of our original variables $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}$.

Proposition 283 (Second Representation of 2SLS Estimator)

The second expression for the 2SLS estimator is:

$$\hat{\beta}_{2SLS} = \left[\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}. \tag{218}$$

Proof. First, notice that in our original expression for the 2SLS estimator:

$$\hat{\beta}_{2SLS} = (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y}$$

we can plug in our expressions for the predicted values:

$$\hat{X} = Z\hat{\pi} = Z(Z^TZ)^{-1}Z^TX.$$

We plug this into our original expression to get:

$$\begin{aligned}
\hat{\beta}_{2SLS} &= (\hat{X}^T\hat{X})^{-1}\hat{X}^Ty && (219)\\
&= \left(\left[Z(Z^TZ)^{-1}Z^TX\right]^T\left[Z(Z^TZ)^{-1}Z^TX\right]\right)^{-1}\left[Z(Z^TZ)^{-1}Z^TX\right]^Ty && (220)\\
&=_{(1)} \left(\left[X^TZ(Z^TZ)^{-1}Z^T\right]\left[Z(Z^TZ)^{-1}Z^TX\right]\right)^{-1}\left[Z(Z^TZ)^{-1}Z^TX\right]^Ty && (221)\\
&= \left(X^TZ(Z^TZ)^{-1}Z^TZ(Z^TZ)^{-1}Z^TX\right)^{-1}\left[Z(Z^TZ)^{-1}Z^TX\right]^Ty && (222)\\
&= \left(X^TZ(Z^TZ)^{-1}I_1Z^TX\right)^{-1}\left[Z(Z^TZ)^{-1}Z^TX\right]^Ty && (223)\\
&= \left(X^TZ(Z^TZ)^{-1}Z^TX\right)^{-1}\left[X^TZ(Z^TZ)^{-1}Z^T\right]y && (224)\\
&= \left(X^TZ(Z^TZ)^{-1}Z^TX\right)^{-1}X^TZ(Z^TZ)^{-1}Z^Ty && (225)\\
& && (226)
\end{aligned}$$

where in $=_{(1)}$ we applied the transpose and used the symmetry of $\left((Z^TZ)^{-1}\right)^T = (Z^TZ)^{-1}$. □

The previous expression is a bit cumbersome to work with. We can again rewrite our 2SLS estimator.

Proposition 284 (Third Representation of 2SLS Estimator)
The third expression for the 2SLS estimator is:

$$\hat{\beta}_{2SLS} = (\hat{X}^TX)^{-1}\hat{X}^Ty. \quad (227)$$

Proof. We take our second expression:

$$\hat{\beta}_{2SLS} = \left[X^TZ(Z^TZ)^{-1}Z^TX\right]^{-1}X^TZ(Z^TZ)^{-1}Z^Ty.$$

We then apply the transpose to the whole expression in front of the y:

$$= \left[(Z(Z^TZ)^{-1}Z^TX)^TX\right]^{-1}\left(Z(Z^TZ)^{-1}Z^TX\right)^Ty$$

Notice that red and blue terms are exactly the expression for the first representation of the 2SLS estimator:

$$\hat{\beta}_{2SLS} = (\hat{X}^T\hat{X})^{-1}\hat{X}^Ty.$$

We therefore plug this in to get our desired expression:

$$\hat{\beta}_{2SLS} = (\hat{X}^TX)^{-1}\hat{X}^Ty.$$

□

Now, we take pause and notice the case for when the 2SLS estimator coincides with the OLS estimator.

Proposition 285 (2SLS and OLS Estimator)
The 2SLS estimator coincides with the OLS estimator:

$$\hat{\beta}_{2SLS} = \hat{\beta}_{OLS} \tag{228}$$

if the instrumental variable used is the explanatory variable itself $z_i = x_i$. This implies that x_i is an exogenous explanatory variable.

Proof. Recall the definition of fitted values for $\hat{x}_i$ and plug in the explanatory variable for the instrumental variable:

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}I_1 = \boldsymbol{X}.$$

Therefore, the predicted value is just the explanatory itself $\hat{\boldsymbol{X}} = \boldsymbol{X}$. □

This is very intuitive since if we project x_i on itself, we obtain perfect predictions of x_i and the second stage of 2SLS coincides with the original model.

10.1.1 Just-Identified System

So far, we have assumed only one valid and informative instrument z_i for a **single** endogenous explanatory variable x_i. In such a scenario, we say that the **scalar parameter** β is **just-identified**. For this subsection, we now extend this to K-dimensional parameter vector.

Assumption 286 (K Endogenous Variables and Parameters)
Let $\boldsymbol{\beta}$ be a $K \times 1$ parameter vector associated to $\boldsymbol{x}_i$ which is a $K \times 1$ vector of explanatory variables.

We now distinguish between the three cases of how many parameters we are estimating and how many instrumental variables we have.

Definition 287 (Just-Identified)
Let $\boldsymbol{z}_i$ contain L valid instruments. The parameter vector $\boldsymbol{\beta}$ is **just-identified** if $L = K$ and $rank(\mathbb{E}[\boldsymbol{z}_i\boldsymbol{x}_i^T]) = K$.

Definition 288 (Over-Identified)
The parameter vector $\boldsymbol{\beta}$ is **over-identified** if $L > K$ and $rank(\mathbb{E}[\boldsymbol{z}_i\boldsymbol{x}_i^T]) = K$.

Definition 289 (Under-Identified)
The parameter vector $\boldsymbol{\beta}$ is **under-identified** if $rank(\mathbb{E}[\boldsymbol{z}_i\boldsymbol{x}_i^T]) < K$. In particular, the order condition $L < K$ implies under-identification.

If we are in the nice case of just-identification, we get a nice expression for the 2SLS estimator.

Proposition 290 (Just-Identified Representation of 2SLS Estimator)
Suppose that the parameter vector $\boldsymbol{\beta}$ is **just-identified**. Then, the expression for the 2SLS estimator simplifies to:

$$\hat{\boldsymbol{\beta}}_{2SLS} = (\boldsymbol{Z}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\boldsymbol{y}. \tag{229}$$

Proof. We show the proof for our previous setting where $K = 1$. Now, note that $\hat{\pi}$ is a scalar in the first stage linear projection as we only have one endogenous variable: $\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\pi}$. Then, we rewrite our third representation of the 2SLS estimator using this result:

$$\begin{aligned}
\hat{\beta}_{2SLS} &= (\hat{\boldsymbol{X}}^T\boldsymbol{X})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y} \\
&= ((\boldsymbol{Z}\hat{\pi})^T\boldsymbol{X})^{-1}(\boldsymbol{Z}\hat{\pi})^T\boldsymbol{y} \\
&= (\hat{\pi}\boldsymbol{Z}^T\boldsymbol{X})^{-1}\hat{\pi}(\boldsymbol{Z}^T\boldsymbol{y}) \\
&=_{(1)} \frac{1}{\hat{\pi}}(\boldsymbol{Z}^T\boldsymbol{X})^{-1}\hat{\pi}(\boldsymbol{Z}^T\boldsymbol{y}) \\
&= (\boldsymbol{Z}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\boldsymbol{y}
\end{aligned}$$

where $=_{(1)}$ makes use of the fact that $\hat{\pi}$ is a scalar. □

10.1.2 L Instruments

So far we looked at the case where we had 1 instrument z_i. We now look at the case where if we had L instruments $\boldsymbol{z}_i = (z_{1i}, z_{2i}, ..., z_{Li})^T$ but we still have one endogenous explanatory variable x_i.

Assumption 291 (L Instruments and 1 Endogenous Variable)
Suppose we had L instrumental variables $\boldsymbol{z}_i = (z_{1i}, z_{2i}, ..., z_{Li})^T$ and one endogenous explanatory variable x_i.

We again have the structural equation:

$$y_i = x_i\beta + u_i \tag{230}$$

for $i = 1, ..., n$. To make life easier for us again, we assume that the variables have zero unconditional mean.

Assumption 292 (Zero Unconditional Mean)
Suppose that we had that zero unconditional mean for the structural error term, endogenous variable, and violation of the orthogonality condition:

$$\mathbb{E}[u_i] = \mathbb{E}[x_i] = 0 \tag{231}$$

$$\mathbb{E}[\boldsymbol{z}_i] = \boldsymbol{0} \tag{232}$$

$$\mathbb{E}[x_i u_i] \neq 0. \tag{233}$$

Now, since we assumed that $\mathbb{E}[x_i] = 0$ and $\mathbb{E}[\boldsymbol{z}_i] = \boldsymbol{0}$, we do not need to include an intercept term in our **first stage linear projection**:

$$x_i = \boldsymbol{z}_i^T\boldsymbol{\pi} + r_i \tag{234}$$

for $i = 1, ..., n$. Here, we have 1 endogenous variable x_i but L instruments $\boldsymbol{z}_i$ and therefore, we have $\boldsymbol{\pi}$ as a $L \times 1$ vector of coefficients on the L instruments. We can then write this in matrix form:

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\pi} + \boldsymbol{r}. \tag{235}$$

Definition 293 (Validity of L Instruments)

Let z_i contain L instrumental variables. Then, z_i is a valid instrument if:

$$\mathbb{E}[z_i u_i] = \mathbf{0} \tag{236}$$

so that **all** the instrumental variables in z_i are valid.

Definition 294 (Informativeness of Instruments)

The z_i instruments are informative for the single explanatory endogenous variable x_i if at least one of the instrumental variables in z_i is informative:

$$\mathbb{E}[z_i x_i] \neq \mathbf{0}. \tag{237}$$

If $\mathbb{E}[z_i z_i^T]$ is non-singular, equivalently, the coefficient vector on the instruments $\boldsymbol{\pi}$ is not equal to the zero vector:

$$\boldsymbol{\pi} \neq \mathbf{0}. \tag{238}$$

Proof. These two definitions are equivalent for informativeness since first, recall the first stage linear projection:

$$x_i = z_i^T \boldsymbol{\pi} + r_i.$$

Now, multiply by z_i on both sides:

$$z_i x_i = z_i z_i^T \boldsymbol{\pi} + z_i r_i.$$

Taking expectations and using the projection orthogonality condition $\mathbb{E}[z_i r_i] = \mathbf{0}$ gives:

$$\mathbb{E}[z_i x_i] = \mathbb{E}[z_i z_i^T]\boldsymbol{\pi}.$$

If $\mathbb{E}[z_i z_i^T]$ is non-singular, it follows that

$$\mathbb{E}[z_i x_i] \neq \mathbf{0} \quad \Longleftrightarrow \quad \boldsymbol{\pi} \neq \mathbf{0}.$$

□

We can restate the property of informativeness in a different manner.

Proposition 295 (Informativeness Rank Condition)

The z_i instruments are informative for the single explanatory endogenous variable x_i if:

$$rank[\boldsymbol{\pi}] = 1 \tag{239}$$

or equivalently,

$$rank[\mathbb{E}[z_i x_i]] = 1. \tag{240}$$

We then get two expressions for the 2SLS estimator with L instruments.

Proposition 296 (2SLS Estimator with L Instruments)
Suppose we had L instruments $\boldsymbol{z}_i$. Then, the 2SLS estimator of β has the form:

$$\hat{\beta}_{2SLS} = \left(\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}\right)^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y} \tag{241}$$

and

$$\hat{\beta}_{2SLS} = \left(\hat{\boldsymbol{X}}^T\boldsymbol{X}\right)^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y}. \tag{242}$$

Proof. Plug in the expression that:

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}$$

and

$$\hat{\boldsymbol{X}}^T = \boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T.$$

Then, follow the steps going from the first representation of the 2SLS estimator to the second representation and then to the third representation. □

10.2 Control Function

We now give another on how we can compute the 2SLS estimator through **control functions**. We make an assumption.

Assumption 297 (L Instruments and 1 Endogenous Variable)
Suppose we had L instrumental variables $\boldsymbol{z}_i = (z_{1i}, z_{2i}, ..., z_{Li})^T$ and one endogenous explanatory variable x_i.

We again have the structural equation:

$$y_i = x_i\beta + u_i \tag{243}$$

for $i = 1, ..., n$ where $\mathbb{E}[u_i] = \mathbb{E}[x_i] = 0$, $\mathbb{E}[\boldsymbol{z}_i] = \boldsymbol{0}$, and the orthogonality condition violated $\mathbb{E}[x_i u_i] \neq 0$.

Now, since we assumed that $\mathbb{E}[x_i] = 0$ and $\mathbb{E}[\boldsymbol{z}_i] = \boldsymbol{0}$, we do not need to include an intercept term in our **first stage linear projection**:

$$x_i = \boldsymbol{z}_i^T\boldsymbol{\pi} + r_i \tag{244}$$

for $i = 1, ..., n$. Here, we have 1 endogenous variable x_i but L instruments $\boldsymbol{z}_i$ and therefore, we have $\boldsymbol{\pi}$ as a $L \times 1$ vector of coefficients on the L instruments. We can then write this in matrix form:

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\pi} + \boldsymbol{r}. \tag{245}$$

To make life easier for us again, we assume that the variables have zero unconditional mean.

Assumption 298 (Zero Unconditional Mean)
Suppose that we had that zero unconditional mean for the structural error term, endogenous variable, and violation of the orthogonality condition:

$$\mathbb{E}[u_i] = \mathbb{E}[x_i] = 0 \tag{246}$$

$$\mathbb{E}[\boldsymbol{z}_i] = \boldsymbol{0} \tag{247}$$

$$\mathbb{E}[x_i u_i] \neq 0. \tag{248}$$

Now, define the OLS residuals from the first stage linear projection equation as $\hat{r}_i$.

$$\hat{r}_i \equiv x_i - \boldsymbol{z}_i^T \hat{\boldsymbol{\pi}}.$$

Definition 299 (First Stage Linear Projection Residuals)
The OLS residuals from the first stage linear projection equation is defined as:

$$\hat{r}_i \equiv x_i - \boldsymbol{z}_i^T \hat{\boldsymbol{\pi}}. \tag{249}$$

With this, we can define what is the control function approach for 2SLS.

Definition 300 (2SLS Using Control Functions)
Let $\hat{r}_i$ be the OLS residuals from the first stage linear projection equation. Then, the 2SLS estimator $\hat{\beta}_{2SLS}$ can be estimated as the OLS estimator of the extended equation:

$$y_i = x_i\beta + \hat{r}_i\gamma + v_i \tag{250}$$

for $i = 1, ..., n$. The first stage residuals $\hat{r}_i$ are referred to the **control function**.

First, notice we are now regressing on x_i and not $\hat{x}_i$! The inclusion of the control function allows us to control for correlation between the endogenous x_i and the error term u_i in the original model. The reason it does this is because we can decompose the endogenous variable:

$$x_i = \boldsymbol{z}_i^T \boldsymbol{\pi} + r_i$$

where we assume $\boldsymbol{z}_i$ is uncorrelated with the structural error u_i by definition of an instrumental variable. As $\boldsymbol{z}_i$ is uncorrelated with u_i, then, the only component of x_i that could be correlated with the structural error u_i is therefore r_i. Therefore, by including r_i into our equation, we now control for this correlation between x_i and the structural error u_i. This control function approach gives us a consistent estimator β, which we shall discuss in the next section.

10.3 Indirect Least Squares

For the next method, we need to assume we are in a just-identified setting.

Assumption 301 (Just-Identified)
The parameter vector $\boldsymbol{\beta}$ is **just-identified** if $L = K$ and $rank(\mathbb{E}[\boldsymbol{z}_i\boldsymbol{x}_i^T]) = K$.

Suppose that we had the number of parameters is 1, $K = 1$, which implies that we have one endogenous explanatory variable x_i. By just-identified system, this means that the number of instrumental variables is 1, $L = 1$. Now, take our first stage linear projection for x_i:

$$x_i = z_i\pi + r_i$$

and we can then substitute this into our structural equation:

$$y_i = x_i\beta + u_i$$

to get the reduced form equation:

$$y_i = z_i(\pi\beta) + (u_i + r_i\beta).$$

Definition 302 (Reduced Form Equation)
The reduced form equation in indirect least squares is given by:

$$y_i = z_i(\pi\beta) + (u_i + r_i\beta) \tag{251}$$

where number of parameters equal the number of instrumental variables.

In the general case with L instruments and K regressors, the first-stage coefficient matrix $\boldsymbol{\Pi}$ is $L \times K$ and the reduced-form coefficient $\boldsymbol{\Pi\beta}$ is $L \times 1$. Thus, this multiplication is well defined whether or not $L = K$. Just-identification is required to recover $\boldsymbol{\beta}$ uniquely by indirect least squares, not for the reduced form to exist.

We can state the indirect least squares method approach.

Definition 303 (Indirect Least Squares)
Let

$$x_i = z_i\pi + r_i \tag{252}$$

be our first stage linear projection equation. Let $\hat{\pi}$ be our OLS estimator in this first stage linear projection equation. Then, let

$$y_i = z_i(\pi\beta) + (u_i + r_i\beta) \tag{253}$$

be the reduced form equation with $\widehat{\pi\beta}$ as the OLS estimator in this reduced form equation. Then, the **indirect least squares estimator** $\hat{\beta}_{ID}$ is given by:

$$\hat{\beta}_{ID} = \frac{\widehat{\pi\beta}}{\hat{\pi}}. \tag{254}$$

This estimator has some nice properties.

Proposition 304 (Consistency of Indirect Least Squares Estimator)
The indirect least squares estimator $\hat{\beta}_{ID}$ is a consistent estimator.

We can also show that this estimator is equivalent to the 2SLS estimator in a just-identified setting.

Theorem 305 (Indirect Least Squares and 2SLS Estimator)
Suppose that we have a just-identified model. Then, the indirect least squares estimator coincides with the 2SLS estimator.

10.4 Consistency of 2SLS Estimator

First, we recall that in finite samples, the 2SLS estimator is a biased estimator.

Theorem 306 (Bias of 2SLS Estimator)
The 2SLS estimator

$$\hat{\boldsymbol{\beta}}_{2SLS} = (\hat{\boldsymbol{X}}^T\boldsymbol{X})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y} \tag{255}$$

is a **biased estimator** of $\boldsymbol{\beta}$.

Whilst we may not have the finite sample property of an unbiased estimator, we at least have a nice asymptotic property for the 2SLS estimator.

Theorem 307 (Consistency of 2SLS Estimator)
Assume that (y_i, x_i, z_i) are independent and identically distributed. Furthermore, suppose that the instrument z_i is:

1. Valid $\mathbb{E}[z_i u_i] = 0$
2. Informative $\mathbb{E}[z_i x_i] \neq 0$ or $\pi \neq 0$.

Then the 2SLS estimator:

$$\hat{\beta}_{2SLS} = (\hat{\boldsymbol{X}}^T \boldsymbol{X})^{-1} \hat{\boldsymbol{X}}^T \boldsymbol{y} \tag{256}$$

is a **consistent** estimator of β.

Proof. First, write the 2SLS estimator as:

$$\begin{aligned}
\hat{\beta}_{2SLS} &= (\hat{\boldsymbol{X}}^T \boldsymbol{X})^{-1} \hat{\boldsymbol{X}}^T \boldsymbol{y} \\
&= (\hat{\boldsymbol{X}}^T \boldsymbol{X})^{-1} \hat{\boldsymbol{X}}^T (\boldsymbol{X}\beta + \boldsymbol{u}) \\
&= (\hat{\boldsymbol{X}}^T \boldsymbol{X})^{-1} \hat{\boldsymbol{X}}^T \boldsymbol{X}\beta + (\hat{\boldsymbol{X}}^T \boldsymbol{X})^{-1} \hat{\boldsymbol{X}}^T \boldsymbol{u} \\
&= \beta + (\hat{\boldsymbol{X}}^T \boldsymbol{X})^{-1} \hat{\boldsymbol{X}}^T \boldsymbol{u}.
\end{aligned}$$

We then multiply and divide by n:

$$\begin{aligned}
\hat{\beta}_{2SLS} &= \beta + (\hat{\boldsymbol{X}}^T \boldsymbol{X})^{-1} \hat{\boldsymbol{X}}^T \boldsymbol{u} \\
&= \beta + n(\hat{\boldsymbol{X}}^T \boldsymbol{X})^{-1} \left(\frac{1}{n}\right) \hat{\boldsymbol{X}}^T \boldsymbol{u} \\
&= \beta + \underbrace{\left(\frac{\hat{\boldsymbol{X}}^T \boldsymbol{X}}{n}\right)^{-1}}_{(*)} \underbrace{\left(\frac{\hat{\boldsymbol{X}}^T \boldsymbol{u}}{n}\right)}_{(**)}
\end{aligned}$$

We first look at (*).

Claim 308.

$$p \lim_{n\to\infty} \left(\frac{\hat{\boldsymbol{X}}^T \boldsymbol{X}}{n}\right)^{-1} = \left(\pi \mathbb{E}[z_i x_i]\right)^{-1} \neq 0.$$

Proof. We have that (x_i, z_i) is i.i.d. We can then use the expression $\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\pi}$ and plug this in to get:

$$p \lim_{n\to\infty} \left(\frac{\hat{\boldsymbol{X}}^T \boldsymbol{X}}{n}\right) = p \lim_{n\to\infty} \left[\hat{\pi} \frac{\boldsymbol{Z}^T \boldsymbol{X}}{n}\right] = \pi \mathbb{E}[z_i x_i].$$

Here, $\hat{\pi}$ is a **consistent estimator** of π by the properties of the OLS estimator in the first-stage equation. Furthermore, as z_i is informative for x_i, we have that $\mathbb{E}[z_i x_i] \neq 0$.

We can then apply the **continuous mapping theorem** to conclude that:

$$p \lim_{n\to\infty} \left(\frac{\hat{\boldsymbol{X}}^T \boldsymbol{X}}{n}\right)^{-1} = \left(\pi \mathbb{E}[z_i x_i]\right)^{-1} \neq 0.$$

□

We now look at (**).

Claim 309.

$$p\lim_{n\to\infty}\left(\frac{\hat{\boldsymbol{X}}^T\boldsymbol{u}}{n}\right)=0.$$

Proof. We know that $\boldsymbol{x}_i$ is i.i.d for all $i=1,\ldots,n$. Furthermore, $\mathbb{E}[z_iu_i]=0$ for all $i=1,\ldots,n$ by validity of the instrumental variable. We can therefore apply the **weak law of large numbers** to get:

$$p\lim_{n\to\infty}\left(\frac{\hat{\boldsymbol{X}}^T\boldsymbol{u}}{n}\right)=p\lim_{n\to\infty}\left(\hat{\pi}\frac{\boldsymbol{Z}^T\boldsymbol{u}}{n}\right)=\pi\mathbb{E}[z_iu_i]=_{(1)}\pi\cdot 0=0$$

whereby $=_{(1)}$ arises from applying Slutsky's theorem using both the validity assumption of the instrumental variable $\boldsymbol{Z}^T$ and consistency of $\hat{\pi}\xrightarrow{P}\pi$. □

Therefore, we can apply the **continuous mapping theorem** to the product of (*) and (**) that we started with, whereby we now have that the **product of the limits is the limit of the products**:

$$p\lim_{n\to\infty}\left[\beta+\underbrace{\left(\frac{\hat{\boldsymbol{X}}^T\boldsymbol{X}}{n}\right)^{-1}}_{(*)}\underbrace{\left(\frac{\hat{\boldsymbol{X}}^T\boldsymbol{u}}{n}\right)}_{(**)}\right]=\beta+\left(\pi\mathbb{E}[z_ix_i]\right)^{-1}\cdot 0=\beta$$

We can therefore conclude that:

$$p\lim_{n\to\infty}\hat{\beta}_{2SLS}=\beta$$

Therefore, we have shown consistency.

□

Looking at the proof for consistency, we can see that we require our instrumental variables to be both valid and informative in order for our 2SLS estimator $\hat{\beta}_{2SLS}$ to be consistent.

10.5 Asymptotics

We get some nice asymptotic results for 2SLS estimators.

Theorem 310 (Asymptotic Normality of 2SLS Estimators)
Let $\{y_i,x_i,z_i\}$ be i.i.d observations. Suppose that the instruments z_i are both valid and informative. Then,

$$\sqrt{n}(\hat{\beta}_{2SLS}-\beta)\xrightarrow{D}\mathcal{N}(0,V) \tag{257}$$

where V is a variance-covariance matrix.

We can do hypothesis testing using chi-squared Wald tests.

Proposition 311 (Variance Estimator Under Conditional Homoskedasticity)
Under conditional homoskedasticity $\mathbb{E}[u_i^2|z_i] = \sigma^2$ for all $i = 1, 2, ..., n$, then a consistent estimator of the limit variance matrix V is given by:

$$n\hat{\sigma}^2(\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1} \tag{258}$$

where σ^2 can be estimated consistently using the residuals from the 2SLS regression:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 \tag{259}$$

where $\hat{u}_i \equiv y_i - x_i^T\hat{\beta}_{2SLS}$.

This implies that the variance of the 2SLS estimator is:

$$\hat{Var}[\hat{\beta}_{2SLS}] = \frac{\hat{V}}{n} = \hat{\sigma}^2(\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}. \tag{260}$$

As usual, we can obtain White standard errors for the 2SLS estimator if we are concerned about conditional heteroskedasticity.

Proposition 312 (Variance Estimator Under Conditional Heteroskedasticity)
Under conditional heteroskedasticity $\mathbb{E}[u_i^2|z_i] \neq \sigma^2$ for all $i = 1, 2, ..., n$, then the variance of $\hat{\beta}_{2SLS}$ is given by:

$$\hat{Var}(\hat{\beta}_{2SLS}) = \frac{\hat{V}}{n} = (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\left(\sum_{i=1}^{n}\hat{u}_i^2\hat{\boldsymbol{x}}_i\hat{\boldsymbol{x}}_i^T\right)(\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1} \tag{261}$$

where

$$\hat{u}_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{2SLS}. \tag{262}$$

10.6 IV with Multiple Regressors

We now generalise where we have multiple explanatory variables that are endogenous. Suppose we had the parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_K)$. Let $\boldsymbol{x}_i^T = (1, x_{2i}, ..., x_{Ki})$ be a $1 \times K$ row vector of explanatory variables such that:

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i \tag{263}$$

where the orthogonality condition no longer holds $\mathbb{E}[\boldsymbol{x}_i u_i] \neq \boldsymbol{0}$. Notice that we now have an intercept term in $\boldsymbol{x}_i$. Furthermore, assume that $\boldsymbol{z}_i^T$ is a $1 \times L$ row vector containing the intercept, any exogenous explanatory variables contained in $\boldsymbol{x}_i$, and any additional **excluded instruments** that are valid so that $\mathbb{E}[\boldsymbol{z}_i u_i] = 0$.

Let M be the number of excluded instruments, let K_{x_1} be the number of included exogenous explanatory variables (including the intercept), and let L be the total number of instruments, so that:

$$L = K_{x_1} + M.$$

We now state the first necessary condition we need for identification.

Definition 313 (Order Condition)
The order condition is that there is at least as many instrumental variables L as there are number of parameters K we are trying to estimate:

$$L \geq K. \tag{264}$$

Recall that instrumental variables here $\boldsymbol{z}_i$ contains both exogenous explanatory variables and outside instrumental variables.

We now state the necessary and sufficient condition for identification of $\boldsymbol{\beta}$.

Definition 314 (Rank Condition)
The **rank condition** is that the $L \times K$ matrix is full rank:

$$rank(\mathbb{E}[\boldsymbol{z}_i \boldsymbol{x}_i^T]) = K. \tag{265}$$

Remark 315. *Recall that by linear algebra,* $rank(\mathbb{E}[\boldsymbol{z}_i \boldsymbol{x}_i^T]) \leq \min\{K, L\}$. *Since* $L \geq K$ *by the order condition, the rank condition requires this rank to equal* K.

The rank condition ensures that the instruments provide K linearly independent sources of variation for the explanatory variables and hence identify $\boldsymbol{\beta}$. The absence of perfect multicollinearity among the explanatory variables is a separate requirement.

Let us define the first stage projection equation.

Definition 316 (First Stage Projection)
The first stage projection equation is:

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{R} \tag{266}$$

where $\boldsymbol{X}$ is a $n \times K$ matrix, $\boldsymbol{Z}$ is a $n \times L$ matrix, $\boldsymbol{\Pi}$ is a $L \times K$ matrix and $\boldsymbol{R}$ is a $n \times K$ matrix.

Here, in this first stage projection equation, we have K equations in which we are projecting. We can now state an equivalent condition for the rank condition.

Definition 317 (Equivalent Rank Condition)
If $\mathbb{E}[\boldsymbol{z}_i \boldsymbol{z}_i^T]$ is non-singular, an equivalent rank condition is:

$$rank(\boldsymbol{\Pi}) = K. \tag{267}$$

A few final notes on this section. The parameter vector $\boldsymbol{\beta}$ is **not identified** if we have one endogenous explanatory variable for which we have no informative excluded instrument. Furthermore, the rank condition also fails if we have one informative instrument for two or more endogenous variables.

We have seen that the rank condition was needed to check for the informativeness of instrumental variables.

10.7 Testing

We are interested in three types of tests for 2SLS estimators. First, some of the instruments are informative. Second, all the instruments valid. Finally, are the explanatory variables exogenous.

10.7.1 Testing for Informativeness of Instrumental Variables

We now want to test whether are our instrumental variables informative for the explanatory variable.

Definition 318 (Testing for Informativeness of Instruments)
Let x_i be a single endogenous variable, let $\boldsymbol{q}_i$ contain the included exogenous regressors, and let $\boldsymbol{z}_i$ contain the excluded instruments. Estimate the first-stage linear projection:

$$x_i = \boldsymbol{q}_i^T \boldsymbol{\delta} + \boldsymbol{z}_i^T \boldsymbol{\pi} + r_i \tag{268}$$

and run the joint F-test:

$$H_0 : \pi_1 = \pi_2 = ... = \pi_M = 0.$$

With several endogenous variables, repeat this test for each first-stage equation.

A rule of thumb suggested by Staiger and Stock (1997) and often used in empirical work is that the value of the **F-test statistic should be at least 10** for reliance on the standard asymptotic approximation to the distribution of the 2SLS estimator. For this linear model with a single endogenous explanatory variable and conditional homoskedasticity, Stock and Yogo (2005) provide alternative critical values for this F-test which seek to control either the bias in the 2SLS estimator relative to the bias of the OLS estimator in the same model or the rejection frequency of standard asymptotic t-tests based on the 2SLS estimator.

10.7.2 Test for Validity of Instrumental Variables

The Sargan/Hansen test is a test for validity of instrumental variables.

Assumption 319 (Sargan Test Assumption)
We have a linear model with conditional **homoskedasticity**. Furthermore, we have an over-identified model whereby $L > K$.

Definition 320 (Null Hypothesis for Validity of Instruments)
The null hypothesis for the testing of the validity of instruments $\boldsymbol{z}_i$ is given by:

$$H_0 : \mathbb{E}[\boldsymbol{z}_i u_i] = \boldsymbol{0}. \tag{269}$$

We can now state the Sargan test.

Definition 321 (Sargan Test for Validity)
Under the null hypothesis $H_0 : \mathbb{E}[\boldsymbol{z}_i u_i] = \boldsymbol{0}$, the **Sargan test statistic** has the form:

$$J = \left(\frac{1}{\hat{\sigma}^2}\right) \hat{\boldsymbol{u}}^T \boldsymbol{Z} (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \hat{\boldsymbol{u}} \xrightarrow{D} \chi^2(L - K) \tag{270}$$

whereby $\hat{u}_i = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{2SLS}$ and $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2 = \mathbb{E}[u_i^2 | z_i]$.

The intuition behind this test is that the scalar test statistic J is proportional to what we minimise to define the 2SLS estimator via the GMM estimator with $(\boldsymbol{Z}^T \boldsymbol{Z})^{-1}$ as the choice for the weight matrix. We try to set the criterion

function of the GMM estimator as close to 0 as possible. A large minimised value of J means that the sample moments cannot be made jointly close to zero, which provides evidence against the overidentifying restrictions.

If we have **conditional heteroskedasticity**, then we should use the **Hansen** test.

10.7.3 Test for Endogeneity

We now want to construct our second test to check for endogeneity of our regressors.

Definition 322 (Null Hypothesis for Endogeneity of Regressors)
The null hypothesis for the testing of the endogeneity of regressors $\boldsymbol{x}_i$ is given by:

$$H_0 : \mathbb{E}[\boldsymbol{x}_i u_i] = \mathbf{0}. \tag{271}$$

Recall that we used instrumental variables because we were worried that our OLS estimators were inconsistent and that by using appropriate instruments, the 2SLS estimator will become consistent. If the instrumental variables were indeed valid and informative, then we would expect the 2SLS estimator to be different to the OLS estimator if one of the explanatory variables were endogenous since the OLS estimator will no longer be consistent. We can therefore compare these two sets of parameter estimates as the basis for the Hausman test.

Under the null hypothesis where we have no endogeneity, then both $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{2SLS}$ are consistent estimators. If there was endogeneity, then only $\hat{\boldsymbol{\beta}}_{2SLS}$ will be a consistent estimator. However, under the null, $\hat{\boldsymbol{\beta}}_{2SLS}$ will be less efficient that $\hat{\boldsymbol{\beta}}_{OLS}$.

Definition 323 (Hausman Test for Endogeneity)
Under the null hypothesis $H_0 : \mathbb{E}[\boldsymbol{x}_i u_i] = \mathbf{0}$, the **Hausman test statistic** has the form:

$$h = (\hat{\boldsymbol{\beta}}_{2SLS} - \hat{\boldsymbol{\beta}}_{OLS})^T \left[\widehat{Var}[\hat{\boldsymbol{\beta}}_{2SLS}] - \widehat{Var}[\hat{\boldsymbol{\beta}}_{OLS}]\right]^{-1} (\hat{\boldsymbol{\beta}}_{2SLS} - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{D} \chi^2(K) \tag{272}$$

provided we have **conditional homoskedasticity** $\mathbb{E}[u_i^2|z_i] = \sigma^2$ and the $K \times K$ variance $\left[\widehat{Var}[\hat{\boldsymbol{\beta}}_{2SLS}] - \widehat{Var}[\hat{\boldsymbol{\beta}}_{OLS}]\right]$ is non-singular.

We can construct an alternative form of the Hausman test.

Definition 324 (Scalar Version of Hausman Test for Endogeneity)
The scalar version of the Hausman test has two forms:

$$h = \frac{(\hat{\beta}_{k,2SLS} - \hat{\beta}_{k,OLS})^2}{[\widehat{Var}[\hat{\beta}_{k,2SLS}] - \widehat{Var}[\hat{\beta}_{k,OLS}]]} \xrightarrow{D} \chi^2(1) \tag{273}$$

and

$$t = \frac{(\hat{\beta}_{k,2SLS} - \hat{\beta}_{k,OLS})}{\sqrt{[\widehat{Var}[\hat{\beta}_{k,2SLS}] - \widehat{Var}[\hat{\beta}_{k,OLS}]]}} \sim \mathcal{N}(0,1), \qquad h = t^2 \tag{274}$$

provided that we have **conditional homoskedasticity** $\mathbb{E}[u_i^2|z_i] = \sigma^2$.

10.8 Finite Sample Problems

We look at two sources of finite sample bias.

10.8.1 Overfitting

If we use too many instrumental variables z_i relative to the number of endogenous variables $\boldsymbol{x}_i$, then when we construct our predicted values $\hat{\boldsymbol{x}}_i$, we may end up having a perfect fit so that $\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i$ and therefore we get back our endogenous variable. From this, we did not remove the endogeneity in the system and therefore end up with endogeneity again.

Proposition 325 (Overfitting)
As the number of L instruments approach the sample size n, then the 2SLS estimator $\hat{\boldsymbol{\beta}}_{2SLS}$ tend towards the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$.

In such a situation, $\hat{\boldsymbol{\beta}}_{2SLS}$ will have exactly the same bias as $\hat{\boldsymbol{\beta}}_{OLS}$ in the limit as the number of instruments converges to the sample size $L \to n$. Therefore, if $\hat{\boldsymbol{\beta}}_{2SLS}$ and $\hat{\boldsymbol{\beta}}_{OLS}$ are quite similar, we should check that we are not using too many instruments.

10.8.2 Weak Instrumental Variables

Weak instruments is when the coefficients on the instruments in the first stage equation is barely significant so that the instrument only weakly identifies the endogenous variable $\boldsymbol{x}_i$. This means that $\boldsymbol{\pi} \approx \mathbf{0}$. In such a situation, we get a finite sample bias in the 2SLS estimator $\hat{\boldsymbol{\beta}}_{2SLS}$ in the direction of the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$, which is also biased.

Under conventional fixed-parameter asymptotics, any fixed non-zero relevance is sufficient for consistency, but weak relevance can produce severe finite-sample bias and poor normal approximations. Under local-to-zero weak-instrument asymptotics, the limiting distribution of the 2SLS estimator is non-standard and the estimator can be far away from the true parameter $\boldsymbol{\beta}$.

As a result, we should avoid weak instruments whenever possible and should not attach much significance to 2SLS estimates in situations where the instruments are very weak. In these situations, conventional t-tests and Wald tests become quite unreliable.

10.8.3 Anderson-Rubin Test

We now give a test for coefficient hypotheses that remains valid under weak instruments. Suppose that we have the linear model with an intercept x_{1i} and a single endogenous explanatory variable x_{Ki}:

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_{K-1} x_{K-1,i} + \beta_K x_{Ki} + u_i \tag{275}$$

with $\mathbb{E}[x_{ki} u_i] = 0$ for all $k = 1, 2, ..., K-1$.

Furthermore, suppose that we have M valid excluded instruments $z_{1i}, z_{2i}, ..., z_{Mi}$ with $\mathbb{E}[z_{mi} u_i] = 0$ for $m = 1, 2, ..., M$.

We require one of these outside instruments to be informative for the explanatory variable x_{Ki} in the first stage linear projection whereby at least one coefficient on the instruments θ_j is non-zero:

$$x_{Ki} = \delta_1 + \delta_2 x_{2i} + ... + \delta_{K-1} x_{K-1,i} + \theta_1 z_{1i} + ... + \theta_M z_{Mi} + r_i. \tag{276}$$

We can run an F-test of $H_0 : \theta_1 = \theta_2 = ... = \theta_M = 0$ with degrees of freedom $F(M, n-L)$. Failure to reject this null indicates a lack of first-stage relevance and possible under-identification, whereas rejection supports relevance. A commonly used rule of thumb suggested by Staiger and Stock is that the first-stage F-statistic should be at least 10.

We may also be interested in testing coefficients in the presence of weak instruments.

Definition 326 (Anderson Rubin Test)

Suppose we had the structural equation:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \tag{277}$$

with one weak instrumental variable z_{1i} available. Suppose we wanted to test $H_0 : \beta_1 = 0$. The Anderson Rubin test says to write the reduced form equation:

$$y_i = (\beta_1 \theta) z_{1i} + (\beta_2 + \beta_1 \delta) x_{2i} + (u_i + \beta_1 r_i) \tag{278}$$

and test the null hypothesis $H_0 : \beta_1 \theta = 0$.

11 Generalised Method of Moments

11.1 Motivating Examples

We give two examples of deriving GMM estimators for linear regression and linear regression with instrumental variables before stating the general framework for GMM.

11.1.1 GMM Applied to Linear Regression

Suppose we have the population model:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i \quad \text{for } i = 1, ..., n$$

whereby y_i is a scalar, $\boldsymbol{x}_i$ is a $K \times 1$ vector of explanatory variable and $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters. Furthermore, it is important to stress that $\boldsymbol{\beta}$ here is the true value of the coefficients of the data generating process for the linear model. We impose the following orthogonality assumption, which we will soon see later is necessary for identification.

Assumption 327 (Orthogonality Condition)
The orthogonality condition is that the explanatory variables are uncorrelated with the true population error terms:

$$\mathbb{E}[\boldsymbol{x}_i u_i] = 0 \quad \text{for } i = 1, ..., n$$

Now, the ordinary least squares estimator is the argument to the solution to the following minimisation program:

$$\hat{\boldsymbol{\beta}}_{OLS} \equiv \arg\min_{\tilde{\boldsymbol{\beta}}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}})^2$$

where we stress that $\tilde{\boldsymbol{\beta}}$ refers to any particular value that the estimator can take, which is in contrast to the fixed true population parameter value $\boldsymbol{\beta}$.

From our minimisation program, we can solve for the OLS estimator simply by taking first order conditions with respect to $\tilde{\boldsymbol{\beta}}$. We end up with the first order condition:

$$\nabla_{\tilde{\boldsymbol{\beta}}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}})^2 = -2 \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}) = 0$$

where $\nabla_{\tilde{\boldsymbol{\beta}}}$ is the $K \times 1$ gradient vector with partial derivatives with respect to $\tilde{\boldsymbol{\beta}}$.

Divide by 2 on both sides and multiply by $1/n$ gives us:

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}) = 0$$

What we can then do is to define a function of the **data and parameters** as:

$$g(w_i; \tilde{\boldsymbol{\beta}}) \equiv \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}) \quad \text{where } w_i = (y_i, \boldsymbol{x}_i)$$

where w_i is the observed variables. We use this notation w_i as this can be for any general variables we may observe (later on, w_i will include instrumental variables). We can therefore write our first order condition as the **sample average of a known function of data and parameters**:

$$\frac{1}{n}\sum_{i=1}^{n} g(w_i, \tilde{\boldsymbol{\beta}}) = 0$$

This sample average in fact as a special name in GMM estimation.

Definition 328 (Sample Moment Condition in Linear Regression)
The **sample moment condition** in linear regression is given by:

$$\frac{1}{n}\sum_{i=1}^{n} g(w_i, \tilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) = 0. \tag{279}$$

Remark 329. *The sample moment condition is the equation that defines what our estimator shall be.*

As we have a sample moment condition, it makes sense to have a population analogue whereby we replace the sample average with the expectations operator $\mathbb{E}$. This also has an important name too.

Definition 330 (Population Moment Condition in Linear Regression)
The **population moment condition** in linear regression is given by:

$$\mathbb{E}[g(w_i; \tilde{\boldsymbol{\beta}})] = \mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})] = 0. \tag{280}$$

We now collect and summarise everything we have done so far before proceeding further. We have assumed that we had a linear model for the data generating process and an orthogonality assumption. With this linear model, we can then construct a population moment condition that needs to hold. We can then construct a corresponding sample moment condition by replacing expectations with averages. Now, to find an estimator of the true population parameter, we can instead find an estimator such that the **sample moment condition holds**!

Proposition 331 (Solution to Population Moment Condition)
The population moment condition is solved:

$$\mathbb{E}[g(w_i; \tilde{\boldsymbol{\beta}})] = \mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})] = 0$$

when the estimator $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

Proof. Start off with the population moment:

$$\mathbb{E}[g(w_i; \tilde{\boldsymbol{\beta}})] = \mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})]$$

Let $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ from our proposition:

$$\mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})] = \mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})]$$

We now want to show that this is equal to zero. Recall that from our linear model specification: $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i$. Plugging this in gives us:

$$\mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})] = \mathbb{E}[\boldsymbol{x}_i u_i] =_{(1)} 0$$

where $=_{(1)}$ holds by our orthogonality assumption. □

We have shown that $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ is **a solution** to the population moment condition. We want to know if this solution is also **unique**?

Proposition 332 (Unique Solution to Population Moment Condition)
The estimator $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ is the **unique** solution to the population moment condition:

$$\mathbb{E}[g(w_i; \tilde{\boldsymbol{\beta}})] = \mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}})] = 0$$

if there is no perfect multicollinearity, so that $rank(\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T]) = K$.

Proof. Start off with our population moment condition and plug in our linear model for y_i:

$$\begin{aligned}
\mathbb{E}[g(w_i; \tilde{\boldsymbol{\beta}})] &= \mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}})] \\
&= \mathbb{E}[\boldsymbol{x}_i(\boldsymbol{x}_i^T \boldsymbol{\beta} + u_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}})] \\
&= \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{x}_i u_i - \boldsymbol{x}_i \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}] \\
&= \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \boldsymbol{x}_i u_i] \\
&= \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] + \mathbb{E}[\boldsymbol{x}_i u_i] \\
&=_{(1)} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] + 0 \\
&= \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})]
\end{aligned}$$

where $=_{(1)}$ invokes the orthogonality assumption again. Therefore, we want to know for what values $\tilde{\boldsymbol{\beta}}$ does the following expression hold:

$$\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] = 0.$$

We claim that the **only** solution to this is when $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$. Now, we need an important result from linear algebra.

Theorem 333 (Product of Matrix and Vector is Zero)
Suppose that $\boldsymbol{A}$ is a $M \times N$ matrix. Suppose that $\boldsymbol{A}$ has $rank(\boldsymbol{A}) = L < N$. Then, there are $N - L$ **linearly independent vectors** $\boldsymbol{b} \in \mathbb{R}^N \setminus \{\boldsymbol{0}\}$ such that:

$$\boldsymbol{A}\boldsymbol{b} = \boldsymbol{0}.$$

Remark 334. *This says that if a matrix $\boldsymbol{A}$ does not have full* **column rank** *N and instead had a rank of L, then exists $N - L$ vectors $\boldsymbol{b}$ such that the product of $\boldsymbol{A}\boldsymbol{b}$ is the zero vector $\boldsymbol{0}$.*

This theorem actually has an important corollary which is what we actually need.

Corollary 335
Suppose that $\boldsymbol{A}$ is a $K \times K$ matrix. Suppose that $rank(\boldsymbol{A}) = K$ is full rank. Then, the only vector $\boldsymbol{b} \in \mathbb{R}^K$ such that:

$$\boldsymbol{A}\boldsymbol{b} = \boldsymbol{0}$$

is when $\boldsymbol{b} = \boldsymbol{0}$ is the zero vector.

We now go back to our expression:

$$\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T](\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$$

and we can map this to our corollary by letting $\boldsymbol{A} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T]$ be our $K \times K$ matrix and $\boldsymbol{b} = \boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}$ be our $K \times 1$ vector. We assumed that $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T]$ is full rank. Therefore, by our corollary, the only way for

$$\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] = 0$$

to hold is if the vector multiplying $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T]$ is the zero vector $(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) = \boldsymbol{0}$ and we only get a zero vector if:

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}.$$

□

We have just shown the linear regression model can be solved by GMM! The GMM estimator takes our model in terms of its population moment conditions, and solves them by choosing an estimator to set their sample analogue equal to 0.

11.1.2 GMM Applied to Linear Regression with Instrumental Variables

We now derive the GMM framework in the context of linear regression with instrumental variables.

We would like to estimate the parameter $\boldsymbol{\beta}$ in the model:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i \quad \text{for } i = 1, \ldots, n. \tag{281}$$

where $\boldsymbol{x}_i, \boldsymbol{\beta}$ are $K \times 1$ vectors. However, **some** of the explanatory variables $\boldsymbol{x}_i$ may be **endogenous** $\mathbb{E}[\boldsymbol{x}_i u_i] \neq \boldsymbol{0}$. Therefore, we will need to use **instrumental variables**.

Assumption 336 (Instrumental Variable Orthogonality)
An instrumental variable $\boldsymbol{z}_i$ is one where:

$$\mathbb{E}[\boldsymbol{z}_i u_i] = 0. \tag{282}$$

Definition 337 (Instrumental Variable)
A variable $\boldsymbol{z}_i$ is a **valid candidate instrument** if it satisfies the **instrumental variable orthogonality** assumption. A usable set of instruments must additionally satisfy the relevance or rank condition.

Therefore, we have our explanatory variables $\boldsymbol{x}_i$ and our instrumental variables $\boldsymbol{z}_i$ whereby:

$$\boldsymbol{x}_i \in \mathbb{R}^K \quad \text{and} \quad \boldsymbol{z}_i \in \mathbb{R}^L.$$

First, we note that both $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ will have at least an intercept term common in both variables. Furthermore, any exogenous variables in $\boldsymbol{x}_i$ will also be included into $\boldsymbol{z}_i$. Finally, $\boldsymbol{z}_i$ will have additional outside instrumental variables. We elaborate this more. First, partition the structural equation as:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i = \boldsymbol{x}_{1i}^T \boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}^T \boldsymbol{\beta}_2 + u_i$$

whereby we have partitioned the explanatory variables as

$$\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{x}_{2i} \end{bmatrix} = \begin{matrix} \text{Exogenous variables, } \mathbb{E}[\boldsymbol{x}_{1i} u_i] = 0 \\ \text{Endogenous variables, } \mathbb{E}[\boldsymbol{x}_{2i} u_i] \neq 0 \end{matrix}$$

Here, $\boldsymbol{x}_{1i}$ explanatory variables are **valid** instruments for themselves as they are exogenous. We can also partition the instruments $\boldsymbol{z}_i$ as:

$$\boldsymbol{z}_i = \begin{bmatrix} \boldsymbol{z}_{1i} \\ \boldsymbol{z}_{2i} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{z}_{2i} \end{bmatrix} = \begin{matrix} \text{Exogenous variables, } \mathbb{E}[\boldsymbol{x}_{1i} u_i] = 0 \\ \text{Excluded instruments, } \mathbb{E}[\boldsymbol{z}_{2i} u_i] = 0 \end{matrix}$$

The variables $\boldsymbol{z}_{1i}$ are the exogenous variables $\boldsymbol{x}_{1i}$ and $\boldsymbol{z}_{2i} \in \mathbb{R}^M$ are the **excluded** (outside) instruments. These variables are called excluded instruments as they are variables that do not appear in the population model in Eq. (281). To be clear on terminology, $\boldsymbol{z}_i$ are referred to the **instruments** and $\boldsymbol{z}_{2i}$ are referred to the **excluded instruments**. Therefore, $\boldsymbol{z}_i$ contains the intercept term, exogenous explanatory variables and the excluded instruments.

Let us also be clear on the dimensions of all our vectors and matrices. The parameter of interest $\boldsymbol{\beta}$ is a $K \times 1$ vector. The explanatory variables, which contains both exogenous and endogenous variables, is a $K \times 1$ vector. We denote K_{x_1} as the number of exogenous variables (including intercept) and K_{x_2} are the number of endogenous variables such that $K_{x_1} + K_{x_2} = K$. The instrumental variables $\boldsymbol{z}_i$ is a $L \times 1$ vector which contains the exogenous variables $\boldsymbol{x}_{1i}$ and the excluded instruments $\boldsymbol{z}_{2i}$. The excluded instruments $\boldsymbol{z}_{2i}$ is a $M \times 1$ vector such that $L = M + K_{x_1}$ whereby K_{x_1} is the number of exogenous explanatory variables and L is the total number of instruments.

Now, we wish to derive the population moment conditions to identify the true population parameter $\boldsymbol{\beta}$. We can start off with our instrumental variable orthogonality condition and substitute in our true population model:

$$\begin{aligned} 0 &= \mathbb{E}[\boldsymbol{z}_i u_i] \\ &= \mathbb{E}[\boldsymbol{z}_i [y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}]] \\ &= \mathbb{E}[g(w_i; \boldsymbol{\beta})] \end{aligned}$$

where we defined $w_i = (\boldsymbol{x}_i, y_i, \boldsymbol{z}_i)$ as our observed variables and $g(w_i; \tilde{\boldsymbol{\beta}}) = \boldsymbol{z}_i [y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}]$.

Proposition 338 (Solution to Population Moment Condition)
The population moment condition is solved:

$$\mathbb{E}[g(w_i; \tilde{\boldsymbol{\beta}})] = \mathbb{E}[\boldsymbol{z}_i [y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}]] = 0$$

when the estimator $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

Proof. Let $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ and use the IV orthogonality assumption:

$$\begin{aligned} \mathbb{E}[\boldsymbol{z}_i [y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}]] &= \mathbb{E}[\boldsymbol{z}_i [y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}]] \\ &= \mathbb{E}[\boldsymbol{z}_i [u_i]] = 0. \end{aligned}$$

□

We can also show that this is an unique solution.

Proposition 339 (Unique Solution to Population Moment Condition)
The estimate $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ is the **unique solution** to the population moment condition:

$$\mathbb{E}[g(w_i; \tilde{\boldsymbol{\beta}})] = \mathbb{E}[\boldsymbol{z}_i [y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}]] = 0$$

if the **rank condition** $rank(\mathbb{E}[\boldsymbol{z}_i \boldsymbol{x}_i^T]) = K$ holds.

Proof. Plug in the true population model and use the IV orthogonality assumption:

$$\begin{aligned}\mathbb{E}[z_i(y_i - x_i^T\tilde{\boldsymbol{\beta}})] &= \mathbb{E}[z_i(x_i^T\boldsymbol{\beta} + u_i - x_i^T\tilde{\boldsymbol{\beta}})] \\ &= \mathbb{E}[z_i x_i^T\boldsymbol{\beta} + z_i u_i - z_i x_i^T\tilde{\boldsymbol{\beta}}] \\ &= \mathbb{E}[z_i x_i^T\boldsymbol{\beta} - z_i x_i^T\tilde{\boldsymbol{\beta}}] + \mathbb{E}[z_i u_i] \\ &= \mathbb{E}[z_i x_i^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] + 0 \\ &= \mathbb{E}[z_i x_i^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})]\end{aligned}$$

Again, we can argue that the only way for $\mathbb{E}[z_i x_i^T](\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) = 0$ is if $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$. The matrix $\mathbb{E}[z_i x_i^T]$ is $L \times K$, where $L \geq K$, and by assumption it has full column rank K. Therefore, its null space contains only the zero vector, so $\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}} = \mathbf{0}$ and hence $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

□

Remark 340. *The assumption $rank(\mathbb{E}[z_i x_i^T]) = K$ was the identification assumption used in the section on instrumental variable estimation.*

Now, we want to construct our sample moment conditions from our population moment function $\mathbb{E}[g(w_i;\tilde{\boldsymbol{\beta}})]$.

Definition 341 (Sample Moment Condition in Linear Regression with IV)

The **sample moment condition** in linear regression with instrumental variables is given by:

$$\frac{1}{n}\sum_{i=1}^{n} g(w_i;\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i^T\tilde{\boldsymbol{\beta}}) = 0. \tag{283}$$

Looking at the sample moment equation, we can see that we have a notable issue. $z_i \in \mathbb{R}^L$ so then $z_i(y_i - x_i^T\tilde{\boldsymbol{\beta}})$ is a $L \times 1$ vector. So we have L equations. However, we only have K parameters we wish to estimate $\boldsymbol{\beta}$. And by assumption, we usually have that $L \geq K$. Therefore, we have an overdetermined system with L equations but K unknowns.

As a result, we have two potential cases that can occur.

Case: L = K. Here, we have as many instruments as parameters to identify. This means that the system is **just identified** and we can solve for the K unknowns with the L equations (assuming linearity of the system).

Theorem 342 (Instrumental Variable OLS Estimator)

Assuming a just-identified system where $L = K$, then the instrumental variable OLS estimator is given by:

$$\hat{\boldsymbol{\beta}}_{IV} = (\boldsymbol{Z}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\boldsymbol{y} \tag{284}$$

Proof. First, look at the sample moment condition:

$$\frac{1}{n}\sum_{i=1}^{n} g(w_i;\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i^T\tilde{\boldsymbol{\beta}}) = 0$$

Drop the n^{-1} scaling:

$$\sum_{i=1}^{n} z_i(y_i - x_i^T\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n} z_i y_i - \sum_{i=1}^{n} z_i x_i^T\tilde{\boldsymbol{\beta}} = 0$$

Re-arrange to get:

$$\sum_{i=1}^{n} z_i x_i^T\tilde{\boldsymbol{\beta}} = \sum_{i=1}^{n} z_i y_i$$

Assuming invertibility of $\sum_{i=1}^{n} z_i x_i^T$, which holds with probability approaching one under the population rank condition, we therefore have that:

$$\hat{\boldsymbol{\beta}}_{IV} = \Big(\sum_{i=1}^{n} z_i x_i^T\Big)^{-1} \sum_{i=1}^{n} z_i y_i$$

or in matrix form:

$$\hat{\boldsymbol{\beta}}_{IV} = (\boldsymbol{Z}^T \boldsymbol{X})^{-1} \boldsymbol{Z}^T \boldsymbol{y}$$

□

Case: $L > K$. Here, we have more instruments than parameters. This means that the system is **over identified** and we can't solve for the K unknowns with the L equations easily. One to way estimate the parameters is to use the method of **two stage least squares** (2SLS).

Theorem 343 (2SLS Estimator)
Assuming an over identified system where $L > K$, then the Two-Stage Least Squares estimator is given by:

$$\hat{\boldsymbol{\beta}}_{2SLS} = (\hat{\boldsymbol{X}}^T \boldsymbol{X})^{-1} \hat{\boldsymbol{X}}^T \boldsymbol{y}. \tag{285}$$

Proof. Here, we can take the sample moment conditions:

$$\frac{1}{n} \sum_{i=1}^{n} g(w_i; \tilde{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} z_i (y_i - x_i^T \tilde{\boldsymbol{\beta}}) = 0$$

where we said that $z_i(y_i - x_i^T \tilde{\boldsymbol{\beta}})$ was a $L \times 1$ vector. What we can do is to pre-multiply this expression by the first stage regression coefficients $\hat{\pi}_n^T$ and this will reduce the system down to a K dimensional system. We will therefore have K equations and K unknowns and hence we can solve for the estimators explicitly.

First, recall that the first stage regression is given by regressing the explanatory variables on the instrumental variables:

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{R}$$

whereby $\boldsymbol{\Pi}$ is a $L \times K$ matrix of the coefficients on the L instruments, $\boldsymbol{Z}$ is a $n \times L$ matrix containing the observed instruments, and $\boldsymbol{X}$ is the $n \times K$ matrix of explanatory variables and $\boldsymbol{R}$ is the error terms. Then, the OLS coefficients in this first stage regression are given by:

$$\hat{\boldsymbol{\Pi}} \equiv (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{X}$$

whereby $\hat{\boldsymbol{\Pi}}$ is a $L \times K$ matrix. Furthermore, recall that the predicted values of the explanatory variables $\boldsymbol{X}$ are then:

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\boldsymbol{\Pi}} = \boldsymbol{Z}(\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{X}.$$

We can then pre-multiply the sample moment conditions with this first stage coefficients:

$$0 = \frac{1}{n} \hat{\boldsymbol{\Pi}}^T \sum_{i=1}^{n} z_i (y_i - x_i^T \tilde{\boldsymbol{\beta}})$$

We now have a system of K equations and K unknowns! We now recognise that we can use the definition of the predicted values of $\hat{\boldsymbol{X}}$ to get:

$$= \frac{1}{n} \sum_{i=1}^{n} \hat{x}_i (y_i - x_i^T \tilde{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} \hat{x}_i y_i - \frac{1}{n} \sum_{i=1}^{n} \hat{x}_i x_i^T \tilde{\boldsymbol{\beta}} = 0$$

So we therefore have that:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i x_i^T\tilde{\beta} = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i y_i$$

We have invertibility of $\sum_{i=1}^{n}\hat{x}_i x_i^T$ and therefore we get the result:

$$\hat{\beta}_{2SLS} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i x_i^T\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i y_i$$

or in matrix form:

$$\hat{\beta}_{2SLS} = (\hat{X}^T X)^{-1}\hat{X}^T y.$$

□

We can also express the 2SLS estimator in an alternative manner.

Proposition 344 (Alternative Expression for 2SLS)
The 2SLS estimator

$$\hat{\beta}_{2SLS} = (\hat{X}^T X)^{-1}\hat{X}^T y$$

can be expressed as:

$$\hat{\beta}_{2SLS} = (\hat{X}^T \hat{X})^{-1}\hat{X}^T y$$

Proof. Recall that the predicted value is: $\hat{X} = Z(Z^T Z)^{-1}Z^T X$. Then, the transpose is:

$$\hat{X}^T = (Z(Z^T Z)^{-1}Z^T X)^T = X^T Z(Z^T Z)^{-1}Z^T.$$

Now, notice that the two 2SLS estimator expressions differ by $\hat{X}^T X$ and $\hat{X}^T\hat{X}$. We can work backwards to see that:

$$\begin{aligned}
\hat{X}^T\hat{X} &= \left[Z(Z^T Z)^{-1}Z^T X\right]^T\left[Z(Z^T Z)^{-1}Z^T X\right] \\
&= \left[X^T Z(Z^T Z)^{-1}Z^T\right]\left[Z(Z^T Z)^{-1}Z^T X\right] \\
&= X^T Z(Z^T Z)^{-1}Z^T Z(Z^T Z)^{-1}Z^T X \\
&= X^T Z(Z^T Z)^{-1}I_L Z^T X \\
&= X^T Z(Z^T Z)^{-1}Z^T X \\
&= \left[Z(Z^T Z)^{-1}Z^T X\right]^T X \\
&= \hat{X}^T X.
\end{aligned}$$

Therefore, we can replace $\hat{X}^T X$ in our original expression of the 2SLS by $\hat{X}^T\hat{X}$ to get the new expression:

$$\hat{\beta}_{2SLS} = (\hat{X}^T\hat{X})^{-1}\hat{X}^T y.$$

□

All of this gives us a nice way to motivate and view that 2SLS as a way of reducing an overdetermined system of sample moment conditions. After doing this reduction, we are then able to explicitly solve this reduced system for the 2SLS estimators.

11.2 GMM General Framework

We can now specify what our general framework is for GMM. Suppose that we have i.i.d. data $w_i \in \mathbb{R}^N$ (this could be $y_i, \boldsymbol{x}_i, \boldsymbol{z}_i$). Then, suppose we have a model that is parameterised by the parameter vector $\tilde{\boldsymbol{\theta}} \in \Theta \subseteq \mathbb{R}^K$. Furthermore, there exists a true population parameter vector $\boldsymbol{\theta} \in \Theta$. Notation-wise, $\tilde{\boldsymbol{\theta}}$ refers to any particular value that the population parameter vector can take, whilst $\boldsymbol{\theta}$ refers to the unique **true** population parameter implied by the model. The unique true vector $\boldsymbol{\theta}$ is what we are trying to estimate. Then, the moment conditions g are a set of functions:

$$g : \mathbb{R}^N \times \Theta \to \mathbb{R}^L \quad \text{or } (w, \tilde{\boldsymbol{\theta}}) \to g(w; \tilde{\boldsymbol{\theta}})$$

such that the population moment conditions hold at the true parameter:

$$\mathbb{E}[g(w_i; \boldsymbol{\theta})] = \mathbf{0}$$

which is a set of L equations. We generally require $L \geq K$ for identification and in the case that $L > K$, then we have an overidentified system whereby we have more moment conditions than parameters. We can then construct sample moment equations based off the population moment conditions:

$$\overline{g}_n(\tilde{\boldsymbol{\theta}}) \equiv \frac{1}{n} \sum_{i=1}^{n} g(w_i; \tilde{\boldsymbol{\theta}})$$

and we try to set these as close to zero as possible. There are **two ways** to set these sample moment conditions as close as possible to zero in the case of an **overidentified system**.

The first way requires us to define a new type of function.

Definition 345 (Criterion Function)
The criterion function $Q_n(\tilde{\boldsymbol{\theta}})$ is defined to be the scalar:

$$Q_n(\tilde{\boldsymbol{\theta}}) = \overline{g}_n(\tilde{\boldsymbol{\theta}})^T \boldsymbol{W}_n \overline{g}_n(\tilde{\boldsymbol{\theta}})$$

whereby $\boldsymbol{W}_n$ is a positive definite $L \times L$ weight matrix and $\overline{g}_n(\tilde{\boldsymbol{\theta}})$ are the sample moment conditions.

We can now define the GMM estimator.

Definition 346 (GMM Estimator)
Given the criterion function Q_n, the GMM estimator $\hat{\boldsymbol{\theta}}_{GMM}$ is defined to be the argument of the minimisation program:

$$\hat{\boldsymbol{\theta}}_{GMM} \equiv \arg\min_{\tilde{\boldsymbol{\theta}} \in \Theta} Q_n(\tilde{\boldsymbol{\theta}})$$

Remark 347. *The weight matrix $\boldsymbol{W}_n$ is how we trade-off between how important it is for 1 moment to be closer to 0 compared to another moment. Choosing different weight matrices $\boldsymbol{W}_n$ will give us different efficiency properties of the GMM estimator.*

The second approach we can choose is to pick a $L \times K$ matrix $\boldsymbol{L}$ and solve the sample moment conditions:

$$\boldsymbol{L}^T \overline{g}_n(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$$

whereby we now have a reduced system of K equations. We now have a system of K equations and K unknowns and we can set this system exactly equal to 0.

11.2.1 GMM Applied to IV Model

We take our GMM estimator and apply it to the linear regression model with instrumental variables to see how it works in practice.

Theorem 348 (GMM Estimator with IV Variables)
Under the linear model with instrumental variables, the GMM estimator has the form:

$$\hat{\boldsymbol{\beta}}_{GMM} = \left[(\boldsymbol{X}^T\boldsymbol{Z})\boldsymbol{W}_n(\boldsymbol{Z}^T\boldsymbol{X})\right]^{-1}(\boldsymbol{X}^T\boldsymbol{Z})\boldsymbol{W}_n(\boldsymbol{Z}^T\boldsymbol{y}) \tag{286}$$

Proof. The population moment conditions are given by:

$$\boldsymbol{0} = \mathbb{E}[g(\boldsymbol{w}_i;\boldsymbol{\beta})] = \mathbb{E}[\boldsymbol{z}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})]$$

The sample analogue is given by:

$$\overline{g}_n(\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) = n^{-1}\boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$$

The GMM estimator is defined as:

$$\hat{\boldsymbol{\beta}}_{GMM} = \arg\min_{\tilde{\boldsymbol{\beta}}} \boldsymbol{Q}_n(\tilde{\boldsymbol{\beta}}) = \arg\min_{\tilde{\boldsymbol{\beta}}} \overline{g}_n(\tilde{\boldsymbol{\beta}})^T\boldsymbol{W}_n\overline{g}_n(\tilde{\boldsymbol{\beta}}).$$

We need this following chain rule for quadratic forms to be used later.

Theorem 349 (Chain Rule of Quadratic Forms)
For any differentiable function $h : \Theta \to \mathbb{R}^M$,

$$\nabla_{\boldsymbol{\theta}}\left[h(\boldsymbol{\theta})^T\boldsymbol{A}h(\boldsymbol{\theta})\right] = [D_{\boldsymbol{\theta}}h(\boldsymbol{\theta})]^T(\boldsymbol{A} + \boldsymbol{A}^T)h(\boldsymbol{\theta}) \tag{287}$$

whereby $\nabla_{\boldsymbol{\theta}}$ is the gradient vector of partial derivatives of $\boldsymbol{\theta}$ and $D_{\boldsymbol{\theta}}h(\boldsymbol{\theta})$ is the Jacobian matrix.

We now compute the first order conditions required for a minimum.

$$\begin{aligned}
\boldsymbol{0} &= \nabla_{\tilde{\boldsymbol{\beta}}}Q_n(\tilde{\boldsymbol{\beta}}) \\
&= \nabla_{\tilde{\boldsymbol{\beta}}}\left[\overline{g}_n(\tilde{\boldsymbol{\beta}})^T\boldsymbol{W}_n\overline{g}_n(\tilde{\boldsymbol{\beta}})\right] \\
&=_{(1)} [D_{\tilde{\boldsymbol{\beta}}}\overline{g}_n(\tilde{\boldsymbol{\beta}})]^T(\boldsymbol{W}_n + \boldsymbol{W}_n^T)\overline{g}_n(\tilde{\boldsymbol{\beta}}) \\
&=_{(2)} 2[D_{\tilde{\boldsymbol{\beta}}}\overline{g}_n(\tilde{\boldsymbol{\beta}})]^T\boldsymbol{W}_n\overline{g}_n(\tilde{\boldsymbol{\beta}}) \\
&=_{(3)} 2[D_{\tilde{\boldsymbol{\beta}}}(n^{-1}\boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}))]^T\boldsymbol{W}_n n^{-1}\boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) \\
&= \frac{2}{n}[D_{\tilde{\boldsymbol{\beta}}}(n^{-1}\boldsymbol{Z}^T\boldsymbol{y} - n^{-1}\boldsymbol{Z}^T\boldsymbol{X}\tilde{\boldsymbol{\beta}})]^T\boldsymbol{W}_n\boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) \\
&= \frac{2}{n}[-n^{-1}\boldsymbol{Z}^T\boldsymbol{X}]^T\boldsymbol{W}_n\boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) \\
&= -\frac{2}{n^2}\boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})
\end{aligned}$$

In $=_{(1)}$, we applied the chain rule of quadratic forms. In $=_{(2)}$, we use the fact that $\boldsymbol{W} = \boldsymbol{W}^T$ since it is a positive definite matrix. In $=_{(3)}$, we plugged in our definition of sample analogues $\overline{g}_n(\tilde{\boldsymbol{\beta}}) = n^{-1}\boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$.

So therefore, we have shown that:

$$\mathbf{0} = -\frac{2}{n^2}\boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GMM}).$$

We then have that:

$$\boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{GMM} = \boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}^T\boldsymbol{y}$$

We can then invert the terms on the left hand side to get our final desired result:

$$\hat{\boldsymbol{\beta}}_{GMM} = \left[(\boldsymbol{X}^T\boldsymbol{Z})\boldsymbol{W}_n(\boldsymbol{Z}^T\boldsymbol{X})\right]^{-1}(\boldsymbol{X}^T\boldsymbol{Z})\boldsymbol{W}_n(\boldsymbol{Z}^T\boldsymbol{y}).$$

□

Proposition 350 (GMM Estimator with the 2SLS Weight Matrix)
Let $\hat{\boldsymbol{\beta}}_{GMM}$ be the GMM estimator in the linear model with instrumental variables. If we set the weight matrix to be:

$$\boldsymbol{W}_n = [\boldsymbol{Z}^T\boldsymbol{Z}]^{-1}$$

then the GMM estimator is equivalent to the Two-Stage Least Squares estimator:

$$\hat{\boldsymbol{\beta}}_{GMM} = [\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}]^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{2SLS}.$$

Proof. First, notice that if $\boldsymbol{W}_n = [\boldsymbol{Z}^T\boldsymbol{Z}]^{-1}$, then we have that:

$$\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}^T = \boldsymbol{Z}[\boldsymbol{Z}^T\boldsymbol{Z}]^{-1}\boldsymbol{Z}^T = P_{\boldsymbol{Z}}$$

where $P_{\boldsymbol{Z}}$ is the projection matrix onto the subspace spanned by $\boldsymbol{Z}$. Finally, recall that the predicted values $\hat{\boldsymbol{X}} = P_{\boldsymbol{Z}}\boldsymbol{X}$. Therefore, we can plug in our expression for $\boldsymbol{W}_n = [\boldsymbol{Z}^T\boldsymbol{Z}]^{-1}$ and notice the projection matrix $P_{\boldsymbol{Z}}$ popping up:

into our expression for the GMM estimator:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{GMM} &= \left[(\boldsymbol{X}^T\boldsymbol{Z})\boldsymbol{W}_n(\boldsymbol{Z}^T\boldsymbol{X})\right]^{-1}(\boldsymbol{X}^T\boldsymbol{Z})\boldsymbol{W}_n(\boldsymbol{Z}^T\boldsymbol{y}) \\
&= \left[\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}^T)\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}^T)\boldsymbol{y} \\
&= \left[\boldsymbol{X}^T P_{\boldsymbol{Z}}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T P_{\boldsymbol{Z}}\boldsymbol{y} \\
&= \left[\boldsymbol{X}^T\hat{\boldsymbol{X}}\right]^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{2SLS}.
\end{aligned}$$

□

11.3 Asymptotics

11.3.1 Consistency of GMM Estimator

We now want to show the consistency of the GMM estimator:

$$\hat{\boldsymbol{\theta}}_{GMM} \equiv \arg\min_{\tilde{\boldsymbol{\theta}} \in \Theta} Q_n(\tilde{\boldsymbol{\theta}}) = \arg\min_{\tilde{\boldsymbol{\theta}} \in \Theta} \overline{g}_n(\tilde{\boldsymbol{\theta}})^T \boldsymbol{W}_n \overline{g}_n(\tilde{\boldsymbol{\theta}})$$

where $\boldsymbol{W}_n$ is the positive definite weight matrix and

$$\overline{g}_n(\tilde{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} g(w_i; \tilde{\boldsymbol{\theta}})$$

are the sample moments.

We have an issue though. For a GMM estimator, we do not have an explicit formula for the GMM estimator unless we have linear moment conditions. Therefore, it is not so straightforward to derive asymptotic properties for the GMM estimator. However, what we can do is, instead of looking at the GMM estimator, which is defined to be the minimiser of the criterion function $Q_n(\tilde{\boldsymbol{\theta}})$, we can look at the asymptotic properties of the criterion function $Q_n(\tilde{\boldsymbol{\theta}})$ and infer asymptotic properties of the GMM estimator from the asymptotic properties of the criterion function. To proceed, we need two assumptions to help us.

Assumption 351 (GMM-ID)

The true parameter $\boldsymbol{\theta}$ uniquely solves the **population moment conditions**:

$$\mathbb{E}[g(w_i; \boldsymbol{\theta})] = \mathbf{0}.$$

Assumption 352 (GMM-WGHT)

The weight matrix $\boldsymbol{W}_n$ is a positive definite matrix and

$$\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{W}$$

where $\boldsymbol{W}$ is a positive definite matrix.

We can now state and show our consistency property of the GMM estimator.

Proposition 353 (Consistency of the GMM Estimator)

Under the assumptions [GMM-ID] and [GMM-WGHT], together with standard regularity conditions ensuring uniform convergence of Q_n to a well-separated population criterion, the GMM estimator is a consistent estimator:

$$\hat{\boldsymbol{\theta}}_{GMM} \xrightarrow{P} \boldsymbol{\theta} \tag{288}$$

where $\boldsymbol{\theta}$ is the true population parameter vector.

Proof. (Heuristic Proof). Unfortunately, we cannot do a proper proof since we need to actually show a form of uniform convergence, but we sketch a heuristic proof.

First, for each fixed parameter $\tilde{\boldsymbol{\theta}} \in \Theta$, we have that by the **weak law of large numbers**, the sample moments converge to the population moments:

$$\overline{g}_n(\tilde{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} g(w_i; \tilde{\boldsymbol{\theta}}) \xrightarrow{P} \mathbb{E}[g(w_i; \tilde{\boldsymbol{\theta}})]$$

We define:

$$g_0(\tilde{\boldsymbol{\theta}}) \equiv \mathbb{E}[g(w_i; \tilde{\boldsymbol{\theta}})]$$

to the be the population moments. We note that by [GMM-ID] assumption, $g_0(\tilde{\boldsymbol{\theta}}) \neq \mathbf{0}$ if $\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}$.

By the **continuous mapping theorem**, for each parameter $\tilde{\boldsymbol{\theta}} \in \Theta$, we have that the sample criterion function converges to the population criterion function:

$$Q_n(\tilde{\boldsymbol{\theta}}) = \overline{g}_n^T(\tilde{\boldsymbol{\theta}})\boldsymbol{W}_n\overline{g}_n(\tilde{\boldsymbol{\theta}}) \xrightarrow{P} g_0(\tilde{\boldsymbol{\theta}})^T\boldsymbol{W}g_0(\tilde{\boldsymbol{\theta}}) =: Q(\tilde{\boldsymbol{\theta}})$$

where we invoked [GMM-WGHT] assumption for the convergence of $\boldsymbol{W}$. By [GMM-ID] assumption and the fact that the weight matrix $\boldsymbol{W}$ is positive definite, we have that:

$$Q(\tilde{\boldsymbol{\theta}}) = \begin{cases} 0 & \text{if } \tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} \\ > 0 & \text{if } \tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta} \end{cases}$$

We can see that the population criterion function $Q(\tilde{\boldsymbol{\theta}})$ will therefore have a minimum at the true population parameter $\boldsymbol{\theta}$ since at $\boldsymbol{\theta}$, we have that $Q(\boldsymbol{\theta}) = 0$ and otherwise, $Q(\tilde{\boldsymbol{\theta}}) > 0$ for $\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}$.

We have shown that the sample criterion function Q_n converges to the population criterion function Q. We heuristically argue that this implies that the GMM estimator $\hat{\boldsymbol{\theta}}_{GMM}$ which is the minimiser of the sample criterion function Q_n (pointwise) converges to the true population parameter $\boldsymbol{\theta}$ which is the minimiser of the population criterion function Q. Hence, we have shown our result that:

$$\hat{\boldsymbol{\theta}}_{GMM} \xrightarrow{P} \boldsymbol{\theta}$$

□

11.3.2 Asymptotic Normality of GMM Estimator

We want to be able to do inference on the models we are estimating and hence we need to be able to say something about the distribution of the estimators. We require two regularity conditions.

Assumption 354 (GMM-DIFF)
The moment conditions $\tilde{\boldsymbol{\theta}} \to g(w; \tilde{\boldsymbol{\theta}})$ are continuously differentiable for all $w \in \mathbb{R}^N$.

Assumption 355 (GMM-INTR)
The true population parameter is in the interior of the parameter space $\boldsymbol{\theta} \in int(\Theta)$.

[GMM-DIFF] will allow us to compute the first order conditions needed when computing a GMM estimator. [GMM-INTR] will ensure that our estimator will lie in the interior of the parameter space and hence ensure we have an interior solution for the first order conditions. This is because we may have some issues if our minimum is on the boundary of the parameter space as then the first order conditions may not necessarily hold.

We state two more assumptions we will invoke later and explain why we need them later.

Assumption 356 (GMM-JAC)
The Jacobian of the population moments under true values is full rank:

$$Rank(\boldsymbol{D}) = Rank(\boldsymbol{D}_{\boldsymbol{\theta}}\mathbb{E}[g(w_i; \boldsymbol{\theta})]) = K.$$

Assumption 357 (GMM-VAR)

The population covariance matrix $\boldsymbol{S}$ of the moment equations $g(w_i;\boldsymbol{\theta})$ under the true parameter $\boldsymbol{\theta}$:

$$\boldsymbol{S} \equiv \mathbb{E}[g(w_i;\boldsymbol{\theta})g(w_i;\boldsymbol{\theta})^T] \tag{289}$$

is a positive definite matrix.

We can now finally state the limiting distribution of the GMM estimator.

Theorem 358 (Asymptotic Normality of GMM Estimator)

Under the assumptions GMM-INTR, GMM-DIFF, GMM-ID, GMM-JAC, GMM-VAR, GMM-WGHT, the GMM estimator $\hat{\boldsymbol{\theta}}_{GMM}$ is **asymptotically normal**:

$$n^{1/2}[\hat{\boldsymbol{\theta}}_{GMM}(\boldsymbol{W}_n) - \boldsymbol{\theta}] \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{W}}) \tag{290}$$

where the variance matrix is given by:

$$\boldsymbol{V}_{\boldsymbol{W}} = (\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}.$$

Proof. First, recall that the GMM estimator satisfies the first order conditions when minimising the criterion function $Q_n(\tilde{\boldsymbol{\theta}})$:

$$\mathbf{0} = \nabla_{\tilde{\boldsymbol{\theta}}} Q_n(\tilde{\boldsymbol{\theta}}) = \nabla_{\tilde{\boldsymbol{\theta}}}\left[\overline{g}_n(\tilde{\boldsymbol{\theta}})^T\boldsymbol{W}_n\overline{g}_n(\tilde{\boldsymbol{\theta}})\right] = 2[\boldsymbol{D}_{\tilde{\boldsymbol{\theta}}}\overline{g}_n(\tilde{\boldsymbol{\theta}})]^T\boldsymbol{W}_n\overline{g}_n(\tilde{\boldsymbol{\theta}}).$$

The GMM estimator $\hat{\boldsymbol{\theta}}_{GMM}$ satisfies the aforementioned first order condition:

$$\mathbf{0} = 2[\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})]^T\boldsymbol{W}_n\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM}) \tag{291}$$

Clearly, we can drop the constant 2 on both sides of the equation. We wish to derive the asymptotic distribution of the GMM estimator from Eq. (291). However, an issue is that Eq. (291) is a **nonlinear** equation and this makes things tricky. Therefore, as we will soon see, we will apply a linearisation to Eq. (291) in order to help us figure out the asymptotic distribution of the GMM estimator. First, let us write Eq. (291) as the product of three terms, which we will analyse the asymptotic properties of separately.

$$[\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})]^T\boldsymbol{W}_n\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \underbrace{[\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})]^T}_{(A)} \times \underbrace{\boldsymbol{W}_n}_{(B)} \times \underbrace{\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})}_{(C)}$$

We now analyse each of the three terms separately.

Claim 359. *The expression (A) converges to the Jacobian of the population moment conditions:*

$$\textbf{(A)} = \boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{P} \boldsymbol{D}_{\boldsymbol{\theta}}\mathbb{E}[g(w_i;\boldsymbol{\theta})]$$

or equivalently:

$$\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \boldsymbol{D}_{\boldsymbol{\theta}}\mathbb{E}[g(w_i;\boldsymbol{\theta})] + o_p(1)$$

Proof. First, we explain what the expression (A) is:

$$\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})$$

This is the Jacobian of the **sample moments** evaluated at the **GMM estimator**. Before analysing its asymptotics, we look at the similar Jacobian of the **sample moments** evaluated at a general parameter $\tilde{\boldsymbol{\theta}} \in \Theta$:

$$\begin{aligned} \boldsymbol{D}_{\boldsymbol{\theta}} \overline{g}_n(\tilde{\boldsymbol{\theta}}) &= \boldsymbol{D}_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} g(w_i; \tilde{\boldsymbol{\theta}}) \\ &= \frac{1}{n} \sum_{i=1}^{n} D_{\boldsymbol{\theta}} g(w_i; \tilde{\boldsymbol{\theta}}) \end{aligned}$$

which holds for a fixed $\tilde{\boldsymbol{\theta}} \in \Theta$. Now, we noticed that as $\tilde{\boldsymbol{\theta}}$ is fixed, and w_i is i.i.d, this implies that $g(w_i; \tilde{\boldsymbol{\theta}})$ is a sequence of i.i.d random variables. Therefore, we can apply the **weak law of large numbers**:

$$\frac{1}{n} \sum_{i=1}^{n} D_{\boldsymbol{\theta}} g(w_i; \tilde{\boldsymbol{\theta}}) \xrightarrow{P} \mathbb{E}[D_{\boldsymbol{\theta}} g(w_i; \tilde{\boldsymbol{\theta}})] =_{(1)} D_{\boldsymbol{\theta}} \mathbb{E}[g(w_i; \tilde{\boldsymbol{\theta}})] = D_{\boldsymbol{\theta}} g_0(\tilde{\boldsymbol{\theta}})$$

whereby we interchanged the expectation and differentiation operator by the **dominating convergence theorem**. This holds for any $\tilde{\boldsymbol{\theta}} \in \Theta$ but we were interested for whether does it holds for our GMM estimator. We know that $\hat{\boldsymbol{\theta}}_{GMM} \xrightarrow{P} \boldsymbol{\theta}$ is a consistent estimator so we can plug in the GMM estimator our the expression we just derived and therefore conclude that:

$$\boldsymbol{D}_{\boldsymbol{\theta}} \overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{P} D_{\boldsymbol{\theta}} g_0(\boldsymbol{\theta}) \equiv \boldsymbol{D}$$

which says that the Jacobian of the sample moment conditions with the GMM estimator converges to the Jacobian of the population moment conditions at the true parameter, which we denote by $\boldsymbol{D}$. Equivalently, by little oh-pee notation, we could have written this as:

$$\boldsymbol{D}_{\boldsymbol{\theta}} \overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \boldsymbol{D} + o_p(1).$$

□

Claim 360. *The expression*

$$\textbf{(B)} = \boldsymbol{W}_n \xrightarrow{P} \boldsymbol{W}.$$

Proof. By the assumption [GMM-WGHT], we immediately have that:

$$\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{W}.$$

□

Claim 361. *The sample moments evaluated at the GMM estimator has a first order Taylor expansion around the true population parameter $\boldsymbol{\theta}$ given by:*

$$\textbf{(C)} = \overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \overline{g}_n(\boldsymbol{\theta}) + (D_{\boldsymbol{\theta}} g_0(\boldsymbol{\theta}) + o_p(1))(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta})$$

Proof. First, we require a stochastic first-order Taylor expansion theorem.

Proposition 362 (Stochastic First-Order Taylor Expansion)
Suppose that $\boldsymbol{\theta} \to h(\boldsymbol{w}; \boldsymbol{\theta})$ is twice continuously differentiable for every $\boldsymbol{w} \in \mathbb{R}^N$ and that $\sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}[||\boldsymbol{D}_{\boldsymbol{\theta}} h(w_i; \boldsymbol{\theta})||] < \infty$. Then, for every $\tilde{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta} \in int(\Theta)$, we have that:

$$\overline{h}_n(\tilde{\boldsymbol{\theta}}_n) - \overline{h}_n(\boldsymbol{\theta}) = [\boldsymbol{D}_{\boldsymbol{\theta}} \overline{h}_n(\boldsymbol{\theta}) + o_p(1)](\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \tag{292}$$

whereby $\overline{h}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} h(w_i; \boldsymbol{\theta})$.

We can immediately apply this Taylor expansion to $\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})$ around the true population parameter $\boldsymbol{\theta}$ to get:

$$\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM}) - \overline{g}_n(\boldsymbol{\theta}) = [\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\boldsymbol{\theta}) + o_p(1)](\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta})$$

However, we have already shown that:

$$\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\boldsymbol{\theta}) \xrightarrow{P} \boldsymbol{D}_{\boldsymbol{\theta}} g_0(\boldsymbol{\theta}) \equiv \boldsymbol{D}$$

and hence we have the result:

$$\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM}) - \overline{g}_n(\boldsymbol{\theta}) = [\boldsymbol{D} + o_p(1)](\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta})$$

□

We collect our three expressions and hence we have that our first order condition:

$$\mathbf{0} = [\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})]^T \boldsymbol{W}_n \overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \underbrace{[\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})]^T}_{(A)} \times \underbrace{\boldsymbol{W}_n}_{(B)} \times \underbrace{\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})}_{(C)}$$

can be written as:

$$\mathbf{0} = \Big[\boldsymbol{D} + o_p(1)\Big]^T \boldsymbol{W}_n \Big[\overline{g}_n(\boldsymbol{\theta}) + [\boldsymbol{D} + o_p(1)](\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta})\Big]$$

Now, we restrict our attention to the sample moment averages at the true parameter:

$$\overline{g}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(w_i; \boldsymbol{\theta})$$

Clearly, by definition, the moments have population mean 0

$$\mathbb{E}[g(w_i; \boldsymbol{\theta})] = \mathbf{0}$$

by the GMM-ID assumption. Therefore, we have a sample average with mean 0. We can divide by $\sqrt{n}$ and invoke the **central limit theorem**

$$n^{1/2}\overline{g}_n(\boldsymbol{\theta}) = \frac{1}{n^{1/2}}\sum_{i=1}^{n} g(w_i; \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{S})$$

whereby $\boldsymbol{S} \equiv \mathbb{E}[g(w_i; \boldsymbol{\theta})g(w_i; \boldsymbol{\theta})^T]$ is the variance-covariance matrix of the sample moments as the expectation is 0. Therefore, we take our original expression and multiply by $n^{1/2}$ and re-arrange to get:

$$\begin{aligned} n^{1/2}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}) &= -\Big[\big[\boldsymbol{D} + o_p(1)\big]^T \boldsymbol{W}_n \big[\boldsymbol{D} + o_p(1)\big]\Big]^{-1} \times \big[\boldsymbol{D} + o_p(1)\big]^T \boldsymbol{W}_n n^{1/2}\overline{g}_n(\boldsymbol{\theta}) \\ = -\Big[\boldsymbol{D}^T\boldsymbol{W}_n\boldsymbol{D} + \boldsymbol{D}^T\boldsymbol{W}_n o_p(1) + o_p(1)\boldsymbol{W}_n\boldsymbol{D} + o_p(1)\boldsymbol{W}_n o_p(1)\Big]^{-1} &\times \Big(\boldsymbol{D}^T\boldsymbol{W}_n n^{1/2}\overline{g}_n(\boldsymbol{\theta}) + o_p(1)\boldsymbol{W}_n n^{1/2}\overline{g}_n(\boldsymbol{\theta})\Big) \\ &= -[\boldsymbol{D}^T\boldsymbol{W}_n\boldsymbol{D}]^{-1} \times \boldsymbol{D}^T\boldsymbol{W}_n n^{1/2}\overline{g}_n(\boldsymbol{\theta}) + o_p(1) \end{aligned}$$

So we have that:

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}) = -[\boldsymbol{D}^T\boldsymbol{W}_n\boldsymbol{D}]^{-1} \times \boldsymbol{D}^T\boldsymbol{W}_n n^{1/2}\overline{g}_n(\boldsymbol{\theta}) + o_p(1)$$

By **Slutsky's theorem** and GMM-WGHT assumption, we have that:

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}) = -[\boldsymbol{D}^T\boldsymbol{W}_n\boldsymbol{D}]^{-1} \times \boldsymbol{D}^T\boldsymbol{W}_n n^{1/2}\overline{g}_n(\boldsymbol{\theta}) + o_p(1) \xrightarrow{D} -[\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D}]^{-1}\boldsymbol{D}^T\boldsymbol{W} \cdot \mathcal{N}(\mathbf{0}, \boldsymbol{S})$$

where we have that:

$$[\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D}]^{-1}\boldsymbol{D}^T\boldsymbol{W}\cdot\mathcal{N}(\boldsymbol{0},\boldsymbol{S})=\mathcal{N}(\boldsymbol{0},(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}).$$

Notice that in these past steps, we assumed that

$$\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D}$$

was invertible. Well, we know that $\boldsymbol{W}$ is a positive-definite matrix by assumption. Therefore, we have that:

$$Rank(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})=Rank(\boldsymbol{D})$$

as $\boldsymbol{W}$ is a positive definite matrix. Therefore, in order for $\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D}$ to be invertible, we require that the matrix $\boldsymbol{D}$ is full-rank. However, this was our assumption in GMM-JAC! Therefore,

$$Rank(\boldsymbol{D})=Rank[\boldsymbol{D}_{\boldsymbol{\theta}}\mathbb{E}[g(w_i;\boldsymbol{\theta})]]=K$$

where $\boldsymbol{D}_{\boldsymbol{\theta}}\mathbb{E}[g(w_i;\boldsymbol{\theta})]$ is a $L\times K$ matrix.

To conclude, we have shown that:

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{GMM}-\boldsymbol{\theta})\xrightarrow{D}\mathcal{N}(\boldsymbol{0},\boldsymbol{V}_{\boldsymbol{W}})$$

where the variance matrix is given by:

$$\boldsymbol{V}_{\boldsymbol{W}}=(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}.$$

□

In practice, the Jacobian $\boldsymbol{D}$ and the variance-of-moments matrix $\boldsymbol{S}$ are unknown because they depend on the true population parameter $\boldsymbol{\theta}$. We therefore replace them with estimators evaluated at $\hat{\boldsymbol{\theta}}_{GMM}$:

$$\hat{\boldsymbol{D}}\equiv\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}_{\boldsymbol{\theta}}g(w_i;\hat{\boldsymbol{\theta}}_{GMM}). \tag{293}$$

$$\hat{\boldsymbol{S}}_n\equiv\frac{1}{n}\sum_{i=1}^{n}g(w_i;\hat{\boldsymbol{\theta}}_{GMM})g(w_i;\hat{\boldsymbol{\theta}}_{GMM})^T-\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})^T \tag{294}$$

11.4 Identification

We now want to look more into the rank of the Jacobian assumption in [GMM-Jacobian]. We can in fact see that the rank of the Jacobian of the population moments are in fact connected to identification.

Under the GMM-ID assumption, we have that $\mathbb{E}[g(w_i;\tilde{\boldsymbol{\theta}})]=\boldsymbol{0}$ if and only if $\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}$. We say that $\boldsymbol{\theta}$ is **globally identified** as there is no other parameter in the model that can fulfill this condition. The distribution of the data uniquely pins down the parameter of interest.

Under the GMM-JAC assumption, we assumed that the rank of the Jacobian of the population moment condition was full column rank:

$$\boldsymbol{D}\equiv\boldsymbol{D}_{\boldsymbol{\theta}}\mathbb{E}[g(w_i;\boldsymbol{\theta})]$$

was rank K. This is a sufficient local rank condition and is necessary for the usual regular, asymptotically normal GMM theory, although it is not necessary for identification in every nonlinear model. We can view this condition as a local

identification property. To see why, we say that $\boldsymbol{\theta}$ is **locally identified** if there exists an $\epsilon > 0$ such that

$$\mathbb{E}[g(w_i; \tilde{\boldsymbol{\theta}})] \neq 0$$

for all $0 < ||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}|| < \epsilon$. This says that there is no other $\tilde{\boldsymbol{\theta}}$ that is a solution to the moment conditions besides the true population parameter $\boldsymbol{\theta}$ within a neighborhood. To see how this comes about, let us take a Taylor expansion of the moment condition around a small neighborhood of the true population parameter:

$$\mathbb{E}[g(w_i; \tilde{\boldsymbol{\theta}})] \approx \mathbb{E}[g(w_i; \boldsymbol{\theta})] + \boldsymbol{D}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

The first term is zero by definition $\mathbb{E}[g(w_i; \boldsymbol{\theta})] = 0$. For the second term, we use the linear algebra result that if $\boldsymbol{D}$ has **full column rank** then the **only** way for $\boldsymbol{D}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = 0$ is if the vector $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathbf{0}$ which only holds if $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$. Therefore, we have that the moment conditions:

$$\mathbb{E}[g(w_i; \tilde{\boldsymbol{\theta}})] \approx 0$$

in a small neighborhood around the true population parameter $\boldsymbol{\theta}$ if $\boldsymbol{D}$ is full rank and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$.

The issue with local identification is that it does not preclude other parameters that lie outside of the neighborhood that satisfies the moment conditions.

Proposition 363 (Global and Local Identification for GMM)
In the GMM framework, a parameter is **globally identified** under the [GMM-ID] assumption. A parameter is **locally identified** if [GMM-JAC] assumption is satisfied.

In practice, we can analyse whether local identification holds by considering the rank of $\boldsymbol{D}_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})$, as this is a necessary condition for the asymptotic normality of the GMM estimator. If this matrix is not full rank, we have big issues.

11.4.1 Identification in Linear Model with Instrumental Variables

We apply our ideas of local identification to the instrumental variables framework. Recall that the [GMM-JAC] assumption needed for local identification was that the rank of the Jacobian of the population moment conditions:

$$Rank(\boldsymbol{D}) = Rank(\boldsymbol{D}_{\boldsymbol{\theta}}\mathbb{E}[g(w_i; \boldsymbol{\theta})]) = K$$

is full column rank. In the IV model, recall that $g(w_i; \boldsymbol{\beta}) = \boldsymbol{z}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$. Therefore, the population moment condition is:

$$\mathbb{E}[g(w_i; \boldsymbol{\beta})] = \mathbb{E}[\boldsymbol{z}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})]$$

Then, the Jacobian of this is:

$$\boldsymbol{D} = \boldsymbol{D}_{\boldsymbol{\beta}}\mathbb{E}[\boldsymbol{z}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})] = -\mathbb{E}[\boldsymbol{z}_i\boldsymbol{x}_i^T]$$

We require that $\mathbb{E}[\boldsymbol{z}_i\boldsymbol{x}_i^T]$ has rank K for local identification. However, recall that this also coincided with the global identification required in IV models as we saw in Proposition 339! Therefore, in the IV framework, global and local identification are equivalent to each other. This is due to the fact that the moment conditions are linear.

Theorem 364 (Global and Local Identification in IV Framework)
In the instrumental variables framework, global and local identification are equivalent.

11.5 Asymptotic Efficiency

We have shown that asymptotic normality holds for GMM estimators. However, in order to do inference, we require more efficient regarding the variance of the estimator.

First, we have already shown that under [GMM-WGHT], the weight matrix

$$\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{W}.$$

From this, the asymptotic variance of the GMM estimator $\hat{\boldsymbol{\theta}}_{GMM}(\boldsymbol{W}_n)$ was shown to be:

$$\boldsymbol{V}_{\boldsymbol{W}} = (\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}$$

as we saw in Theorem 358 whereby $\boldsymbol{D}$ is the Jacobian of the moments at the true population parameters

$$\boldsymbol{D} \equiv \boldsymbol{D}_{\boldsymbol{\theta}}\mathbb{E}[g(w_i;\boldsymbol{\theta})]$$

and $\boldsymbol{S}$ is the variance of the moments

$$\boldsymbol{S} = \mathbb{E}[g(w_i;\boldsymbol{\theta})g(w_i;\boldsymbol{\theta})^T].$$

The question we wish to ask is what is the weight matrix $\boldsymbol{W}$ that minimises the variance $\boldsymbol{V}_{\boldsymbol{W}}$.

Theorem 365 (Optimal Weight Matrix for GMM)
Setting the weight matrix as the inverse of the population covariance matrix of the moments,

$$\boldsymbol{W} = \boldsymbol{S}^{-1},$$

minimises the variance of the GMM estimator in the sense that

$$\boldsymbol{V}_{\boldsymbol{W}} - \boldsymbol{V}_{\boldsymbol{S}^{-1}}$$

is positive semidefinite for any other weight matrix $\boldsymbol{W}$.

The variance of the GMM estimator under this weight matrix $\boldsymbol{W} = \boldsymbol{S}^{-1}$ is:

$$\boldsymbol{V}_{\boldsymbol{S}^{-1}} = (\boldsymbol{D}^T\boldsymbol{S}^{-1}\boldsymbol{D})^{-1} \tag{295}$$

To interpret what $\boldsymbol{V}_{\boldsymbol{W}} - \boldsymbol{V}_{\boldsymbol{S}^{-1}}$ is doing, the optimal weighting scheme downweights noisy linear combinations of moments and also accounts for covariance across the moments. This ensures that the GMM estimator $\hat{\boldsymbol{\theta}}_{GMM}$ is informed most strongly by moments that are estimated more precisely.

However, one issue is that we don't know what $\boldsymbol{S} = \mathbb{E}[g(w_i;\boldsymbol{\theta})g(w_i;\boldsymbol{\theta})^T]$ is in practice since we do not know the true population parameter $\boldsymbol{\theta}$. There are two ways around this.

The first way around this is to compute the **two-step GMM** estimator. In the first step, estimate $\boldsymbol{\theta}$ using any consistent initial positive-definite weight matrix, often the identity matrix. Then estimate $\boldsymbol{S}$ from the moments evaluated at this first-step estimate. Finally, re-estimate $\boldsymbol{\theta}$ using $\hat{\boldsymbol{S}}^{-1}$ as the weight matrix.

The second way is to use **continuously-updated GMM**. Here, for each conjectured value $\tilde{\boldsymbol{\theta}}$, we calculate the variance of sample moments at that conjecture $\tilde{\boldsymbol{\theta}}$, which we denote by $S_n(\tilde{\boldsymbol{\theta}})$. The continuously updated estimator is then the quadratic form that is allowed to depend on this conjecture and weight matrix now depends on $\tilde{\boldsymbol{\theta}}$.

11.5.1 Efficiency in Linear IV Model

We now analyse the efficiency of the GMM estimator in the instrumental variable framework.

Theorem 366 (Optimal Weight Matrix for Linear IV Model Under Homoskedasticity)
Under homoskedasticity, the optimal weight matrix for the linear IV model is:

$$\boldsymbol{W}_n^* = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{z}_i^T\right)^{-1} = n\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \propto \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \tag{296}$$

Proof. First, we write out our sample moment conditions:

$$\overline{g}_n(\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^n g(w_i;\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) =_{(1)} \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i u_i$$

where $=_{(1)}$ holds if $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ which is the true population parameter.

Therefore, the variance of the sample moments at the true population parameters is then given by:

$$\boldsymbol{S} = \mathbb{E}[g(w_i;\boldsymbol{\beta})g(w_i;\boldsymbol{\beta})^T] = \mathbb{E}[u_i^2\boldsymbol{z}_i\boldsymbol{z}_i^T] =_{(1)} \sigma_u^2\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T]$$

where $=_{(1)}$ requires the assumption of **homoskedasticity**.

We require an estimator of the variance $\boldsymbol{S}$, whereby we can use:

$$\hat{\boldsymbol{S}}_n \equiv \hat{\sigma}_n^2\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T\right)$$

whereby

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \hat{u}_i^2$$

is the residual-based estimator from 2SLS, which gives us a consistent estimator of σ^2.

Then, recall that the optimal weight matrix is the inverse of the variance but we now plug in our estimated variance:

$$\boldsymbol{W}_n^* = \hat{\boldsymbol{S}}_n^{-1} = \hat{\sigma}_n^{-2}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T\right)^{-1}$$

Now, notice that scaling the weight matrix $\boldsymbol{W}_n^*$ has no effect on the optimisation routine of GMM:

$$\min_{\tilde{\boldsymbol{\beta}}} \overline{g}_n(\tilde{\boldsymbol{\beta}})^T\boldsymbol{W}_n^*\overline{g}_n(\tilde{\boldsymbol{\beta}})$$

since it is only the relative importance of weights on the moments that matters, not their absolute levels. Therefore, we can drop the scalar factor $\hat{\sigma}_n^{-2}$ from the optimal weight matrix:

$$\boldsymbol{W}_n^* = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T\right)^{-1} = n\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \propto \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}$$

□

Now, recall that we have shown that in Proposition 350, such a weight matrix gives us a two-stage least squares estimator! Hence, under homoskedasticity, the 2SLS estimator is the asymptotically efficient GMM estimator.

Under **heteroskedasticity**, we have that the variance matrix

$$\boldsymbol{S} = \mathbb{E}[u_i^2 \boldsymbol{z}_i \boldsymbol{z}_i^T]$$

can be estimated by:

$$\hat{\boldsymbol{S}}_n = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{z}_i \boldsymbol{z}_i^T$$

The weight matrix is then set to be:

$$\boldsymbol{W}_n = \left(\hat{\boldsymbol{S}}_n\right)^{-1}$$

We can then compute the GMM estimator under this weight matrix.

11.6 Test For Overidentifying Restrictions

We are now interested in looking at testing overidentifying restrictions in the GMM framework. This is analogous to instrument exogeneity testing as we shall soon see.

Definition 367 (Hypothesis Test for Overidentifying Restrictions)
The hypothesis test for overidentifying restrictions is:

$$H_0 : \mathbb{E}[g(w_i; \boldsymbol{\theta})] = 0$$

against

$$H_1 : \mathbb{E}[g(w_i; \boldsymbol{\theta})] \neq 0$$

Notice that this is similar to testing instrument exogeneity:

$$H_0 : \mathbb{E}[\boldsymbol{z}_i(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})] = 0.$$

We base our test for overidentifying restrictions on the GMM criterion function we defined in Definition 345:

$$Q_n(\tilde{\boldsymbol{\theta}}) = \overline{g}_n(\tilde{\boldsymbol{\theta}})^T \boldsymbol{W}_n \overline{g}_n(\tilde{\boldsymbol{\theta}}) \tag{297}$$

whereby $\boldsymbol{W}_n$ is a positive definite $L \times L$ weight matrix and $\overline{g}_n(\tilde{\boldsymbol{\theta}})$ are the sample moment conditions. Recall that Q_n is a scalar!

Theorem 368 (Limiting Distribution for Test for Overidentifying Restriction in GMM)
Let Q_n be the criterion function. Then, if the optimal weight matrix $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{S}^{-1}$, the limiting distribution of the rescaled GMM criterion function is:

$$nQ_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{D} \chi^2(L - K) \tag{298}$$

where $L - K$ is the number of overidentified restrictions.

Proof. First, look at the sample moments $\overline{g}_n(\tilde{\boldsymbol{\theta}})$. We shall examine them under the true population parameters $\overline{g}_n(\boldsymbol{\theta})$.

Theorem 369 (Asymptotic Normality of Sample Moments at True Parameter)
The sample moments under the true population parameters $\boldsymbol{\theta}$ is asymptotically normal:

$$n^{1/2}\overline{g}_n(\boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{S}) \tag{299}$$

where $\boldsymbol{S}$ is the variance of the moments $\boldsymbol{S} = \mathbb{E}[g(w_i;\boldsymbol{\theta})g(w_i;\boldsymbol{\theta})^T]$.

Proof. The sample moment averages at the true parameter is given by:

$$\overline{g}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(w_i;\boldsymbol{\theta})$$

Clearly, by definition, the moments have population mean 0 at the true parameter

$$\mathbb{E}[g(w_i;\boldsymbol{\theta})] = \mathbf{0}$$

by the GMM-ID assumption. Therefore, we have a sample average with mean 0. By assumption, w_i are i.i.d. We can divide by $\sqrt{n}$ and invoke the **central limit theorem**

$$n^{1/2}\overline{g}_n(\boldsymbol{\theta}) = \frac{1}{n^{1/2}}\sum_{i=1}^{n} g(w_i;\boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{S})$$

whereby $\boldsymbol{S} \equiv \mathbb{E}[g(w_i;\boldsymbol{\theta})g(w_i;\boldsymbol{\theta})^T]$ is the variance-covariance matrix of the sample moments as the expectation is 0. □

We can rewrite the result we have just shown as:

$$n^{1/2}\overline{g}_n(\boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{S}) = \boldsymbol{S}^{1/2}\boldsymbol{\eta}$$

whereby $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_L)$ is a multivariate normal distribution and $\boldsymbol{S}^{1/2}$ is the square root of the variance matrix.

We now assume that we know the true value $\boldsymbol{\theta}$. Now, we multiply the GMM criterion function Eq. (297) by n and apply **Slutsky's theorem**, the GMM criterion at the true population parameter $\boldsymbol{\theta}$ is:

$$nQ_n(\boldsymbol{\theta}) = \left(\sqrt{n}\overline{g}_n(\boldsymbol{\theta})\right)^T \boldsymbol{W}_n \left(\sqrt{n}\overline{g}_n(\boldsymbol{\theta})\right) \xrightarrow{D} \boldsymbol{\eta}^T \boldsymbol{S}^{1/2}\boldsymbol{W}\boldsymbol{S}^{1/2}\boldsymbol{\eta}$$

under the null hypothesis $H_0 : \mathbb{E}[g(w_i;\boldsymbol{\theta})] = 0$.

Now, suppose that the optimal weight matrix is chosen: $\boldsymbol{W} = \boldsymbol{S}^{-1}$. Then, our limiting expression for the rescaled criterion function is:

$$nQ_n(\boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{\eta}^T \boldsymbol{S}^{1/2}\boldsymbol{S}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{\eta} = \boldsymbol{\eta}^T\boldsymbol{\eta}$$

Now, recall that $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_L)$ is a multivariate standard normal vector. Its squared Euclidean norm therefore has a χ^2 distribution:

$$\boldsymbol{\eta}^T\boldsymbol{\eta} \sim \chi^2(L)$$

However, we don't know the true value $\boldsymbol{\theta}$, so we base our test on the GMM estimator $\hat{\boldsymbol{\theta}}_{GMM}$, which is a consistent estimator for the true parameter $\boldsymbol{\theta}$ under the null H_0. First, recall that the first order condition for the GMM criterion:

$$\mathbf{0} = \nabla_{\tilde{\boldsymbol{\theta}}} Q_n(\tilde{\boldsymbol{\theta}}) = \left[D_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})\right]^T \boldsymbol{W}_n \overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})$$

This expression is a $K \times 1$ vector because $D_{\boldsymbol{\theta}}\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})$ is an $L \times K$ matrix, $\boldsymbol{W}_n$ is an $L \times L$ matrix, and $\overline{g}_n(\hat{\boldsymbol{\theta}}_{GMM})$

is an $L \times 1$ vector. By what we have shown before, $K \leq L$. Regardless of whether H_0 is true, the first-order conditions set K linear combinations of the sample moments to zero by construction.

If we have as many moments as parameters (K = L), then all moment conditions can be set to 0 and we have no overidentifying restrictions to test. When $L > K$, estimation uses K independent directions in the moments, leaving $L - K$ independent overidentifying restrictions. Therefore, the test has $L - K$ degrees of freedom. Hence,

$$nQ_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{D} \chi^2(L-K).$$

□

11.7 Hypothesis Testing

We are now interested in testing hypotheses involving linear or nonlinear restrictions.

11.7.1 Linear Testing

We have shown that the GMM estimator is asymptotically normal:

$$n^{1/2}[\hat{\boldsymbol{\theta}}_{GMM}(\boldsymbol{W}_n) - \boldsymbol{\theta}] \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{W}})$$

where the variance matrix is given by:

$$\boldsymbol{V}_{\boldsymbol{W}} = (\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}.$$

Definition 370 (Linear Restrictions)
The null hypothesis of testing linear restrictions is given by:

$$H_0 : \boldsymbol{R}\boldsymbol{\theta} = \rho \quad H_1 : \boldsymbol{R}\boldsymbol{\theta} \neq \rho$$

where $\boldsymbol{R}$ is a $r \times K$ matrix, ρ is an $r \times 1$ vector, and r is the number of restrictions, with $r \leq K$.

Under asymptotic normality, we can construct the Wald test.

Definition 371 (Wald Statistic)
The **Wald Statistic** is given by:

$$W_n = n(\boldsymbol{R}\hat{\boldsymbol{\theta}}_{GMM} - \rho)^T(\boldsymbol{R}\hat{\boldsymbol{V}}_n\boldsymbol{R}^T)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\theta}}_{GMM} - \rho). \tag{300}$$

With the Wald statistic, we can derive the asymptotic distribution needed to conduct hypothesis testing.

Theorem 372 (Wald Test)
Let W_n be the Wald Statistic. The limiting distribution of the Wald Statistic is a chi-squared distribution:

$$W_n \xrightarrow{D} \chi^2(r) \tag{301}$$

where r is the number of restrictions made in the null hypothesis.

Proof. Recall the Wald statistic

$$W_n = n(\boldsymbol{R}\hat{\boldsymbol{\theta}}_{GMM} - \rho)^T(\boldsymbol{R}\hat{\boldsymbol{V}}_n\boldsymbol{R}^T)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\theta}}_{GMM} - \rho)$$

First, by consistency, we have that the middle expression is:

$$\boldsymbol{R}\hat{\boldsymbol{V}}_n\boldsymbol{R}^T \xrightarrow{P} \boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T.$$

We now look at the *bread* of the Wald statistic and see we can rewrite it:

$$\begin{aligned} n^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\theta}}_{GMM} - \rho) &=_{(1)} n^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{R}\boldsymbol{\theta}) \\ &= n^{1/2}(\boldsymbol{R}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta})) \xrightarrow{D} \boldsymbol{R}\mathcal{N}(\boldsymbol{0}, \boldsymbol{V}) \\ &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T) \\ &= (\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T)^{1/2}\boldsymbol{\eta} \qquad \text{where } \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r) \end{aligned}$$

whereby $=_{(1)}$ holds under the null hypothesis of the linear restriction and we invoked the asymptotic normality of the GMM estimator.

So going back to our Wald Statistic, we combine our results with Slutsky's theorem to get:

$$\begin{aligned} W_n &= n(\boldsymbol{R}\hat{\boldsymbol{\theta}}_{GMM} - \rho)^T(\boldsymbol{R}\hat{\boldsymbol{V}}_n\boldsymbol{R}^T)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\theta}}_{GMM} - \rho) \\ &= \left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\theta}}_{GMM} - \rho)\right)^T(\boldsymbol{R}\hat{\boldsymbol{V}}_n\boldsymbol{R}^T)^{-1}\left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\theta}}_{GMM} - \rho)\right) \\ &\xrightarrow{D} \left[(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T)^{1/2}\boldsymbol{\eta}\right]^T\left(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T\right)^{-1}\left[(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T)^{1/2}\boldsymbol{\eta}\right] \\ &= \boldsymbol{\eta}^T(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T)^{1/2}\left(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T\right)^{-1}(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T)^{1/2}\boldsymbol{\eta} \\ &=_{(1)} \boldsymbol{\eta}^T\boldsymbol{\eta} \sim \chi^2(r) \end{aligned}$$

where $=_{(1)}$ uses the fact that the squared Euclidean norm of an r-dimensional standard normal vector has a $\chi^2(r)$ distribution. □

Therefore, for linear restrictions, we construct a Wald statistic and compare its value to a $\chi^2(r)$ distribution and reject the null hypothesis if the Wald statistic is greater than the specified critical value for the $\chi^2(r)$ distribution.

11.7.2 Nonlinear Testing

We now look at testing nonlinear restrictions. First, we change notation and let d_r be the number of restrictions we are testing and r be the restrictions themselves. Now, we assume that the restrictions $r : \Theta \to \mathbb{R}^{d_r}$ are represented by a differentiable function with the Jacobian of the restrictions defined as:

$$\boldsymbol{R} \equiv D_{\boldsymbol{\theta}}r(\boldsymbol{\theta})$$

whereby $D_{\boldsymbol{\theta}}r(\boldsymbol{\theta})$ is a $d_r \times K$ matrix having rank $d_r \leq K$.

Definition 373 (Hypothesis Test)
The hypothesis test for nonlinear restrictions are the hypotheses:

$$H_0 : r(\boldsymbol{\theta}) = \rho \quad \text{vs} \quad H_1 : r(\boldsymbol{\theta}) \neq \rho$$

We can now define the Wald statistic which shall be a quadratic form of the restrictions evaluated at the consistent estimators of the true population parameter.

Definition 374 (Wald Statistic)
The Wald Statistic under nonlinear restrictions is given by:

$$W_n = n(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho)^T (\boldsymbol{R}\hat{\boldsymbol{V}}_n\boldsymbol{R}^T)^{-1}(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho) \tag{302}$$

We now state the asymptotic distribution for the Wald statistic under nonlinear restrictions.

Theorem 375 (Limiting Distribution of Wald Statistic for Nonlinear Restrictions)
The Wald statistic under nonlinear restrictions has a limiting chi-squared distribution:

$$W_n \xrightarrow{D} \chi^2(d_r). \tag{303}$$

Proof. First, we rewrite the Wald Statistic:

$$\begin{aligned} W_n &= n(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho)^T (\boldsymbol{R}\hat{\boldsymbol{V}}_n\boldsymbol{R}^T)^{-1}(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho) \\ &= \left(n^{1/2}(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho)\right)^T (\boldsymbol{R}\hat{\boldsymbol{V}}_n\boldsymbol{R}^T)^{-1}\left(\sqrt{n}(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho)\right) \end{aligned}$$

The *meat* of our sandwich uses the fact that $\hat{\boldsymbol{V}}$ is consistent to show:

$$\boldsymbol{R}\hat{\boldsymbol{V}}_n\boldsymbol{R}^T \xrightarrow{P} \boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T.$$

We now look at the bread of our sandwich.

$$n^{1/2}(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho) = n^{1/2}(r(\hat{\boldsymbol{\theta}}_{GMM}) - r(\boldsymbol{\theta}))$$

whereby we imposed the null hypothesis $r(\boldsymbol{\theta}) = \rho$. Now, we know that the GMM estimator is asymptotically normal:

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$$

and we are interested in the distribution of a function $r(\cdot)$ of $\hat{\boldsymbol{\theta}}_{GMM}$. We can apply a matrix version of the **delta method** to derive the limiting distribution of $r(\hat{\boldsymbol{\theta}}_{GMM})$ as:

$$n^{1/2}(r(\hat{\boldsymbol{\theta}}_{GMM}) - r(\boldsymbol{\theta})) \xrightarrow{D} \left[D_{\boldsymbol{\theta}}r(\boldsymbol{\theta})\right]\mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$$

However, by definition, we said that:

$$\boldsymbol{R} \equiv D_{\boldsymbol{\theta}}r(\boldsymbol{\theta})$$

so we have that:

$$\left[D_{\boldsymbol{\theta}}r(\boldsymbol{\theta})\right]\mathcal{N}(\boldsymbol{0}, \boldsymbol{V}) = \boldsymbol{R}\mathcal{N}(\boldsymbol{0}, \boldsymbol{V}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T) = (\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T)^{1/2}\boldsymbol{\eta}$$

whereby $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_r})$. We can combine our two results and apply Slutsky's theorem to conclude that:

$$
\begin{aligned}
W_n &= \left(n^{1/2}(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho)\right)^T (\boldsymbol{R}\hat{\boldsymbol{V}}_n\boldsymbol{R}^T)^{-1}\left(\sqrt{n}(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho)\right) \\
&\xrightarrow{D} \left((\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T)^{1/2}\boldsymbol{\eta}\right)^T \left(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T\right)^{-1}\left((\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^T)^{1/2}\boldsymbol{\eta}\right) \\
&= \boldsymbol{\eta}^T\boldsymbol{\eta} \sim \chi^2(d_r)
\end{aligned}
$$

whereby we used the fact that the squared Euclidean norm of a d_r-dimensional standard normal vector has a chi-squared distribution. □

In practice, we don't know what $\boldsymbol{R}$ is as it can vary with the true population value $\boldsymbol{\theta}$. Therefore, we can construct a consistent estimator of $\boldsymbol{R}$ with:

$$\hat{\boldsymbol{R}}_n = \boldsymbol{D}_{\boldsymbol{\theta}} r(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{P} \boldsymbol{D}_{\boldsymbol{\theta}} r(\boldsymbol{\theta}) = \boldsymbol{R}$$

using the GMM estimator $\hat{\boldsymbol{\theta}}_{GMM}$. From this, we can plug this into our Wald Statistic:

$$W_n = n(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho)^T(\hat{\boldsymbol{R}}_n\hat{\boldsymbol{V}}_n\hat{\boldsymbol{R}}_n^T)^{-1}(r(\hat{\boldsymbol{\theta}}_{GMM}) - \rho)$$

and this can be shown to still converge to an asymptotic χ^2 distribution.

11.8 Criterion-Based Test

There are two issues with the Wald statistic. The first is that it is not invariant to equivalent representations of the restrictions, so restrictions that are essentially equivalent can give drastically different test results. Second, the Jacobian of the restrictions may fail to have full rank. The criterion-based test addresses the invariance issue, although a rank failure can still invalidate the usual chi-squared approximation.

We can construct something called the **QLR test** which is similar to the likelihood ratio test.

First, define the null hypotheses:

$$H_0 : r(\boldsymbol{\theta}) = \rho \qquad H_1 : r(\boldsymbol{\theta}) \neq \rho$$

Recall that the **GMM criterion** is given by:

$$Q_n(\tilde{\boldsymbol{\theta}}) \equiv \overline{g}_n(\tilde{\boldsymbol{\theta}})^T\hat{\boldsymbol{S}}_n^{-1}\overline{g}_n(\tilde{\boldsymbol{\theta}}).$$

The QLR test works by evaluating the fit of the model under the null hypothesis and not under the null hypothesis. That is, we compare the criterion function values between the restricted and unrestricted model.

Definition 376 (QLR Test)

Define a subset of the parameter space:

$$\Theta_p \equiv \{\boldsymbol{\theta} \in \Theta : r(\boldsymbol{\theta}) = \rho\}$$

as the parameters that satisfy the null hypothesis. Then, the QLR statistic is given by:

$$QLR_n \equiv n\left[\min_{\tilde{\boldsymbol{\theta}}\in\Theta_p} Q_n(\tilde{\boldsymbol{\theta}}) - \min_{\tilde{\boldsymbol{\theta}}\in\Theta} Q_n(\tilde{\boldsymbol{\theta}})\right] \geq 0$$

where the first term is the criterion function under the restricted model and the second term is under the unrestricted model.

We can also derive the limiting distribution of this test.

Proposition 377 (Limiting Distribution of QLR Test)
The limiting distribution of the QLR test is:

$$QLR_n \xrightarrow{D} \chi^2(d_r) \tag{304}$$

where d_r is the number of restrictions.

Whilst the Wald statistic and QLR test statistic are asymptotically equivalent under standard regularity conditions, the QLR test is often preferable because it is invariant to equivalent reparameterisations of the restrictions.

12 Time Series

12.1 Introduction

The big issue in time series data is that observations and error terms may be serially dependent, unlike in the simplest cross-sectional setting. Lagged variables can help us model temporal dependence in the conditional mean, whilst the remaining dependence in the errors must also be handled by the specification or by appropriate inference.

12.2 Cross Sectional Data Review

We now review concepts we have previously seen in cross-sectional data.

12.2.1 Ordinary Least Squares

We first revisit regressions that we saw in the cross sectional case. In particular, it is very important to understand how the Frisch-Waugh-Lovell theorem works as that is the key behind how other concepts in time series operate.

We assume a statistical model in matrix notation given by:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}.$$

The **OLS estimator** is given by:

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

whereby we require the assumption that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.

The **least squares fitted values** is given by:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$

whereby we defined the **projection matrix**:

$$\boldsymbol{P} \equiv \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

The **least squares residuals** is given by:

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y}$$

where we defined the **annihilator matrix**:

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}.$$

We have that $\boldsymbol{P}, \boldsymbol{M}$ are idempotent and symmetric matrices, and are orthogonal to each other. The **residual variance** can be estimated by:

$$s^2 = \frac{1}{n-k}\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}.$$

12.2.2 Maximum Likelihood Estimation

We now revisit maximum likelihood estimation which we will also need. We state the maximised likelihood function.

Proposition 378 (Likelihood Function under Normality)

Suppose that y_i, x_i are i.i.d and $u_i \sim \mathcal{N}(0, \sigma^2)$. Then, the likelihood function is given by:

$$\hat{\ell} = -\frac{n}{2}\log\left(2\pi e\hat{\sigma}^2\right) \tag{305}$$

whereby:

$$\hat{\sigma}^2 = \frac{1}{n}\hat{u}^T\hat{u}.$$

Proof. The log-likelihood function is given by:

$$\ell_{\boldsymbol{y}|\boldsymbol{X}}(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 \tag{306}$$

The maximum likelihood estimation for $\boldsymbol{\beta}$ clearly involves minimising the second term as that is the term that depends on $\boldsymbol{\beta}$. Doing so gives us:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

The estimate for the variance is:

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}.$$

We plug these back into Eq. (306):

$$\begin{aligned}
\hat{\ell} &= -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(\hat{u}_i)^2 \\
&= -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}n\hat{\sigma}^2 \\
&= -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2} \\
&= -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}\log(e) \\
&= -\frac{n}{2}\left[\log(2\pi\hat{\sigma}^2) + \log(e)\right] \\
&= -\frac{n}{2}\log\left(2\pi e\hat{\sigma}^2\right).
\end{aligned}$$

□

This is useful because we now have an analytical expression for the likelihood ratio test. Suppose we had the model:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$$

and we wish to test $\beta_2 = 0$. We can conduct a log-likelihood ratio test instead of a t-test.

Proposition 379 (Log-Likelihood Ratio Test)

The log-likelihood ratio test is given by:

$$LR_{\boldsymbol{\beta}_2=\boldsymbol{0}} = n\log\left(\frac{\hat{\sigma}^2_{Restricted}}{\hat{\sigma}^2_{Unrestricted}}\right) \geq 0.$$

Proof. The log-likelihood ratio test is given by:

$$LR_{\boldsymbol{\beta}_2=\mathbf{0}} = -2\left\{ \max_{\boldsymbol{\beta}_1,\sigma^2} \ell(\boldsymbol{\beta}_1, \mathbf{0}, \sigma^2) - \max_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\sigma^2} \ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2) \right\} \geq 0$$

We then can just plug in our expression for the maximised log-likelihood $\hat{\ell} = -\frac{n}{2}\log\left(2\pi e\hat{\sigma}^2\right)$ to get the desired result. □

12.2.3 Partial Least Squares

We are now interested in **partial least squares estimators**. The first major result we state is the Frisch-Waugh-Lovell theorem.

Theorem 380 (Frisch-Waugh-Lovell Theorem)
Suppose we had the model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

Then suppose we partitioned the design matrix:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix}$$

whereby $\boldsymbol{X}_1$ is a $n \times k_1$ matrix and $\boldsymbol{X}_2$ is a $n \times k_2$ matrix. Likewise, partition the coefficients:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

whereby $\boldsymbol{\beta}_1$ is a $k_1 \times 1$ vector and $\boldsymbol{\beta}_2$ is a $k_2 \times 1$ vector.

The **Frisch-Waugh-Lovell theorem** says that:

$$\hat{\boldsymbol{\beta}}_2 = (\hat{\boldsymbol{u}}_{2\cdot1}^T \hat{\boldsymbol{u}}_{2\cdot1})^{-1} \hat{\boldsymbol{u}}_{2\cdot1}^T \hat{\boldsymbol{u}}_{\boldsymbol{y}\cdot1}.$$

where $\hat{\boldsymbol{u}}_{\boldsymbol{y}\cdot1}$ are the residuals from regressing $\boldsymbol{y}$ on $\boldsymbol{X}_1$ and $\hat{\boldsymbol{u}}_{2\cdot1}$ are the residuals from regression $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$.

Proof. First, let us partition our design matrix:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix}$$

whereby $\boldsymbol{X}_1$ is a $n \times k_1$ matrix and $\boldsymbol{X}_2$ is a $n \times k_2$ matrix. Likewise, we partition the coefficients:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

whereby $\boldsymbol{\beta}_1$ is a $k_1 \times 1$ vector and $\boldsymbol{\beta}_2$ is a $k_2 \times 1$ vector. The least squares estimator is given by:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

We are interested in asking the question on whether can we retrieve the last k_2 entries of $\hat{\boldsymbol{\beta}}$ directly. The answer is yes we can by the **Frisch-Waugh-Lovell theorem**.

The first step is to regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$ and retrieve the residuals. This is equivalent to applying the annihilator matrix

$M_{\boldsymbol{X}_1}$ onto $\boldsymbol{y}$:

$$\hat{\boldsymbol{u}}_{\boldsymbol{y}\cdot 1} = \left[\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\right]\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y}.$$

The second step is to regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$ and retrieve the residuals:

$$\hat{\boldsymbol{u}}_{2\cdot 1} = \left[\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\right]\boldsymbol{X}_2 = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{X}_2 = \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2.$$

The final step is to regress the residuals from the first step $\hat{\boldsymbol{u}}_{\boldsymbol{y}\cdot 1}$ on the residuals from the second step $\hat{\boldsymbol{u}}_{2\cdot 1}$ and this gives us the OLS estimator:

$$\hat{\boldsymbol{\beta}}_2 = (\hat{\boldsymbol{u}}_{2\cdot 1}^T\hat{\boldsymbol{u}}_{2\cdot 1})^{-1}\hat{\boldsymbol{u}}_{2\cdot 1}^T\hat{\boldsymbol{u}}_{\boldsymbol{y}\cdot 1}.$$

This final equation is known as the **Frisch-Waugh-Lovell** theorem. □

We now look at a special case whereby we have a single explanatory variable (or two if you count the constant term on the intercept).

Theorem 381 (OLS with a Single Explanatory Variable)
Suppose we had the model:

$$y_i = \beta_1 + \beta_2 x_i + u_i \quad \text{for } i = 1, ..., n.$$

Then, by the Frisch-Waugh-Lovell theorem,

$$\hat{\boldsymbol{\beta}}_{FWL} = \frac{\sum_{i=1}^n (x_i - \overline{x})y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{307}$$

Proof. We are interested on how we can obtain an estimator of β_2 using a similar procedure as the Frisch-Waugh-Lovell theorem. We shall denote $\mathbf{1} = \begin{bmatrix} 1 & 1... & 1 \end{bmatrix}$ to be a $n \times 1$ matrix of 1's.

There are now four steps.

The first step is to regress $\boldsymbol{y}$ on the intercept term:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\boldsymbol{y}\cdot\mathbf{1}} &= (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\boldsymbol{y} \\
&= (n)^{-1}\mathbf{1}^T\boldsymbol{y} \\
&= (n)^{-1}\sum_{i=1}^n y_i \\
&= \overline{y}.
\end{aligned}$$

The second step is to regress $\boldsymbol{X}$ on the intercept term too:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\boldsymbol{X}\cdot\mathbf{1}} &= (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\boldsymbol{X} \\
&= (n)^{-1}\mathbf{1}^T\boldsymbol{X} \\
&= (n)^{-1}\sum_{i=1}^n x_i \\
&= \overline{\boldsymbol{x}}.
\end{aligned}$$

The third step is to retrieve the residuals (observed minus fitted value) from step one and two so we have that the

individual observation is:

$$\hat{u}_{\boldsymbol{y}\cdot 1,i} = y_i - \overline{y} \cdot 1 = y_i - \overline{y}$$
$$\hat{u}_{x\cdot 1,i} = x_i - \overline{x} \cdot 1 = x_i - \overline{x}$$

We can stack these residuals into vectors. The final step is then to regress the first set of residuals $\hat{\boldsymbol{u}}_{y\cdot 1}$ on the other set of residuals $\hat{\boldsymbol{u}}_{x\cdot 1}$ to get:

$$\hat{\boldsymbol{\beta}}_{FWL} = (\hat{\boldsymbol{u}}_{x\cdot 1}^T \hat{\boldsymbol{u}}_{x\cdot 1})^{-1} \hat{\boldsymbol{u}}_{x\cdot 1}^T \hat{\boldsymbol{u}}_{y\cdot 1} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} =_{(1)} \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

whereby $=_{(1)}$ holds by:

$$\begin{aligned}
\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^n (x_i - \overline{x}) y_i - \sum_{i=1}^n (x_i - \overline{x})\overline{y} \\
&= \sum_{i=1}^n (x_i - \overline{x}) y_i - \overline{y} \sum_{i=1}^n (x_i - \overline{x}) \\
&= \sum_{i=1}^n (x_i - \overline{x}) y_i - \overline{y}(n\overline{x} - n\overline{x}) \\
&= \sum_{i=1}^n (x_i - \overline{x}) y_i - 0
\end{aligned}$$

□

12.2.4 Correlation

Now, we revisit the concept of **correlation**. Recall that the population correlation is given by:

$$Corr[X, Y] = \frac{Cov[X, Y]}{\sqrt{Var[X] Var[Y]}}$$

We can now define the sample analogue.

Definition 382 (Sample Correlation)
The sample correlation between y_i and x_i is given by:

$$r_{y,x} = Corr(y_i, x_i) = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2}} \tag{308}$$

We are now interested in **sample partial correlations**. Consider the partitioned equation again:

$$\boldsymbol{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$$

whereby $\boldsymbol{X}_2$ now has dimension $k_2 = 1$.

Definition 383 (Partial Correlation for one dimensional regressor)

The sample partial correlation is given by:

$$Corr(y_i, x_{2,i}|x_{1,i}) = \frac{\hat{\boldsymbol{u}}_{2\cdot1}^T \hat{\boldsymbol{u}}_{\boldsymbol{y}\cdot1}}{\sqrt{(\hat{\boldsymbol{u}}_{2\cdot1}^T \hat{\boldsymbol{u}}_{2\cdot1})(\hat{\boldsymbol{u}}_{y\cdot1}^T \hat{\boldsymbol{u}}_{y\cdot1})}}. \tag{309}$$

whereby $\hat{\boldsymbol{u}}_{2\cdot1}$ is the residual vector from regressing $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$ and $\hat{\boldsymbol{u}}_{y\cdot1}$ is the residual vector from regressing $\boldsymbol{y}$ on $\boldsymbol{X}_1$.

This definition of partial correlation follows a similar argument to the Frisch-Waugh-Lovell theorem whereby we are partialling out the effects of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$ and $\boldsymbol{y}$ respectively and looking at the correlation between them.

Remark 384. *The sample correlation Eq.* (308) *is the partial correlation whereby we partial out the constant term* $\mathbf{1}$ *as* $\hat{u}_{y\cdot1} = y - \overline{y}$ *and* $\hat{u}_{x\cdot1} = x - \overline{x}$,

$$r_{y,x} = Corr(y_i, x_i) = Corr(y_i, x_i|1).$$

We can extend our definitions so far to looking at the correlation of $\boldsymbol{y}$ with a vector of explanatory variables $\boldsymbol{X}$.

Definition 385 (Sample Multiple Correlation)

Suppose that $\boldsymbol{x}$ is now a vector of explanatory variables. Then, the sample multiple correlation is:

$$Corr(\boldsymbol{y}, \boldsymbol{X}) = \left[\frac{\boldsymbol{y}^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}}\right]^{1/2} \tag{310}$$

Definition 386 (Sample Multiple Partial Correlations)

Suppose that $\boldsymbol{x}$ is now a vector of explanatory variables. Then, the sample multiple correlation is:

$$Corr(y_i, x_{2,i}|x_{1,i}) = \left[\frac{\hat{\boldsymbol{u}}_{\boldsymbol{y}\cdot1}^T \hat{\boldsymbol{u}}_{2\cdot1}\left(\hat{\boldsymbol{u}}_{2\cdot1}^T \hat{\boldsymbol{u}}_{2\cdot1}\right)^{-1} \hat{\boldsymbol{u}}_{2\cdot1}^T \hat{\boldsymbol{u}}_{\boldsymbol{y}\cdot1}}{\hat{\boldsymbol{u}}_{y\cdot1}^T \hat{\boldsymbol{u}}_{y\cdot1}}\right]^{1/2}. \tag{311}$$

Proposition 387 (Coefficient of Determination)

The coefficient of determination R^2 is the squared sample multiple partial correlation after partialling out the intercept. That is, for the model with an intercept:

$$\boldsymbol{y} = \boldsymbol{\beta}_1 + \boldsymbol{X}\boldsymbol{\beta}_2 + \boldsymbol{u}$$

we have that:

$$R^2 = Corr(y_i, \boldsymbol{x}_i|1)^2 = \frac{\hat{\boldsymbol{u}}_{\boldsymbol{y}\cdot1}^T \hat{\boldsymbol{u}}_{\boldsymbol{X}\cdot1}\left(\hat{\boldsymbol{u}}_{\boldsymbol{X}\cdot1}^T \hat{\boldsymbol{u}}_{\boldsymbol{X}\cdot1}\right)^{-1} \hat{\boldsymbol{u}}_{\boldsymbol{X}\cdot1}^T \hat{\boldsymbol{u}}_{\boldsymbol{y}\cdot1}}{\hat{\boldsymbol{u}}_{y\cdot1}^T \hat{\boldsymbol{u}}_{y\cdot1}} = \frac{ESS}{TSS}. \tag{312}$$

12.3 Autocorrelation

We move back to time series data. We want to capture the temporal dependence we observe in the data. To do so, we can modify the formula for sample correlation found in Eq. (308).

Definition 388 (Sample Autocorrelation)

For a fixed s, the sample autocorrelation is given by:

$$r_{0,s} = \frac{\sum_{t=s+1}^{T}(Y_t - \overline{Y}_{s+1}^{T})(Y_{t-s} - \overline{Y}_{1}^{T-s})}{\sqrt{\sum_{t=s+1}^{T}(Y_t - \overline{Y}_{s+1}^{T})^2 \sum_{t=s+1}^{T}(Y_{t-s} - \overline{Y}_{1}^{T-s})^2}} \tag{313}$$

whereby $\overline{Y}_a^b$ is the average of $Y_a, ..., Y_b$.

Remark 389. *The first bracket in the numerator looks the variation from time s + 1 up to time T. The second bracket in the numerator looks at the variation from time 1 to time T − s. Therefore, the first pair of observations is Y_{1+s} and Y_1.*

One can show that the sample autocorrelation is in fact asymptotically normal. However, we currently do not have all the techniques needed to show this.

Proposition 390 (Asymptotic Normality of Sample Autocorrelation under i.i.d)

Assume mean-zero i.i.d data and ignore mean correction. Then, the sample autocorrelation is given by:

$$r_{0,s} = \frac{\sum_{t=s+1}^{T} Y_t Y_{t-s}}{\sqrt{\sum_{t=s+1}^{T} Y_t^2 \sum_{t=s+1}^{T} Y_{t-s}^2}}.$$

Under proper rescaling, the sample autocorrelation is asymptotically normal:

$$\sqrt{T} r_{0,s} = \frac{T^{-1/2} \sum_{t=s+1}^{T} Y_t Y_{t-s}}{\sqrt{\left(T^{-1} \sum_{t=s+1}^{T} Y_t^2\right)\left(T^{-1} \sum_{t=s+1}^{T} Y_{t-s}^2\right)}} \xrightarrow{D} \mathcal{N}(0,1).$$

Proof. (Sketch). We can apply a weak law of large numbers for the second moment in the denominator to both terms respectively. However, for the numerator, we have that the product $Y_t \cdot Y_{t-s}$ is a temporal sequence as $Y_t \cdot Y_{t-s}$ can share observations with other products in the sum. We will later establish that the scaled numerator converges to a normal distribution by a modified central limit theorem. Therefore, we can conclude from Slutsky's theorem that the scaled sample autocorrelation of an i.i.d sequence is asymptotically normal. □

Remark 391. *The issue in our proof is that we do not yet have a central limit theorem for dependent variables. We will work our way towards this in this course.*

We now define our first stochastic process.

Definition 392 (Moving Average of Order 1)

A time series Y_t is a **moving average of order one** if:

$$Y_t = \epsilon_t + \theta \epsilon_{t-1} \tag{314}$$

for ϵ_t that are i.i.d with $\mathbb{E}[\epsilon_t] = 0$ and $Var[\epsilon_t] = \sigma^2$.

We state an important property that we will make use of repeatedly.

Theorem 393 (Property of Covariance)
The covariance operator has the property that for random variables X, Y, W, V:

$$Cov(aX + bY, cW + dV) = acCov(X, W) + adCov(X, V) + bcCov(Y, W) + bdCov(Y, V). \tag{315}$$

for constants a, b, c, d.

We can now state the variance and covariance of a MA(1) process.

Proposition 394 (Moments of a MA(1) Process)
Let Y_t be a MA(1) process. Then, we have that:

$$\mathbb{E}[Y_t] = 0 \tag{316}$$
$$Var(Y_t) = (1 + \theta^2)\sigma^2 \tag{317}$$
$$Cov(Y_t, Y_{t-1}) = \theta\sigma^2 \tag{318}$$
$$Cov(Y_t, Y_{t-j}) = 0 \quad \text{for } j \geq 2 \tag{319}$$

Proof. We repeatedly make use of the fact that ϵ_t are i.i.d.

$$\mathbb{E}[Y_t] = \mathbb{E}[\epsilon_t + \theta\epsilon_{t-1}] = 0.$$

$$Var[Y_t] = \mathbb{E}[Y_t^2] - \mathbb{E}[Y_t]^2 = \mathbb{E}[(\epsilon_t + \theta\epsilon_{t-1})^2] - 0 = \mathbb{E}[\epsilon_t^2] + 2\theta\mathbb{E}[\epsilon_t\epsilon_{t-1}] + \mathbb{E}[\theta^2\epsilon_{t-1}^2] = \sigma^2 + 0 + \theta^2\sigma^2.$$

$$\begin{aligned}
Cov(Y_t, Y_{t-1}) &= \mathbb{E}[Y_t Y_{t-1}] \\
&= \mathbb{E}[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2})] \\
&= \mathbb{E}[\epsilon_t\epsilon_{t-1}] + \theta\mathbb{E}[\epsilon_t\epsilon_{t-2}] + \theta\mathbb{E}[\epsilon_{t-1}^2] + \theta^2\mathbb{E}[\epsilon_{t-1}\epsilon_{t-2}] \\
&= 0 + \theta\mathbb{E}[\epsilon_{t-1}\epsilon_{t-1}] + 0 + 0 \\
&= \theta Var(\epsilon_{t-1}) \\
&= \theta\sigma^2
\end{aligned}$$

$$Cov(Y_t, Y_{t-j}) = Cov(\epsilon_t + \theta\epsilon_{t-1}, \epsilon_{t-j} + \theta\epsilon_{t-j-1}) = 0 \quad \text{for } j \geq 2$$

because the two sets of innovations are disjoint and independent when $j \geq 2$. □

Proposition 395 (Sample Autocorrelation for MA(1) Process)
The sample autocorrelation of a MA(1) process converges to the true population autocorrelation:

$$r_{0,1} \xrightarrow{P} \frac{\theta}{(1 + \theta^2)}. \tag{320}$$

Proof. (Sketch). We first compute the population autocorrelation:

$$\begin{aligned} Corr(Y_t, Y_{t-1}) &= \frac{Cov(Y_t, Y_{t-1})}{Var(Y_t)} \\ &= \frac{\theta Var(\epsilon_{t-1})}{(1+\theta^2)Var(\epsilon_{t-1})} \\ &= \frac{\theta}{(1+\theta^2)}. \end{aligned}$$

We will later argue using a modified law of large numbers that the sample autocorrelation converges to the true population autocorrelation. □

Corollary 396

Higher-lag population autocorrelations for a MA(1) process are 0:

$$Corr(Y_t, Y_{t-j}) = 0 \quad \text{for } j \geq 2. \tag{321}$$

Proof. For $j \geq 2$, Y_t and Y_{t-j} depend on disjoint sets of independent innovations. Therefore their correlation is zero. □

Definition 397 (Moving Average Of Order q)

A stochastic process Y_t is a moving average of order q if:

$$Y_t = \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{322}$$

for ϵ_t that are i.i.d.

Proposition 398 (Autocorrelation of MA(Q) Process)

Let Y_t be a moving average of order q. Then, the autocorrelation with all variables past the qth lag is equal to 0:

$$Corr(Y_t, Y_{t-s}) = 0$$

for $s \geq q+1$.

12.4 Partial Autocorrelation

Now, suppose that we have our time series Y_t and this time, we want to find the **partial correlation** between Y_t and Y_{t-s} whilst partialling out all the Y_t's in between Y_t and Y_{t-s}. These values will include $1, Y_{t-1}, Y_{t-2}, ..., Y_{t-s+1}$.

Definition 399 (Partial Autocorrelation)

The partial autocorrelation between Y_t and Y_{t-s} is given by:

$$Corr(Y_t, Y_{t-s}|1, Y_{t-1}, ..., Y_{t-s+1}) = r_{0,s\cdot 1,...,s-1} = \frac{\hat{u}_{0|1,...,s-1}^T \hat{u}_{s|1,...,s-1}}{\sqrt{\left(\hat{u}_{0|1,...,s-1}^T \hat{u}_{0|1,...,s-1}\right) \cdot \left(\hat{u}_{s|1,...,s-1}^T \hat{u}_{s|1,...,s-1}\right)}} \tag{323}$$

Remark 400. *A partial correlogram simply partials out the effects of all lags in between our values of interest.*

We can state useful properties of the partial autocorrelation.

Proposition 401 (Asymptotic Normality of Partial Autocorrelation)
Under the i.i.d white-noise null and for a fixed lag s, the partial autocorrelation is asymptotically normal.

$$T^{1/2} r_{0,s \cdot 1,\ldots,s-1} \xrightarrow{D} \mathcal{N}(0,1). \tag{324}$$

We now introduce another class of stochastic processes called autoregressive models.

Definition 402 (Autoregressive Model)
A time series Y_t is an autoregressive model of order 1 if:

$$Y_t = \alpha Y_{t-1} + \epsilon_t \tag{325}$$

for i.i.d ϵ_t.

Proposition 403 (Autocorrelation of AR(1) Process)
Let Y_t be a covariance-stationary AR(1) process with $|\alpha| < 1$. Then, the population autocorrelation is given by:

$$Corr(Y_t, Y_{t-s}) = \alpha^s \quad \text{for s } \geq 1. \tag{326}$$

Proof. First, we compute the population autocovariance for one lag s = 1:

$$\begin{aligned} Cov(Y_t, Y_{t-1}) &= Cov(\alpha Y_{t-1} + \epsilon_t, Y_{t-1}) \\ &= Cov(\alpha Y_{t-1}, Y_{t-1}) \\ &= \alpha Cov(Y_{t-1}, Y_{t-1}) \\ &= \alpha Var(Y_{t-1}) \end{aligned}$$

Then, we have that:

$$Corr(Y_t, Y_{t-1}) = \frac{Cov(Y_t, Y_{t-1})}{Var(Y_{t-1})} = \alpha.$$

By induction, this holds for:

$$Corr(Y_t, Y_{t-s}) = \alpha^s.$$

□

So for $|\alpha| < 1$ we have that the autocorrelation for an AR(1) process goes to 0 as $s \to \infty$.

Proposition 404 (Population Partial Autocorrelation of AR(1) Process)
The population partial autocorrelation of a covariance-stationary AR(1) process is 0:

$$Corr(Y_t, Y_{t-s} | Y_{t-1}, \ldots, Y_{t-s+1}) = 0 \tag{327}$$

for $s \geq 2$.

Proof. Look at the AR(1) process:

$$Y_t = \alpha Y_{t-1} + \epsilon_t.$$

Y_t only depends on its past through Y_{t-1}. This means that:

$$f(Y_t|Y_{t-1}, Y_{t-2}) = f(Y_t|Y_{t-1}).$$

Then, we have that the conditional density is:

$$\begin{aligned} f(Y_t, Y_{t-2}|Y_{t-1}) &= \frac{f(Y_t, Y_{t-1}, Y_{t-2})}{f(Y_{t-1})} \\ &= \frac{f(Y_t, Y_{t-1}, Y_{t-2})}{f(Y_{t-1}, Y_{t-2})} \times \frac{f(Y_{t-1}, Y_{t-2})}{f(Y_{t-1})} \\ &= f(Y_t|Y_{t-1}, Y_{t-2}) \times f(Y_{t-2}|Y_{t-1}) \\ &= f(Y_t|Y_{t-1}) \times f(Y_{t-2}|Y_{t-1}) \end{aligned}$$

Hence, we have shown that:

$$f(Y_t, Y_{t-2}|Y_{t-1}) = f(Y_t|Y_{t-1}) \times f(Y_{t-2}|Y_{t-1})$$

and therefore we have shown that Y_t, Y_{t-2} are independent conditioned on Y_{t-1}. This implies that the partial autocorrelation is 0:

$$Corr(Y_t, Y_{t-s}|Y_{t-1}, ..., Y_{t-s+1}) = 0.$$

□

Once we have partialled out the effects of Y_{t-1}, then we have that Y_t and Y_{t-2} are uncorrelated. This implies that the sample partial autocorrelation function of an AR(1) process should go to 0 by weak law of large numbers which we shall establish later.

To summarise the results for stochastic processes and correlations:

1. A cutoff in the ACF after q lags suggests an MA(q) series.
2. A cutoff in the PACF after p lags indicates AR(p) series.

12.5 Linear Projection in Time Series

Assumption 405 (Stationarity)
For this chapter, we assume that the time series Y_t is covariance stationary.

Cross Sectional Data Projection

First, we recall projections in cross-sectional data.

A linear predictor for $\boldsymbol{Y}$ is a function $\boldsymbol{X}^T\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$. The **mean squared prediction error** is:

$$S(\boldsymbol{\beta}) = \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{\beta})^2]. \tag{328}$$

The **best linear predictor** of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is:

$$\mathcal{P}[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{X}^T\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ minimises the mean squared prediction error:

$$S(\boldsymbol{\beta}) = \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{\beta})^2].$$

Such a $\boldsymbol{\beta}$ is called the **linear projection coefficient**.

The mean squared prediction error is:

$$S(\boldsymbol{\beta}) = \mathbb{E}[\boldsymbol{Y}^2] - 2\boldsymbol{\beta}^T\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}] + \boldsymbol{\beta}^T\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\boldsymbol{\beta}.$$

Take first order conditions with respect to $\boldsymbol{\beta}$ to get:

$$-2\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}] + 2\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\boldsymbol{\beta} = \boldsymbol{0}.$$

We then solve to get:

$$\boldsymbol{\beta} = (\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T])^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}]. \tag{329}$$

Go back to our expression for best linear predictor and plug in the linear projection coefficient:

$$\mathcal{P}[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{X}^T\boldsymbol{\beta} = \boldsymbol{X}^T(\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T])^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}].$$

This is the **linear projection** of $\boldsymbol{Y}$ onto $\boldsymbol{X}$.

The **projection error** is given by:

$$\epsilon = \boldsymbol{Y} - \mathcal{P}[\boldsymbol{Y}|\boldsymbol{X}].$$

Time Series Data Projection

We now take this idea and think about projecting Y_t onto the closed linear span of its past:

$$\mathcal{H}_{t-1} = \overline{\text{span}}\{1, Y_{t-1}, Y_{t-2}, ...\}.$$

We write the projection of Y_t onto $\mathcal{H}_{t-1}$ as:

$$\mathcal{P}[Y_t|\mathcal{H}_{t-1}]$$

This projection is unique and has a unique projection error (by the Hilbert projection theorem)

$$\epsilon_t = Y_t - \mathcal{P}[Y_t|\mathcal{H}_{t-1}].$$

$$\mathbb{E}[\epsilon_t] = 0$$
$$\mathbb{E}[\epsilon_t^2] = \sigma^2 < \infty$$

ϵ_t is serially uncorrelated.

If Y_t is strictly stationary, then $\mathcal{P}[Y_t|\mathcal{H}_{t-1}]$ and ϵ_t are strictly stationary too.

Theorem 406 (Covariance Stationary Process Have White Noise Projection Error)

If Y_t is a covariance-stationary process, then it has the projection equation:

$$Y_t = \mathcal{P}[Y_t|\mathcal{H}_{t-1}] + \epsilon_t \tag{330}$$

whereby the projection error ϵ_t satisfies:

$$\mathbb{E}[\epsilon_t] = 0$$
$$\mathbb{E}[\epsilon_{t-j}\epsilon_t] = 0 \quad j \geq 1$$
$$\mathbb{E}[\epsilon_t^2] = \sigma^2 < \infty.$$

If Y_t is strictly-stationary, then ϵ_t is strictly stationary.

Definition 407 (White Noise)

The process ϵ_t is a **white noise** if:

$$\mathbb{E}[\epsilon_t] = 0$$
$$\mathbb{E}[\epsilon_{t-j}\epsilon_t] = 0 \quad j \geq 1$$
$$\mathbb{E}[\epsilon_t^2] = \sigma^2 < \infty.$$

Proposition 408 (MDS is a White Noise)

A MDS with a finite, time-invariant second moment $\mathbb{E}[m_t^2] = \sigma^2$ is a white noise.

Remark 409. *The converse does not hold. A white noise is not necessarily a MDS.*

We list the order of processes from most restrictive to most general:

1. i.i.d sequence.
2. MDS.
3. White noise.

We have shown that covariance-stationary processes have a white noise projection error. We can extend upon this to an infinite number of projection errors.

Theorem 410 (Wold Decomposition)

Suppose that Y_t is a purely nondeterministic covariance-stationary process and $\mathbb{E}[\epsilon_t^2] = \sigma^2 < \infty$ where ϵ_t is the projection error. Then, the **Wold decomposition** of Y_t says that Y_t has the **linear representation**:

$$Y_t = \mu + \sum_{j=0}^{\infty} b_j \epsilon_{t-j} \tag{331}$$

whereby ϵ_{t-j} are the white noise projection errors:

$$\epsilon_{t-j} = Y_{t-j} - \mathcal{P}[Y_{t-j}|\mathcal{H}_{t-j-1}]$$

with

$$b_0 = 1, \qquad \sum_{j=0}^{\infty} b_j^2 < \infty$$

and μ being the unconditional mean of Y_t.

Remark 411. *For a general covariance-stationary process, the Wold decomposition may also contain a deterministic component given by the projection onto the remote past. The purely nondeterministic assumption rules out this additional component after demeaning.*

The Wold decomposition says that a purely nondeterministic covariance-stationary process can be represented as a linear function of current and lagged projection errors. Here, μ is the **unconditional mean** of Y_t. This justifies the use of linear models as approximations to covariance-stationary processes.

We now define a useful algebraic construct for analysis in time series.

Definition 412 (Lag Operator)

The **lag operator** L satisfies:

$$L^k Y_t = Y_{t-k}. \tag{332}$$

We use the lag operator to compactly write the Wold representation.

Proposition 413 (Compact Form of Wold Representation)

The compact form of the Wold representation for a covariance stationary process is:

$$Y_t = \mu + b(L)\epsilon_t \tag{333}$$

where $b(L) = b_0 + b_1 L + b_2 L^2 + ...$ is an infinite-order polynomial.

Proof. The Wold decomposition can then be written using the lag operator L as:

$$\begin{aligned} Y_t &= \mu + \sum_{j=0}^{\infty} b_j \epsilon_{t-j} \\ &= \mu + b_0\epsilon_t + b_1 L\epsilon_t + b_2 L^2\epsilon_t + ... \\ &= \mu + (b_0 + b_1 L + b_2 L^2 + ...)\epsilon_t \\ &= \mu + b(L)\epsilon_t \end{aligned}$$

where $b(z) = b_0 + b_1 z + b_2 z^2 + ...$ is an infinite-order polynomial. □

To recap so far, first we have shown that Y_t satisfies a projection onto its infinite past. Then, the Wold representation showed that in fact, this projection is a linear function of lagged projection errors. Now, we want to show an alternative way to write Y_t involving lagged values of Y_t. We now write an infinite order representation of Y_t using lagged values of Y_t.

Theorem 414 (Autoregressive Wold Representation)
If Y_t is covariance-stationary with Wold representation $Y_t = b(L)\epsilon_t$ satisfying some technical conditions, then Y_t has the representation:

$$Y_t = \mu + \sum_{j=1}^{\infty} a_j Y_{t-j} + \epsilon_t \tag{334}$$

for some coefficients μ and a_j whereby

$$\sum_{k=0}^{\infty} k^s |a_k| < \infty$$

and whereby ϵ_{t-j} are white noise projection errors.

Remark 415. *The technical assumptions says that for the Wold representation $Y_t = b(L)\epsilon_t$, the root of the polynomials $b(z)$ lie outside the unit circle $|z| > 1$. Furthermore, we require that the Wold coefficients satisfy:*

$$\sum_{j=0}^{\infty} (\sum_{k=0}^{\infty} k^s b_{j+k})^2 < \infty.$$

The summability of the Wold coefficients ensure convergence of the autoregressive coefficients a_j.

To summarise, we have shown two representations for a covariance stationary time series. The first is the Wold decomposition:

$$Y_t = \mu + \sum_{j=0}^{\infty} b_j \epsilon_{t-j}$$

whereby ϵ_{t-j} are white noise projection errors. The second representation is the autoregressive Wold representation:

$$Y_t = \mu + \sum_{j=1}^{\infty} a_j Y_{t-j} + \epsilon_t$$

whereby ϵ_t is also a white noise projection error. These two representations help us understand why we can approximate stationary time series with linear models, in particular autoregressive and moving-average models.

12.5.1 Linear Time Series Models

Definition 416 (AR(1) Process)
A process Y_t is an AR(1) process if:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \epsilon_t \tag{335}$$

where ϵ_t is a strictly stationary white noise process.

If we back substitute Y_{t-1}, we end up with:

$$\begin{aligned} Y_t &= \alpha_0 + \alpha_1(\alpha_0 + \alpha_1 Y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= \alpha_0(1+\alpha_1) + \alpha_1^2 Y_{t-2} + \alpha_1 \epsilon_{t-1} + \epsilon_t \end{aligned}$$

Continuously doing it for j-periods gives us:

$$Y_t = \alpha_0 \sum_{j=0}^{t-1} \alpha_1^j + \alpha_1^t Y_0 + \sum_{j=0}^{t-1} \alpha_1^j \epsilon_{t-j}.$$

Here, Y_t equals an intercept plus the scaled initial condition $\alpha_1^t Y_0$ and moving average expression. Suppose we can do this recursively into the infinite past and $|\alpha_1| < 1$. We then get the following result.

Proposition 417 (Convergent AR(1) Representation)
Suppose that Y_t is an AR(1) process and $|\alpha_1| < 1$. Then, its unique causal stationary solution has the convergent representation:

$$Y_t = \mu + \sum_{j=0}^{\infty} \alpha_1^j \epsilon_{t-j}. \tag{336}$$

where $\mu = \alpha_0/(1-\alpha_1)$. This solution is strictly stationary; a process started from an arbitrary fixed initial value converges to this stationary solution.

Proof. Look at the recursive form:

$$Y_t = \alpha_0 \sum_{j=0}^{t-1} \alpha_1^j + \alpha_1^t Y_0 + \sum_{j=0}^{t-1} \alpha_1^j \epsilon_{t-j}.$$

Use the result that if $|\beta| < 1$ then:

$$\sum_{k=0}^{\infty} \beta^k = \frac{1}{1-\beta}.$$

Then since $|\alpha_1| < 1$

$$\alpha_0 \lim_{t\to\infty} \sum_{j=0}^{t-1} \alpha_1^j = \frac{\alpha_0}{1-\alpha_1} \equiv \mu.$$

$$\lim_{t\to\infty} \alpha_1^t Y_0 = 0.$$

□

We now use the lag operator to derive the same result. First, we can rewrite the AR(1) process as:

$$(1-\alpha_1 L) Y_t = \alpha_0 + \epsilon_t.$$

Define the **autoregressive lag polynomial** $a(L) = 1 - \alpha_1 L$ to get:

$$a(L) Y_t = \alpha_0 + \epsilon_t.$$

As $a(L) = 1 - \alpha_1 L$, we want to know whether we can invert this. Recall that the geometric series is such that:

$$\sum_{j=0}^{\infty} x^j = \frac{1}{1-x} = (1-x)^{-1}$$

when $|x| < 1$. This gives us the following definition.

Definition 418 (Invertible)
A polynomial $\alpha(z)$ is invertible if

$$[\alpha(z)]^{-1} = \sum_{j=0}^{\infty} \alpha_j z^j \tag{337}$$

is absolutely convergent.

We now apply this to the AR(1) model. We currently have:

$$a(L)Y_t = \alpha_0 + \epsilon_t$$

where $a(L) = 1 - \alpha_1 L$. Now, if $a(L)$ is invertible, we have that:

$$[a(L)]^{-1} = [1 - \alpha_1 L]^{-1} = \frac{1}{1 - \alpha_1 L} = \sum_{j=0}^{\infty} \alpha_1^j L^j$$

Therefore, we apply this to the AR(1) model:

$$\begin{aligned}
&a(L)Y_t = \alpha_0 + \epsilon_t \\
&\leftrightarrow Y_t = a(L)^{-1}\alpha_0 + a(L)^{-1}\epsilon_t \\
&\leftrightarrow Y_t = \left(\sum_{j=0}^{\infty} \alpha_1^j L^j\right)\alpha_0 + \left(\sum_{j=0}^{\infty} \alpha_1^j L^j\right)\epsilon_t \\
&\leftrightarrow Y_t = \sum_{j=0}^{\infty} \alpha_1^j \alpha_0 + \left(\sum_{j=0}^{\infty} \alpha_1^j L^j\right)\epsilon_t \\
&\leftrightarrow Y_t = \sum_{j=0}^{\infty} \alpha_1^j \alpha_0 + \sum_{j=0}^{\infty} \alpha_1^j \epsilon_{t-j} \\
&\leftrightarrow Y_t = \frac{\alpha_0}{1 - \alpha_1} + \sum_{j=0}^{\infty} \alpha_1^j \epsilon_{t-j} \\
&\leftrightarrow Y_t = \mu + \sum_{j=0}^{\infty} \alpha_1^j \epsilon_{t-j}
\end{aligned}$$

We ended back at the convergent AR(1) representation!
We now move onto the class of MA(q) processes.

Definition 419 (MA(q) Process)
Y_t is a MA(q) process if it is of the form:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} \tag{338}$$

where ϵ_t is a strictly stationary white noise process.

Definition 420 (Infinite-Order Moving Average Process)

An infinite-order moving average process, denoted by $MA(\infty)$ is:

$$Y_t = \mu + \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}. \tag{339}$$

Remark 421. *This infinite-order moving average process is known as a* **linear process**.

Suppose we had the linear process:

$$Y_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}.$$

The coefficients of the moving average representation are known as the impulse response function.

Definition 422 (Impulse Response Function)

The **impulse response function** is defined as the change in Y_{t+j} due to a shock in time:

$$\frac{\partial}{\partial \epsilon_t} Y_{t+j} = \theta_j. \tag{340}$$

θ_j is the magnitude of the impact of a time t shock on the time $t+j$ variable.

We have defined the IRF for a linear process. We now want to know how to calculate the impulse response b_j from the coefficients of an autoregressive model of order p:

$$Y_t = \alpha_1 Y_{t-1} + ... + \alpha_p Y_{t-p} + \epsilon_t.$$

The impulse responses can be computed recursively from:

$$b_0 = 1, \qquad b_j = \sum_{\ell=1}^{\min(p,j)} \alpha_\ell b_{j-\ell} \quad \text{for } j \geq 1. \tag{341}$$

12.6 Autoregression

We go back to our autoregressive equation of order 1.

Definition 423 (AR(1))

An AR(1) model is given by:

$$X_t = \alpha X_{t-1} + \epsilon_t \tag{342}$$

whereby $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is i.i.d and independent of X_{t-j} for $j \geq 1$.

whereby X_0 is a **fixed value**. We can solve for a *solution* to this model.

Theorem 424 (Solution to AR(1))

The solution to an AR(1) model without deterministic terms is:

$$X_t = \sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0. \tag{343}$$

Proof. We can solve this equation through back-substitution:

$$\begin{aligned} X_t &= \alpha X_{t-1} + \epsilon_t \\ &= \alpha(\alpha X_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= \alpha^2 X_{t-2} + \alpha\epsilon_{t-1} + \epsilon_t \\ &= \alpha^2(\alpha X_{t-3} + \epsilon_{t-2}) + \alpha\epsilon_{t-1} + \epsilon_t \\ &= \alpha^3 X_{t-3} + \alpha^2\epsilon_{t-2} + \alpha\epsilon_{t-1} + \epsilon_t \end{aligned}$$

and via induction, we can arrive at:

$$X_t = \sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0.$$

□

Now, we recall a useful property of geometric series.

Lemma 425

The geometric series:

$$\sum_{j=0}^{t-1} \alpha^j = \begin{cases} \frac{1-\alpha^t}{1-\alpha} & \alpha \neq 1 \\ t & \alpha = 1 \end{cases} \tag{344}$$

$$\sum_{j=0}^{\infty} \alpha^j = \frac{1}{1-\alpha} \quad \text{for } |\alpha| < 1 \tag{345}$$

Theorem 426 (Conditional Mean and Variance of Conditional AR(1) Model)
Let X_t be an AR(1) model

$$X_t = \alpha X_{t-1} + \epsilon_t \tag{346}$$

with a fixed initial value X_0. Suppose that ϵ_t is i.i.d and normally distributed $\mathcal{N}(0, \sigma^2)$. Assume that ϵ_t and the fixed initial value X_0 are independent. Finally, assume that $|\alpha| < 1$.

Then, the conditional mean is given by:

$$\mu_t \equiv \mathbb{E}[X_t|X_0] = \alpha^t X_0 \tag{347}$$

The conditional variance is given by:

$$\sigma_t^2 \equiv Var[X_t|X_0] = \sigma^2 \frac{1-\alpha^{2t}}{1-\alpha^2}. \tag{348}$$

Proof. First, recall our recursive solution:

$$X_t = \sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0.$$

We then take the conditional expectation of this:

$$\begin{aligned}
\mathbb{E}[X_t|X_0] &= \mathbb{E}[\sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0|X_0] \\
&= \sum_{j=0}^{t-1} \alpha^j \mathbb{E}[\epsilon_{t-j}|X_0] + \alpha^t X_0 \\
&=_{(1)} \sum_{j=0}^{t-1} \alpha^j \mathbb{E}[\epsilon_{t-j}] + \alpha^t X_0 \\
&=_{(2)} \sum_{j=0}^{t-1} \alpha^j (0) + \alpha^t X_0 \\
&= \alpha^t X_0
\end{aligned}$$

whereby $=_{(1)}$ uses the independence of ϵ_t and X_0 and $=_{(2)}$ uses the mean of ϵ_t being 0.

For the conditional variance, we follow a similar approach by analysing the recursive solution.

$$\begin{aligned}
Var[X_t|X_0] &= Var[\sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0 | X_0] \\
&= Var[\sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} | X_0] \\
&= \sum_{j=0}^{t-1} \alpha^{2j} Var[\epsilon_{t-j}|X_0] \\
&= \sum_{j=0}^{t-1} \alpha^{2j} Var[\epsilon_{t-j}] \\
&= \sum_{j=0}^{t-1} \alpha^{2j} \sigma^2 \\
&= \sigma^2 \sum_{j=0}^{t-1} \alpha^{2j} \\
&= \sigma^2 \frac{1-\alpha^{2t}}{1-\alpha^2}.
\end{aligned}$$

□

We have derived the conditional mean and conditional variance for the AR(1) model. We now want to derive the limiting distribution now.

Theorem 427 (Asymptotic Normality of Conditional AR(1) Model)
Let X_t be an AR(1) model with a fixed initial value X_0. Then, the conditional limiting distribution of X_t is:

$$(X_t|X_0) \xrightarrow{D} \mathcal{N}(0, \frac{\sigma^2}{1-\alpha^2}) \tag{349}$$

Proof. By the recursive solution, we know:

$$X_t = \sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0.$$

We also know that ϵ_t are i.i.d normal random variables. Therefore, the sum of i.i.d normal random variable is also a normal random variable. Furthermore, we have already derived the conditional mean and conditional variance for an AR(1) model, so we can conclude that:

$$X_t|X_0 \sim \mathcal{N}(\alpha^t X_0, \sigma^2 \frac{1-\alpha^{2t}}{1-\alpha^2}).$$

Now, we know that $|\alpha| < 1$ so we can take the limit to conclude that:

$$X_t|X_0 \to \mathcal{N}(0, \sigma^2 \frac{1}{1-\alpha^2}).$$

□

Pausing to reflect what we have just done, we have shown that the AR(1) conditional model is asymptotically

normal:

$$(X_t|X_0) \xrightarrow{D} \mathcal{N}(0, \frac{\sigma^2}{1-\alpha^2})$$

The variance $\frac{\sigma^2}{1-\alpha^2}$ is the **stationary variance**. This is in contrast to the innovation variance σ^2. The long-run variance relevant for the sample mean will instead be $\frac{\sigma^2}{(1-\alpha)^2}$.

We now want to examine the autocorrelation of the AR(1) model. Before we do that, we first need the autocovariance of the AR(1) model.

Proposition 428 (Conditional Autocovariance of AR(1) Model)
Let X_t be a conditional AR(1) process:

$$X_t = \alpha X_{t-1} + \epsilon_t$$

for a fixed X_0, $|\alpha| < 1$ and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d and independent of all X_{t-j} for $j \geq 1$.

Then, for a fixed $s \geq 1$, the conditional autocovariance of the AR(1) model is:

$$Cov[X_t, X_{t-s}|X_0] = \alpha^s Var[X_{t-s}|X_0]. \tag{350}$$

Proof. We plug in the recursive formula for the AR(1) model into the conditional autocovariance function:

$$\begin{aligned} Cov[X_t, X_{t-s}|X_0] &= Cov[\sum_{j=0}^{s-1} \alpha^j \epsilon_{t-j} + \alpha^s X_{t-s}, X_{t-s}|X_0] \\ &=_{(1)} Cov[\alpha^s X_{t-s}, X_{t-s}|X_0] \\ &= \alpha^s Cov[X_{t-s}, X_{t-s}|X_0] \\ &= \alpha^s Var[X_{t-s}|X_0] \end{aligned}$$

where in $=_{(1)}$, we use the fact that ϵ_t was from time subscript $t-s+1$ onward and hence is independent of X_{t-s}, implying a covariance of zero. □

We now look at the autocorrelation of the AR(1) model whereby we perform asymptotics on the limiting distribution of the AR(1) model.

Proposition 429 (Autocorrelation for Limiting Distribution of AR(1) Model)
Let X_t be a conditional AR(1) process:

$$X_t = \alpha X_{t-1} + \epsilon_t$$

for a fixed X_0, $|\alpha| < 1$ and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d and independent of all X_{t-j} for $j \geq 1$.

Then, for a fixed $s \geq 1$, the conditional autocorrelation for the limiting distribution of the AR(1) model is given by:

$$Corr(X_t, X_{t-s}|X_0) \to \alpha^s \tag{351}$$

as $t \to \infty$.

Proof. First, recall that the conditional autocovariance of the AR(1) model is:

$$Cov(X_t, X_{t-s}|X_0) = \alpha^s Var[X_{t-s}|X_0]$$

and the conditional variance of the AR(1) model is:

$$Var[X_{t-s}|X_0] = \sigma^2 \frac{1-\alpha^{2(t-s)}}{1-\alpha^2}.$$

Therefore, the limit of the conditional autocovariance is:

$$\lim_{t\to\infty} \alpha^s Var[X_{t-s}|X_0] = \lim_{t\to\infty} \alpha^s \sigma^2 \frac{1-\alpha^{2(t-s)}}{1-\alpha^2} = \alpha^s \frac{\sigma^2}{1-\alpha^2}.$$

Finally, recall that we also have shown that the long-run variance was:

$$\lim_{t\to\infty} Var[X_t|X_0] = \frac{\sigma^2}{1-\alpha^2}.$$

Now, we compute the formula for the conditional autocorrelation:

$$\begin{aligned} Corr(X_t, X_{t-s}|X_0) &= \frac{Cov(X_t, X_{t-s}|X_0)}{\sqrt{Var[X_t|X_0]Var[X_{t-s}|X_0]}} \\ &= \frac{\alpha^s Var[X_{t-s}|X_0]}{\sqrt{Var[X_t|X_0]Var[X_{t-s}|X_0]}} \\ &= \frac{\alpha^s Var[X_{t-s}|X_0]}{\sqrt{Var[X_t|X_0]Var[X_{t-s}|X_0]}} \end{aligned}$$

We then take limits of the numerator and denominator:

$$\frac{\lim_{t\to\infty} \alpha^s Var[X_{t-s}|X_0]}{\lim_{t\to\infty} \sqrt{Var[X_t|X_0]Var[X_{t-s}|X_0]}} = \frac{\alpha^s \frac{\sigma^2}{1-\alpha^2}}{\frac{\sigma^2}{1-\alpha^2}} = \alpha^s.$$

□

Hence, the autocorrelation for the limiting distribution is α^s.

Remark 430. *When $0 < \alpha < 1$, we have an exponential decay in the autocorrelation of the AR(1) model.*

12.7 Autoregressive Model With Intercepts and Time Trends

We now move onto analysing an AR(1) model with an intercept term.

Definition 431 (AR(1) Model with Intercept)
The AR(1) model with an intercept term is:

$$Y_t = \alpha Y_{t-1} + \epsilon_t + \mu \tag{352}$$

where μ is an intercept term and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d and independent of all Y_{t-j} for $j \geq 1$ and with a fixed initial value Y_0.

The issue is that we cannot use our previous results as we now have a μ term that appears and by recursion, this μ term will also appear in X_{t-1} and so on. We now stress the following trick we will use to solve this issue because this trick is used all the time in time series. **The trick is to rewrite the AR(1) model with an intercept into a AR(1) model without an intercept plus the intercept term.** After we have rewritten the model, we can then apply the results we have already derived.

Proposition 432 (AR(1) Model without an Intercept)

Let Y_t be an AR(1) model with an intercept term. Then, we can rewrite Y_t as an AR(1) model without an intercept term by decomposing it into the sum of a mean-zero process X_t and long-run mean μ_Y:

$$X_t = \alpha X_{t-1} + \epsilon_t \tag{353}$$

$$Y_t = X_t + \mu_Y \tag{354}$$

$$\mu_Y = \frac{\mu}{1-\alpha}. \tag{355}$$

Proof. First, let us take our AR(1) model and subtract off a constant μ_Y (this is an arbitrary constant for now but we can pin down what this constant will need to be later):

$$Y_t - \mu_Y = \alpha Y_{t-1} + \epsilon_t + \mu - \mu_Y$$

Now, we have a $Y_t - \mu_Y$ on the left hand side, and to get something similar on the right hand side (but dated one index back), we then add and subtract $\alpha\mu_Y$ on the right hand side:

$$Y_t - \mu_Y = \alpha Y_{t-1} + \epsilon_t + \mu - \mu_Y + \alpha\mu_Y - \alpha\mu_Y$$

We now group the terms as follows:

$$Y_t - \mu_Y = \alpha Y_{t-1} + \epsilon_t + \mu - \mu_Y - \alpha\mu_Y + \alpha\mu_Y$$

$$Y_t - \mu_Y = \alpha(Y_{t-1} - \mu_Y) + \epsilon_t + [\mu - (1-\alpha)\mu_Y]$$

We can see that the terms in red actually form an AR(1) model without an intercept! To make this more explicit, we can define the variable:

$$X_t \equiv Y_t - \mu_Y$$

which then allows us to then rewrite the expression

$$Y_t - \mu_Y = \alpha(Y_{t-1} - \mu_Y) + \epsilon_t + [\mu - (1-\alpha)\mu_Y]$$

to now be:

$$X_t = \alpha X_{t-1} + \epsilon_t + [\mu - (1-\alpha)\mu_Y]$$

where the terms in blue are a constant that we now need to get rid off. That is, if we make the terms in blue $= 0$, then we end up with an AR(1) model without an intercept:

$$X_t = \alpha X_{t-1} + \epsilon_t$$

which is something we have already previously derived results for! Hence, going back to the blue term, we are required to pick a constant μ_Y such that:

$$\mu - (1-\alpha)\mu_Y = 0$$

and clearly, this algebraically holds when:

$$\mu_Y = \frac{\mu}{1-\alpha}.$$

Therefore, we set μ_Y to be this value. Hence, we can see that we have decomposed the AR(1) with an intercept

Y_t (just add and subtract μ_Y to Y_t) to be:

$$Y_t = \underbrace{(Y_t - \mu_Y)}_{=X_t} + \underbrace{\mu_Y}_{=\frac{\mu}{1-\alpha}} = X_t + \mu_Y$$

where X_t is an AR(1) without an intercept $X_t = \alpha X_{t-1} + \epsilon_t$ and μ_Y is just a constant. □

We can informally refer to this procedure as an AR(1) decomposition.

We now have the tools to derive the limiting distribution of an AR(1) model with an intercept term.

Proposition 433 (Limiting Distribution of AR(1) With Intercept)
The limiting distribution of an AR(1) model with an intercept Y_t is:

$$Y_t|Y_0 \xrightarrow{D} \mathcal{N}\left(\mu_Y, \frac{\sigma^2}{1-\alpha^2}\right) \tag{356}$$

Proof. Rewrite the AR(1) model with an intercept term:

$$Y_t = \alpha Y_{t-1} + \epsilon_t + \mu$$

into an AR(1) model without an intercept term plus a constant:

$$Y_t = X_t + \mu_Y$$

whereby $X_t = \alpha X_{t-1} + \epsilon_t$ and $\mu_Y = \frac{\mu}{1-\alpha}$. We have previously shown that the conditional distribution for an AR(1) without an intercept is:

$$X_t|X_0 \xrightarrow{D} \mathcal{N}(0, \frac{\sigma^2}{1-\alpha^2})$$

We can use this to derive the limiting distribution of Y_t as adding a constant μ_Y shifts the mean but does not change the variance and hence we can conclude:

$$Y_t|Y_0 \xrightarrow{D} \mathcal{N}\left(\mu_Y, \frac{\sigma^2}{1-\alpha^2}\right)$$

□

We now introduce the AR(1) model with an intercept **and** a linear time trend.

Definition 434 (AR(1) Model with Intercept and Linear Time Trend)
The AR(1) model with an intercept and linear time trend is given by:

$$Y_t = \alpha Y_{t-1} + \mu_c + \mu_l t + \epsilon_t \tag{357}$$

where μ_c is an intercept term, t is the linear time trend and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d and independent of all Y_{t-j} for $j \geq 1$ and with a fixed initial value Y_0.

We now want to rewrite this model to remove the intercept and time trend.

Proposition 435 (AR(1) Model without an Intercept and Time Trend)

Let Y_t be an AR(1) model with an intercept term and a linear time trend term. Then, we can rewrite Y_t as an AR(1) model without an intercept term and linear time trend term by decomposing it into the sum of a mean-zero process X_t, long-run intercept mean τ_c and long-run time trend mean τ_l:

$$X_t = \alpha X_{t-1} + \epsilon_t \tag{358}$$

$$Y_t = X_t + \tau_c + \tau_l t \tag{359}$$

$$\tau_c = \frac{\mu_c - \alpha\tau_l}{1-\alpha} \tag{360}$$

$$\tau_l = \frac{\mu_l}{1-\alpha} \tag{361}$$

Proof. (Sketch). The proof is similar to removing the intercept from an AR(1) model. First, subtract τ_c and $\tau_l t$ from both sides of:

$$Y_t = \alpha Y_{t-1} + \mu_c + \mu_l t + \epsilon_t$$

to get:

$$Y_t - \tau_c - \tau_l t = \alpha Y_{t-1} + \mu_c + \mu_l t + \epsilon_t - \tau_c - \tau_l t$$

Now, add and subtract $\alpha\tau_c + \alpha\tau_l + \alpha\tau_l t$ to the right-hand side to get:

$$Y_t - \tau_c - \tau_l t = \alpha Y_{t-1} + \mu_c + \mu_l t + \epsilon_t - \tau_c - \tau_l t + \alpha\tau_c + \alpha\tau_l + \alpha\tau_l t - \alpha\tau_c - \alpha\tau_l - \alpha\tau_l t.$$

Now, start to group terms together to get:

$$Y_t - \tau_c - \tau_l t = \alpha(Y_{t-1} - \tau_c + \tau_l - \tau_l t) + \mu_c + \mu_l t - \tau_c - \tau_l t + \alpha\tau_l + \alpha\tau_l t - \alpha\tau_l + \epsilon_t.$$

and we end up with:

$$Y_t - \tau_c - \tau_l t = \alpha\Big[Y_{t-1} - \tau_c - \tau_l(t-1)\Big] + \Big[\mu_c - \tau_c + \alpha\tau_c - \alpha\tau_l\Big] + \Big[\mu_l - \tau_l + \alpha\tau_l\Big]t + \epsilon_t.$$

We then define the mean-zero process:

$$X_t \equiv Y_t - \tau_c - \tau_l t.$$

Therefore, we have that:

$$\underbrace{Y_t - \tau_c - \tau_l t}_{=X_t} = \alpha \underbrace{\Big[Y_{t-1} - \tau_c - \tau_l(t-1)\Big]}_{X_{t-1}} + \underbrace{\Big[\mu_c - \tau_c + \alpha\tau_c - \alpha\tau_l\Big]}_{=0} + \Big[\underbrace{\mu_l - \tau_l + \alpha\tau_l}_{=0}\Big]t + \epsilon_t.$$

to get an AR(1) process without an intercept or time-trend:

$$X_t = \alpha X_{t-1} + \epsilon_t$$

whereby we set the two terms equal to 0 to arrive at:

$$\mu_c - \tau_c + \alpha\tau_c - \alpha\tau_l = 0 \quad \leftrightarrow \quad \tau_c = \frac{\mu_c - \alpha\tau_l}{1-\alpha}$$

and

$$\mu_l - \tau_l + \alpha\tau_l = 0 \quad \leftrightarrow \quad \tau_l = \frac{\mu_l}{1-\alpha}.$$

□

We then can derive the limiting distribution of an AR(1) model with an intercept and time-trend.

Proposition 436 (Limiting Distribution of AR(1) With Intercept and Time-Trend)
The detrended AR(1) model has the conditional limiting distribution:

$$Y_t - (\tau_c + \tau_l t)|Y_0 \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2}{1-\alpha^2}\right) \tag{362}$$

whereby

$$\tau_c = \frac{\mu_c - \alpha\tau_l}{1-\alpha}$$
$$\tau_l = \frac{\mu_l}{1-\alpha}.$$

Proof. First, decompose

$$Y_t = \alpha Y_{t-1} + \mu_c + \mu_l t + \epsilon_t$$

into an AR(1) without an intercept and time-trend:

$$X_t = \alpha X_{t-1} + \epsilon_t \tag{363}$$
$$Y_t = X_t + \tau_c + \tau_l t \tag{364}$$
$$\tau_c = \frac{\mu_c - \alpha\tau_l}{1-\alpha} \tag{365}$$
$$\tau_l = \frac{\mu_l}{1-\alpha} \tag{366}$$

Then, apply our result that the conditional limiting distribution of an AR(1) model is:

$$X_t|X_0 \xrightarrow{D} \mathcal{N}(0, \frac{\sigma^2}{1-\alpha^2})$$

when $|\alpha| < 1$. Now if we look at:

$$Y_t = X_t + \tau_c + \tau_l t$$

we can then immediately conclude that the detrended process satisfies:

$$Y_t - (\tau_c + \tau_l t)|Y_0 \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2}{1-\alpha^2}\right).$$

□

Remark 437. *In this scenario, we have that Y_t is* **trend stationary**. *That is, if we subtract a linear time trend from Y_t, we will get a stationary process.*

Hence, the big takeaway in this section is how to analyse time series models with intercepts and time-trends by rewriting them in terms of models we already know properties about and adding a constant.

12.7.1 Random Walks

Definition 438 (Covariance Stationary)

A scalar process $\{y_t\}_{t\geq 1}$ is **covariance-stationary** if:

1. $\mathbb{E}[y_t] = \mu$
2. $Cov(y_t, y_{t+s}) = \gamma_s$ for all s.

Remark 439. *The important thing is that the first moment and covariance does not depend on time t.*

Definition 440 (White Noise)

The process u_t is a **white noise** if:

$$\mathbb{E}[u_t] = 0 \tag{367}$$

$$\mathbb{E}[u_{t-j}u_t] = 0 \quad j \geq 1 \tag{368}$$

$$\mathbb{E}[u_t^2] = \sigma_u^2 < \infty. \tag{369}$$

We now view what it means for a variable to be driven by a **stochastic trend**.

Definition 441 (AR(1) Process)

The AR(1) process is:

$$y_t = \alpha y_{t-1} + u_t \tag{370}$$

where u_t is a **white-noise process.**

When $\alpha = 1$, we get a random walk process.

Definition 442 (Random Walk)

A random walk process is:

$$y_t = y_{t-1} + u_t. \tag{371}$$

We can recursively substitute to get:

$$y_t = y_0 + \sum_{j=1}^{t} u_j.$$

Then:

$$\mathbb{E}[y_t] = \mathbb{E}[y_0]$$

$$Var[y_t] = t\sigma_u^2 + Var[y_0]$$

The variance of a random walk tends to infinity.

Proposition 443 (Persistence of Random Walk Process)

For any given integer s,

$$Corr(y_t, y_{t+s}) \to 1 \tag{372}$$

as $t \to \infty$.

Proof. For simplicity, take y_0 to be fixed; this does not affect the limiting correlation. Then:

$$Corr(y_t, y_{t+s}) = \frac{\mathbb{E}\left[\left(\sum_{j=1}^{t} u_j\right)\left(\sum_{j=1}^{t+s} u_j\right)\right]}{\left[t\sigma_u^2(t+s)\sigma_u^2\right]^{1/2}} = \frac{t}{(t^2+ts)^{1/2}} \to 1$$

as $t \to \infty$. □

Random variables y_t and y_{t+s} are still strongly correlated even when they are far apart in time. This indicates a strong persistence in the time process.

Definition 444 (Random Walk with Drift)
A random walk with a drift is:

$$y_t = \nu + y_{t-1} + u_t \tag{373}$$

where ν is a constant term and u_t is a white-noise process.

12.8 Stationarity

We now move onto the topic of stationarity. Essentially, stationarity imposes structure on our time series and makes things much easier for analysis.

Definition 445 (Strict Stationarity)
A process $\{X_t\}_t$ is strictly stationary if for any s, the joint distribution of $(X_t, X_{t+1}, ..., X_{t+s})$ does not depend on time t.

Remark 446. *Sometimes, an $*$ is used X_t^* to indicate that a process is a strictly stationary process.*

Hence, if we have a strictly stationary process, then the joint distribution of $(X_{t+1}, ..., X_{t+s})$ will be the same distribution of $(X_1, ..., X_s)$. That is, if we have a block of variables and a joint density associated to this block, the distribution of the block will be the same no matter where we move it around.

Proposition 447 (Strict Stationarity is Preserved Under Measurability)
If X_t^* is a strictly stationary process and $g : \mathbb{R}^n \to \mathbb{R}$ is a measurable function, then

$$Z_t = g(X_t^*, X_{t-1}^*, ..., X_{t-n}^*) \tag{374}$$

is a strictly stationary process.

We give an important class of strictly stationary processes.

Proposition 448 (i.i.d Sequences are Strictly Stationary)
If Y_t is i.i.d, then it is **strictly stationary**.

Example 449
Suppose we had i.i.d ϵ_t sequence. Then, ϵ_t is strictly stationary. Now, if we take a transformation $g(x) = x^2$ so that $g(\epsilon_t) = \epsilon_t^2$, we have that ϵ_t^2 is also a strictly stationary process.

We now define another notion of stationarity.

Definition 450 (Covariance Stationary)
A process $\{X_t\}_t$ is covariance stationary if for all s, t

1. $\mathbb{E}[X_t] = \mu$
2. $Cov(X_t, X_{t+s}) = \psi_s$.

Remark 451. *The first condition says that the expectation does not depend on the time t. Additionally, the covariance does not depend on time t as well. As a corollary, the unconditional variance does not depend on t, but the conditional variance can.*

Something important to note is that covariance stationarity is not preserved under transformations. An intuition for this is because if X_t has second moments, this does not imply that X_t^2 has second moments.

We now want to relative strict stationarity and covariance stationarity to each other. In fact, neither strict stationarity nor covariance stationarity is stronger than the other. However, if we impose some assumptions, we can relate the two concepts of stationarity.

Definition 452 (Gaussian Process)
A process X_t is a Gaussian process if for all s and time t, the joint distribution $(X_{t+1}, ..., X_{t+s})$ is a Gaussian distribution.

We can now relate the two concepts of stationarity.

Proposition 453 (Sufficient Conditions for Strict Stationarity)
A Gaussian covariance stationary process is strictly stationary.

Proof. A finite-dimensional Gaussian distribution is characterised by its mean vector and covariance matrix. Hence, if these are invariant to a common shift in time, then the joint distribution is also invariant to that shift. Therefore, covariance stationarity implies strict stationarity for a Gaussian process. □

Proposition 454 (Sufficient Conditions for Covariance Stationarity)
A strictly stationary process with second moments is a covariance stationary process.

Proof. We need to show that if X_t is strictly stationary, then its first moment and autocovariance is time-invariant.

First, we look at the first moment of X_t:

$$\mathbb{E}[X_t] = \int x f_t(x) dx =_{(1)} \int x f_1(x) dx = \mathbb{E}[X_1]$$

whereby $=_{(1)}$ uses strict stationarity of X_t. Hence, we have shown that:

$$\mathbb{E}[X_t] = \mathbb{E}[X_1]$$

for all time t.

Now look at the variance:

$$\begin{aligned} Var[X_t] &= \mathbb{E}[X_t^2] - \mathbb{E}[X_t]^2 \\ &= \int x^2 f_t(x) dx - \mathbb{E}[X_t]^2 \\ &=_{(1)} \int x^2 f_1(x) dx - \mathbb{E}[X_t]^2 \\ &= \mathbb{E}[X_1^2] - \mathbb{E}[X_t]^2 \\ &= Var[X_1]. \end{aligned}$$

Hence, we have that $Var[X_t] = Var[X_1]$ for all time t.

Now, we look at the autocovariance:

$$\begin{aligned} Cov[X_{t+h}, X_t] &= \mathbb{E}[X_{t+h}X_t] - \mathbb{E}[X_{t+h}]\mathbb{E}[X_t] \\ &= \int_{\mathbb{R}^2} xy f_{t+h,t}(x,y) dx dy - \mathbb{E}[X_{t+h}]\mathbb{E}[X_t] \\ &=_{(1)} \int_{\mathbb{R}^2} xy f_{1+h,1}(x,y) dx dy - \mathbb{E}[X_{t+h}]\mathbb{E}[X_t] \\ &= \mathbb{E}[X_{1+h}X_1] - \mathbb{E}[X_{t+h}]\mathbb{E}[X_t] \\ &=_{(2)} \mathbb{E}[X_{1+h}X_1] - \mathbb{E}[X_{1+h}]\mathbb{E}[X_1] \\ &= Cov[X_{1+h}, X_1] \end{aligned}$$

whereby $=_{(1)}$ and $=_{(2)}$ uses strict stationarity. Hence we have shown that $Cov[X_{t+h}, X_t] = Cov[X_{1+h}, X_1]$ for all time t.

Therefore, we have shown that a strictly stationary process with second moments satisfy the conditions for covariance stationarity. □

Example 455

(Cauchy Distribution). A series $\{X_t\}_{t\geq 1}$ can be strictly stationary if each X_t is Cauchy distributed. However, the Cauchy distribution does not have any moments. Therefore, such a sequence cannot be covariance stationary. That is why we require the assumption of the existence of second moments for a strictly stationary series to be covariance stationary.

Example 456

(MA(1) Process) Consider the MA(1) process:

$$X_t = \theta\epsilon_{t-1} + \epsilon_t$$

where $\{\epsilon_t\}_{t\geq 1}$ is i.i.d with $\mathbb{E}[\epsilon_t] = 0$ and $Var[\epsilon_t] = \sigma^2$.

Then, we can argue for covariance stationarity in two different ways.

The first way is to note that as ϵ_t is i.i.d, then we know that ϵ_t is strictly stationarity. Furthermore, summing two strictly stationary series preserves strict stationarity so then:

$$X_t = \theta\epsilon_{t-1} + \epsilon_t$$

is strictly stationary. Finally, we know that X_t has finite second moments as ϵ_t as finite second moments and hence we can conclude that X_t is then covariance stationary.

The second way is verify the MA(1) process is covariance stationary directly. First, compute the first moment:

$$\mathbb{E}[X_t] = \mathbb{E}[\theta\epsilon_{t-1} + \epsilon_t] = 0.$$

The variance is:

$$Var[X_t] = \mathbb{E}[X_t^2] = (1 + \theta^2)\sigma^2.$$

Therefore, the covariance is:

$$Cov(X_t, X_{t-1}) = \theta\sigma^2$$
$$Cov(X_t, X_{t-s}) = 0 \quad \text{for } s \geq 2.$$

Hence we have shown directly that X_t is covariance stationary.

A lot of times, we have a partial sum of random variables:

$$\sum_{j=0}^{t} \alpha^j X_j$$

and we want to know whether can we simply take limits to get the series:

$$\sum_{j=0}^{\infty} \alpha^j X_j.$$

The following theorem tells us when we can do so.

Theorem 457 (Kolmogorov's 2 Series Theorem)

Let $\{Z_i\}_{i\geq 1}$ be independent random variables. Then suppose that:

1. the series $\sum_{i=0}^{\infty}\mathbb{E}[Z_i]$ converges
2. $\sum_{i=0}^{\infty}Var[Z_i]<\infty$.

Then, the series

$$\sum_{i=0}^{\infty}Z_i \tag{375}$$

converges almost surely.

Remark 458. *We will give an example of applying this theorem in the next section.*

12.9 Unconditional Stationary Autoregression

Previously, we have looked at the **conditional model** of AR(1) whereby:

$$X_t=\alpha X_{t-1}+\epsilon_t$$

where $|\alpha|<1$ and we had a fixed initial value X_0. This then gave us the limiting distribution:

$$X_t|X_0\xrightarrow{D}\mathcal{N}(0,\frac{\sigma^2}{1-\alpha^2}).$$

We now discuss a second formulation of the AR(1) model known as the **unconditional** AR(1) model. The idea is that instead of fixing an initial value X_0, we let X_0 be a random variable whose distribution is the stationary limiting distribution X_0^*.

Definition 459 (Unconditional AR(1) Model)

The **unconditional** AR(1) model is given by:

$$X_t^*=\alpha X_{t-1}^*+\epsilon_t \tag{376}$$

$$X_0^*\sim\mathcal{N}(0,\frac{\sigma^2}{1-\alpha^2}) \tag{377}$$

whereby $\{X_t^*\}_{t\geq 1}$ is a stationary sequence, ϵ_t are i.i.d $\mathcal{N}(0,\sigma^2)$ and independent of X_0^*.

We can then derive the unconditional joint distribution and moments of such a process.

Proposition 460 (Joint Distribution of Unconditional AR(1) Model)

Let $\{X_t^*\}_{t\geq 1}$ be the unconditional AR(1) model. Then, for $s\geq 1$, the joint distribution is:

$$\begin{bmatrix}X_t^*\\X_{t+s}^*\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\frac{\sigma^2}{1-\alpha^2}\begin{bmatrix}1&\alpha^s\\\alpha^s&1\end{bmatrix}\right) \tag{378}$$

Proof. First, we use backward substitution to get:

$$X_t^*=\sum_{j=0}^{t-1}\alpha^j\epsilon_{t-j}+\alpha^tX_0^*.$$

We now compute the **unconditional** moments of X_t^*. The first moment is:

$$\mathbb{E}[X_t^*] = \mathbb{E}[\sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0^*] = 0.$$

The variance is:

$$\begin{aligned} Var[X_t^*] &= Var[\sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0^*] \\ &=_{(1)} Var[\sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j}] + Var[\alpha^t X_0^*] \\ &= \sum_{j=0}^{t-1} \alpha^{2j} \sigma^2 + \alpha^{2t} Var[X_0^*] \\ &= \frac{\sigma^2(1-\alpha^{2t})}{1-\alpha^2} + \alpha^{2t} \frac{\sigma^2}{1-\alpha^2} \\ &= \frac{\sigma^2}{1-\alpha^2}. \end{aligned}$$

where $=_{(1)}$ uses the independence of ϵ_{t-j} and X_0.

The autocovariance is:

$$Cov[X_t^*, X_{t+s}^*] = \mathbb{E}[X_t^* X_{t+s}^*] - \mathbb{E}[X_t^*]\mathbb{E}[X_{t+s}^*] = \mathbb{E}[X_t^* X_{t+s}^*] - 0$$

We now substitute in our recursive formula:

$$\begin{aligned} Cov[X_t^*, X_{t+s}^*] = \mathbb{E}[X_t^* X_{t+s}^*] &= \mathbb{E}[X_t^* (\sum_{j=0}^{s-1} \alpha^j \epsilon_{t+s-j} + \alpha^s X_t^*)] \\ &= \mathbb{E}[X_t^* \sum_{j=0}^{s-1} \alpha^j \epsilon_{t+s-j}] + \mathbb{E}[\alpha^s (X_t^*)^2] \\ &= \sum_{j=0}^{s-1} \alpha^j \mathbb{E}[X_t^* \epsilon_{t+s-j}] + \alpha^s \mathbb{E}[(X_t^*)^2] \\ &=_{(1)} 0 + \alpha^s \mathbb{E}[(X_t^*)^2] \\ &= \alpha^s Var[X_t^*] \\ &= \alpha^s \frac{\sigma^2}{1-\alpha^2} \end{aligned}$$

where $=_{(1)}$ uses the independence of ϵ_{t+s-j} and X_t^* for $j = 0, \ldots, s-1$. We see that the covariance does not depend on t.

Now, we look at the joint distribution of X_t^* and X_{t+s}^*:

$$X_t^* = \sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0^*$$
$$X_{t+s}^* = \sum_{j=0}^{t+s-1} \alpha^j \epsilon_{t+s-j} + \alpha^{t+s} X_0^*$$

We know that ϵ_t and X_0^* are normally distributed. Furthermore, we know that the sum of normal random variables is a normal random variable. Therefore, they are jointly normal with the joint distribution:

$$\begin{bmatrix} X_t^* \\ X_{t+s}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{\sigma^2}{1-\alpha^2} \begin{bmatrix} 1 & \alpha^s \\ \alpha^s & 1 \end{bmatrix} \right).$$

□

Remark 461. *Something important to notice is that this is an unconditional distribution. Previously, in the conditional AR(1) model, we derived a limiting conditional distribution. This is again due to the set up of the unconditional AR(1) model depending initially on a stationary X_0^*.*

Now, we know that the stationary solution to our unconditional AR(1) model $X_t = \alpha X_{t-1} + \epsilon_t$ is:

$$X_t = \sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0.$$

If $|\alpha| < 1$, we want to know whether can we take $t \to \infty$ and write:

$$X_t^* = \sum_{j=0}^{\infty} \alpha^j \epsilon_{t-j}.$$

The following results tells us yes.

Theorem 462 (Moving Average Representation of Unconditional AR(1) Model)
Let X_t be the unconditional AR(1) model. Then, if $|\alpha| < 1$, X_t has the moving average representation:

$$X_t^* = \sum_{j=0}^{\infty} \alpha^j \epsilon_{t-j}. \tag{379}$$

Proof. This follows from **Kolmogorov's 2 series theorem**. We verify that the conditions for the theorem hold. First, the innovations ϵ_{t-i} are independent by assumption. Now, we check the first condition:

$$\sum_{i=0}^{\infty} \mathbb{E}[\alpha^i \epsilon_{t-i}] = \sum_{i=0}^{\infty} \alpha^i \mathbb{E}[\epsilon_{t-i}] = 0$$

as $\mathbb{E}[\epsilon_i] = 0$.

We then check the variance:

$$\sum_{i=0}^{\infty} Var[\alpha^i \epsilon_{t-i}] = \sum_{i=0}^{\infty} \alpha^{2i} \sigma^2 = \frac{\sigma^2}{1-\alpha^2} < \infty. \tag{380}$$

Both conditions for Kolmogorov's 2 series theorem hold and hence:

$$\sum_{i=0}^{\infty} \alpha^i \epsilon_{t-i}$$

converges almost surely. □

To conclude this section, we look at the lag operator and show a different way to derive the moving average representation using lag operators.

Definition 463 (Lag Operator)
The lag operator is the operator L such that:

$$LX_t = X_{t-1}. \tag{381}$$

Now, if we have the AR(1) model:

$$X_t = \alpha X_{t-1} + \epsilon_t$$

we can express this in the **moving average representation**.

Theorem 464 (AR(1) Moving Average Representation)
Let X_t be an AR(1) model

$$X_t = \alpha X_{t-1} + \epsilon_t.$$

Then, if $|\alpha| < 1$, we can write X_t in its infinite order moving average representation:

$$X_t = \sum_{j=0}^{\infty} \alpha^j \epsilon_{t-j}. \tag{382}$$

Proof. First take the AR(1) model:

$$X_t = \alpha X_{t-1} + \epsilon_t.$$

We can rewrite this using the lag operator:

$$\begin{aligned} X_t - \alpha X_{t-1} &= \epsilon_t \\ \leftrightarrow (1 - \alpha L) X_t &= \epsilon_t \end{aligned}$$

Now, let us define the polynomial:

$$A(z) = 1 - \alpha z \tag{383}$$

where if $|\alpha z| < 1$, then $A(z)$ has an inverse:

$$\left[A(z)\right]^{-1} = \sum_{j=0}^{\infty} \alpha^j z^j. \tag{384}$$

Then, we have that if we applied this to the AR(1) model:

$$
\begin{aligned}
&(1-\alpha L)X_t = \epsilon_t \\
&\leftrightarrow X_t = [(1-\alpha L)]^{-1}\epsilon_t \\
&\leftrightarrow X_t = \sum_{j=0}^{\infty} \alpha^j \epsilon_{t-j}.
\end{aligned}
$$

□

12.10 Asymptotic Theory for Time Series

First, we recall the definition of a $\sigma - algebra$.

Definition 465 ($\sigma - algebra$)

A $\sigma - algebra$ is a collection of sets $\mathcal{F}$ such that:

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$
3. If $A_1, A_2, ... \in \mathcal{F}$ then
$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

Definition 466 (Filtration)

A filtration is a sequence of $\sigma - algebra$'s $\mathcal{F}_t$ such that:

$$\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \tag{385}$$

for all time.

A filtration captures the idea that the amount of information you have is *growing* over time.

Definition 467 (Martingale)

A process $\{X_t\}_{t\geq 1}$ is a martingale if

$$\mathbb{E}[X_{t+1}|\mathcal{F}_t] = X_t. \tag{386}$$

Remark 468. *The best guess for the stock price tomorrow is the stock price today, on average.*

Now, suppose we construct a new process consisting of taking the difference of martingales:

$$m_t = X_t - X_{t-1}$$

whereby $\{X_t\}_{t\geq 1}$ is a martingale. Then, the conditional expectation of m_t is:

$$\begin{aligned}
\mathbb{E}[m_t|\mathcal{F}_{t-1}] &= \mathbb{E}[X_t - X_{t-1}|\mathcal{F}_{t-1}] \\
&= \mathbb{E}[X_t|\mathcal{F}_{t-1}] - \mathbb{E}[X_{t-1}|\mathcal{F}_{t-1}] \\
&= \mathbb{E}[X_t|\mathcal{F}_{t-1}] - X_{t-1} \\
&= X_{t-1} - X_{t-1} = 0.
\end{aligned}$$

We now define an important concept that we will use frequently.

Definition 469 (Martingale Difference Sequence)

A pair $(m_t, \mathcal{F}_t)$ for $t \in \mathbb{N}$ is a **martingale difference sequence** if:

1. m_t is $\mathcal{F}_t$-measurable
2. $\mathbb{E}[|m_t|] < \infty$
3. $\mathbb{E}[m_t|\mathcal{F}_{t-1}] = 0$.

As we will soon see, martingale differences are a relaxation of the assumption of independence. To see this, we derive properties of the MDS.

Proposition 470 (Unconditional Mean of MDS)

Let $(m_t, \mathcal{F}_t)_{t\geq 1}$ be a MDS. The unconditional mean of a MDS is:

$$\mathbb{E}[m_t] = 0. \tag{387}$$

Proof. Apply the law of iterated expectations and use the property of a MDS:

$$\mathbb{E}[m_t] = \mathbb{E}[\mathbb{E}[m_t|\mathcal{F}_{t-1}]] = \mathbb{E}[0] = 0.$$

□

Proposition 471 (Conditional Mean of a MDS)

Let $(m_t, \mathcal{F}_t)_{t\geq 1}$ be a MDS. Suppose that $s < t$. Then, the conditional mean is:

$$\mathbb{E}[m_t|\mathcal{F}_s] = 0. \tag{388}$$

Proof. Apply the law of iterated expectations:

$$\begin{aligned}\mathbb{E}[m_t|\mathcal{F}_s] &= \mathbb{E}[\mathbb{E}[m_t|\mathcal{F}_{t-1}]|\mathcal{F}_s] \\ &= \mathbb{E}[0|\mathcal{F}_s] \\ &= 0\end{aligned}$$

□

Proposition 472 (Autocovariance of a MDS)

Let $(m_t, \mathcal{F}_t)_{t\geq 1}$ be a MDS. Assume that $\mathbb{E}[m_t^2] < \infty$. Then, for $s < t$,

$$Cov[m_s, m_t] = 0. \tag{389}$$

Proof. Use the formula for covariance and the fact that the conditional mean of a MDS is 0:

$$\begin{aligned} Cov[m_s, m_t] &= \mathbb{E}[m_s m_t] - \mathbb{E}[m_s]\mathbb{E}[m_t] \\ &= \mathbb{E}[m_s m_t] - 0 \\ &= \mathbb{E}[\mathbb{E}[m_s m_t | \mathcal{F}_s]] \\ &= \mathbb{E}[m_s \underbrace{\mathbb{E}[m_t | \mathcal{F}_s]}_{=0}] \\ &= \mathbb{E}[m_s \cdot 0] \\ &= 0 \end{aligned}$$

□

Proposition 473 (A MDS Process is Serially Uncorrelated)
A MDS $(m_t, \mathcal{F}_t)_{t \geq 1}$ with $\mathbb{E}[m_t^2] < \infty$ is a serially **uncorrelated** sequence.

Proof. We have shown that the covariance of a MDS is zero. Therefore it implies it is uncorrelated. □

Hence, we can see that with a MDS, it is a relaxation of an i.i.d sequence.

Remark 474. *Note that the converse does not hold. That is, a serially uncorrelated process is not necessarily a MDS.*

Proposition 475 (i.i.d Sequence with Mean Zero)
Every i.i.d sequence with mean zero is a MDS.

Proof. We use the fact that m_t is independent and has mean zero to see:

$$\mathbb{E}[m_t | \mathcal{F}_{t-1}] = \mathbb{E}[m_t] = 0.$$

□

12.11 Limit Theorems for Martingale Difference Sequences

We can now state limit theorems for MDS.

Theorem 476 (Law of Large Numbers for Martingale Difference Sequence)
Let $(m_t, \mathcal{F}_t)$ be a martingale difference sequence. If either of:

1. $\sum_{t=1}^{\infty} \frac{\mathbb{E}[|m_t|^{1+p}]}{t^{1+p}} < \infty$ for some $p \in [0, 1]$
2. $\lim_{T \to \infty} \max_{t \leq T} [\mathbb{E}[|m_t|^{1+p}]] < \infty$ for some $p \in [0, 1]$,

then we have that:

$$\frac{1}{T} \sum_{t=1}^{T} m_t \xrightarrow{P} \mathbb{E}[m_t] = 0. \tag{390}$$

Remark 477. *Recall that we have shown that the unconditional mean of a MDS $\mathbb{E}[m_t] = 0$. Hence, this result makes sense.*

We now give an example of how to check these conditions.

Example 478

Let m_t be an i.i.d sequence with $\mathbb{E}[m_t^2] = \sigma^2$. Then, let $p = 1$. We then check the first condition.

$$\begin{aligned}\sum_{t=1}^{\infty} \frac{\mathbb{E}[|m_t|^2]}{t^2} &= \sum_{t=1}^{\infty} \frac{\mathbb{E}[m_t^2]}{t^2} \\ &= \sum_{t=1}^{\infty} \frac{\sigma^2}{t^2} \\ &= \sigma^2 \sum_{t=1}^{\infty} \frac{1}{t^2} \\ &= \sigma^2 \frac{\pi^2}{6} < \infty.\end{aligned}$$

Hence, the first condition is satisfied so we can apply the LLN to m_t.

We now show an application of the second condition.

$$\begin{aligned}\lim_{T\to\infty} \max_{t\leq T} [\mathbb{E}[|m_t|^2]] &= \lim_{T\to\infty} \max_{t\leq T} \sigma^2 \\ &= \sigma^2 < \infty.\end{aligned}$$

Therefore, we see that the second condition is satisfied. Therefore we can apply the law of large number to m_t.

We can now state the central limit theorem for MDS.

Theorem 479 (Central Limit Theorem for Martingale Difference Sequence)

Let $(m_t, \mathcal{F}_t)$ be a martingale difference sequence. Let $S_T^2 = \sum_{t=1}^{T} \mathbb{E}[m_t^2]$. Then, if:

1. $\frac{1}{S_T^2} \sum_{t=1}^{T} \mathbb{E}[m_t^2 \mid \mathcal{F}_{t-1}] \xrightarrow{P} 1$ and
2. $\frac{1}{S_T^2} \sum_{t=1}^{T} \mathbb{E}\left[m_t^2 1\{|m_t| > \delta S_T\}\right] \to 0$ for every $\delta > 0$,

then we have that:

$$\frac{1}{S_T} \sum_{t=1}^{T} m_t \xrightarrow{D} \mathcal{N}(0, 1). \tag{391}$$

Remark 480. *The second condition is known as a* **Lindeberg condition**. *Its intuition is that no unusually large martingale difference should dominate the normalized sum.*

We give an example for checking this.

Example 481

Suppose that m_t is i.i.d with mean zero and constant second moment $\mathbb{E}[m_t^2] = \sigma^2$. Then,

$$S_T^2 = \sum_{t=1}^{T} \mathbb{E}[m_t^2] = \sum_{t=1}^{T} \sigma^2 = T\sigma^2.$$

Then,

$$S_T = \sqrt{T\sigma^2}$$

and hence:

$$\frac{1}{\sqrt{T\sigma^2}} \sum_{t=1}^{T} m_t \xrightarrow{D} \mathcal{N}(0,1).$$

12.12 Limit Theorems for AR(1) Model

We have stated the limit theorems for MDS. We now want to derive limit theorems for the AR(1) model which is what is generally used in this course. To derive these AR(1) limit theorem, we can use the limit theorems for MDS.

First we give a motivation for why we need limit theorems for AR(1) models. Suppose we are interested in testing $H_0 : \mu_Y = 0$ in the model:

$$Y_t = \mu_Y + u_t$$

where we have a correlated error term:

$$u_t = \alpha u_{t-1} + \epsilon_t.$$

If we **incorrectly assume** that u_t is i.i.d, then we would then incorrectly assume that Y_t is i.i.d. Then, when testing $H_0 : \mu_Y = 0$, we would construct the ordinary test statistic:

$$t = \frac{T^{1/2}(\overline{Y} - \mu_Y)}{S_y} \xrightarrow{D} \mathcal{N}(0,1)$$

where S_y is the standard deviation of Y. However, as we will later prove, if we assumed instead correlated error terms, we would instead end with the limiting distribution:

$$t = \frac{T^{1/2}(\overline{Y} - \mu_Y)}{S_y} \xrightarrow{D} \mathcal{N}(0, \frac{1+\alpha}{1-\alpha}).$$

We can see that we end up with two different distributions:

$$\mathcal{N}(0,1) \quad \text{and} \quad \mathcal{N}(0, \frac{1+\alpha}{1-\alpha}).$$

We plot these distributions in Fig. 2. If $\alpha > 0$, then $\frac{1+\alpha}{1-\alpha} > 1$, so the correctly derived limiting distribution has thicker tails than $\mathcal{N}(0,1)$. Using standard-normal critical values then leads to rejecting the null hypothesis too often. If $\alpha < 0$, the variance ratio is below one and the standard-normal test is instead conservative. The two distributions coincide when $\alpha = 0$, in which case the errors are uncorrelated.

With this motivation in mind, we now aim to derive the limit theorems for AR(1) models.

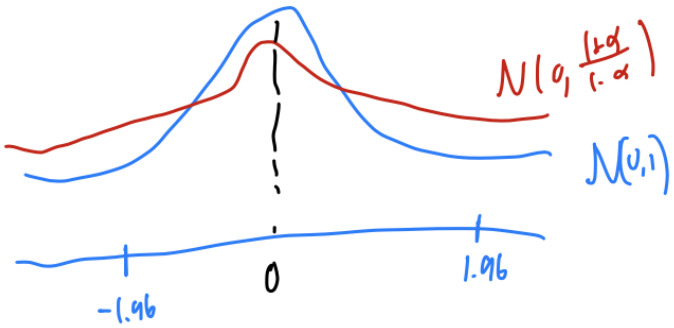

Figure 2: Comparison of the two limiting distributions of the same test statistic under different assumptions.

Theorem 482 (Law of Large Numbers for AR(1))

Let X_t be an AR(1) process

$$X_t = \alpha X_{t-1} + \epsilon_t.$$

Assume that

1. $|\alpha| < 1$
2. ϵ_t is i.i.d with mean zero, variance σ^2, and finite fourth moment
3. X_0 is fixed, or is independent of the innovations and has a finite fourth moment.

Then, for any fixed non-negative integer j, the three laws of large numbers for AR(1) processes are:

$$\frac{1}{T}\sum_{t=1}^{T} X_{t-j} \xrightarrow{P} 0 \tag{392}$$

$$\frac{1}{T}\sum_{t=1}^{T} X_{t-j}^2 \xrightarrow{P} \frac{\sigma^2}{1-\alpha^2} \equiv \sigma_X^2 \tag{393}$$

$$\frac{1}{T}\sum_{t=1}^{T} X_{t-1}\epsilon_t \xrightarrow{P} \mathbb{E}[X_{t-1}\epsilon_t] = 0. \tag{394}$$

Proof. The general strategy for proving each of these results is to first rewrite the expression as a MDS. Then, apply either the MDS law of large numbers or MDS central limit theorem.

Proof of 1

First subtract αX_t from both sides of the AR(1) model:

$$\begin{aligned}
&X_t = \alpha X_{t-1} + \epsilon_t \\
&\leftrightarrow (1-\alpha)X_t = -\alpha(X_t - X_{t-1}) + \epsilon_t \\
&\leftrightarrow X_t = \frac{-\alpha}{(1-\alpha)}(X_t - X_{t-1}) + \frac{1}{(1-\alpha)}\epsilon_t \\
&\leftrightarrow X_t = \frac{\alpha}{(\alpha-1)}(X_t - X_{t-1}) + \frac{1}{(1-\alpha)}\epsilon_t
\end{aligned}$$

where the second last equation is valid as $|\alpha| < 1$. Then we sum both sides up to T:

$$\sum_{t=1}^{T} X_t = \frac{\alpha}{\alpha - 1} \sum_{t=1}^{T} \Big(X_t - X_{t-1} \Big) + \frac{1}{1-\alpha} \sum_{t=1}^{T} \epsilon_t$$
$$\leftrightarrow \sum_{t=1}^{T} X_t = \frac{\alpha}{\alpha - 1} \Big(X_T - X_0 \Big) + \frac{1}{1-\alpha} \sum_{t=1}^{T} \epsilon_t.$$

where we used the fact that $\sum_{t=1}^{T} \Big(X_t - X_{t-1} \Big)$ is a telescoping sum. Then divide by T on both sides:

$$\frac{1}{T} \sum_{t=1}^{T} X_t = \frac{1}{T} \frac{\alpha}{\alpha - 1} \Big(X_T - X_0 \Big) + \frac{1}{1-\alpha} \frac{1}{T} \sum_{t=1}^{T} \epsilon_t.$$

The final step is to now show that in:

$$\frac{1}{T} \sum_{t=1}^{T} X_t = \underbrace{\frac{1}{T} \frac{\alpha}{\alpha - 1} \Big(X_T - X_0 \Big)}_{(*)} + \underbrace{\frac{1}{1-\alpha} \frac{1}{T} \sum_{t=1}^{T} \epsilon_t}_{(**)}.$$

that $(*) = o_p(1)$ and $(**) \xrightarrow{p} 0$.

First, we look at (*):

$$\frac{\alpha}{\alpha - 1} \frac{1}{T} \Big(X_T - X_0 \Big)$$

As X_0 does not depend on T, it is a fixed random variable and therefore $X_0 = O_p(1)$ and therefore $\frac{1}{T} X_0 = o_p(1)$. We now look at the second term:

$$\frac{1}{T} X_T.$$

We need to show that this is also $o_p(1)$ which means showing:

$$\lim_{T \to \infty} \mathbb{P}\big[\frac{|X_T|}{T} > \delta \big] = 0.$$

To do so, we can apply **Markov's inequality**:

$$\mathbb{P}\left[\frac{|X_T|}{T} > \delta \right] \leq \frac{1}{\delta^2 T^2} \mathbb{E}[X_T^2]$$
$$\leq \frac{C}{\delta^2 T^2} \to 0$$

as $T \to \infty$, where stability and the moment assumptions imply $\sup_T \mathbb{E}[X_T^2] \leq C < \infty$. Hence, we have just shown that $\frac{1}{T} X_T = o_p(1)$. We then use the fact that:

$$o_p(1) + o_p(1) = o_p(1)$$

to conclude that

$$\frac{\alpha}{\alpha - 1} \frac{1}{T} \Big(X_T - X_0 \Big) = o_p(1).$$

Now look at (**):

$$\frac{1}{T} \sum_{t=1}^{T} \epsilon_t.$$

This is a MDS. We can therefore apply the **MDS law of large numbers**:

$$\frac{1}{T}\sum_{t=1}^{T}\epsilon_t \xrightarrow{P} \mathbb{E}[\epsilon_t] = 0.$$

Hence, we shown the desired results for (*) and (**) so we have that:

$$\frac{1}{T}\sum_{t=1}^{T}X_t = \underbrace{\frac{1}{T}\frac{\alpha}{\alpha-1}\Big(X_T - X_0\Big)}_{(*)} + \underbrace{\frac{1}{1-\alpha}\frac{1}{T}\sum_{t=1}^{T}\epsilon_t}_{(**)}$$

$$\leftrightarrow \frac{1}{T}\sum_{t=1}^{T}X_t = o_p(1) + o_p(1)$$

and hence:

$$\frac{1}{T}\sum_{t=1}^{T}X_t \xrightarrow{P} 0.$$

Proof of 3

We first prove the third AR(1) law of large numbers before proving the second AR(1) law of large numbers. We want to show that:

$$\frac{1}{T}\sum_{t=1}^{T}X_{t-1}\epsilon_t \xrightarrow{P} \mathbb{E}[X_{t-1}\epsilon_t] = 0.$$

First, we notice that the summand is a martingale difference sequence:

$$\mathbb{E}[X_{t-1}\epsilon_t|\mathcal{F}_{t-1}] = X_{t-1}\mathbb{E}[\epsilon_t|\mathcal{F}_{t-1}] = 0.$$

Therefore, we want to apply the law of large numbers for martingale differences to $X_{t-1}\epsilon_t$. We check the bounded-moment condition with $p = 1$:

$$\begin{aligned}\mathbb{E}[(X_{t-1}\epsilon_t)^2] &= \mathbb{E}[\mathbb{E}[(X_{t-1}\epsilon_t)^2|\mathcal{F}_{t-1}]] \\ &= \mathbb{E}[X_{t-1}^2\mathbb{E}[\epsilon_t^2|\mathcal{F}_{t-1}]] \\ &= \sigma^2\mathbb{E}[X_{t-1}^2].\end{aligned}$$

Stability and the moment assumptions give $\sup_t \mathbb{E}[X_{t-1}^2] < \infty$, so

$$\sup_t \mathbb{E}[(X_{t-1}\epsilon_t)^2] = \sigma^2 \sup_t \mathbb{E}[X_{t-1}^2] < \infty.$$

Therefore, we can apply the WLLN for MDS and conclude that:

$$\frac{1}{T}\sum_{t=1}^{T}X_{t-1}\epsilon_t \xrightarrow{P} \mathbb{E}[X_{t-1}\epsilon_t] = 0.$$

Proof of 2

We now want to show:

$$\frac{1}{T}\sum_{t=1}^{T}X_{t-j}^2 \xrightarrow{P} \mathbb{E}[X_t^2] = Var[X_t] = \frac{\sigma^2}{1-\alpha^2} \equiv \sigma_X^2$$

First, if we expand out the square we get:

$$X_t^2 = \alpha^2 X_{t-1}^2 + \epsilon_t^2 + 2\alpha X_{t-1}\epsilon_t.$$

Now, if we take the sum of both sides, we get:

$$\sum_{t=1}^{T} X_t^2 = \alpha^2 \sum_{t=1}^{T} X_{t-1}^2 + \sum_{t=1}^{T} \epsilon_t^2 + 2\alpha \sum_{t=1}^{T} X_{t-1}\epsilon_t.$$

We then factorise this to get:

$$\sum_{t=1}^{T} X_t^2 = \alpha^2 \left[\sum_{t=1}^{T} X_t^2 + X_0^2 - X_T^2 \right] + \sum_{t=1}^{T} \epsilon_t^2 + 2\alpha \sum_{t=1}^{T} X_{t-1}\epsilon_t.$$

whereby we used the fact that: $\sum_{t=1}^{T} X_{t-1}^2 = \sum_{t=1}^{T} X_t^2 + X_0^2 - X_T^2$. Re-arrange to get:

$$\sum_{t=1}^{T} X_t^2 - \alpha^2 \sum_{t=1}^{T} X_t^2 = \alpha^2 \left[X_0^2 - X_T^2 \right] + \sum_{t=1}^{T} \epsilon_t^2 + 2\alpha \sum_{t=1}^{T} X_{t-1}\epsilon_t$$

Factorise and divide through to get:

$$\sum_{t=1}^{T} X_t^2 = \frac{\alpha^2}{1-\alpha^2} \left[X_0^2 - X_T^2 \right] + \frac{\sum_{t=1}^{T} \epsilon_t^2}{1-\alpha^2} + \frac{2\alpha}{1-\alpha^2} \sum_{t=1}^{T} X_{t-1}\epsilon_t$$

We then divide by T on both sides:

$$\frac{1}{T} \sum_{t=1}^{T} X_t^2 = \frac{\alpha^2}{1-\alpha^2} \left[\frac{X_0^2}{T} - \frac{X_T^2}{T} \right] + \frac{1}{T} \frac{\sum_{t=1}^{T} \epsilon_t^2}{1-\alpha^2} + \frac{1}{T} \frac{2\alpha}{1-\alpha^2} \sum_{t=1}^{T} X_{t-1}\epsilon_t$$

We now look at each term respectively. First, we know that:

$$\frac{X_0^2}{T} = o_p(1)$$

since $o(1)O_p(1) = o(1) = o_p(1)$. The second term is:

$$\frac{X_T^2}{T} = o_p(1)$$

through Markov's inequality. We then can conclude that:

$$\left[\frac{X_0^2}{T} - \frac{X_T^2}{T} \right] = o_p(1).$$

For the third term, we can apply the WLLN for i.i.d sequence:

$$\frac{1}{T} \sum_{t=1}^{T} \epsilon_t^2 \xrightarrow{P} \sigma^2.$$

The final term, we see that

$$\frac{1}{T} \sum_{t=1}^{T} X_{t-1}\epsilon_t \xrightarrow{P} 0$$

whereby we applied the part 3 result proved earlier for the LLN for the AR(1) model.

By Slutsky theorem, we combine all these results to conclude that:

$$\frac{1}{T}\sum_{t=1}^{T} X_t^2 = \frac{\alpha^2}{1-\alpha^2}\left[\frac{X_0^2}{T} - \frac{X_T^2}{T}\right] + \frac{1}{T}\frac{\sum_{t=1}^{T}\epsilon_t^2}{1-\alpha^2} + \frac{1}{T}\frac{2\alpha}{1-\alpha^2}\sum_{t=1}^{T} X_{t-1}\epsilon_t \xrightarrow{P} 0 + \frac{\sigma^2}{1-\alpha^2} + 0$$

□

Theorem 483 (Central Limit Theorem for AR(1))
Let $X_t = \alpha X_{t-1} + \epsilon_t$, where $|\alpha| < 1$, the innovations are i.i.d with mean zero, variance σ^2, and finite fourth moment, and X_0 satisfies the moment assumptions above. For any fixed non-negative integer j, the two central limit theorems for AR(1) processes are:

$$T^{-1/2}\sum_{t=1}^{T} X_{t-j} \xrightarrow{D} \mathcal{N}\big(0, \frac{\sigma^2}{(1-\alpha)^2}\big). \tag{395}$$

$$T^{-1/2}\sum_{t=1}^{T} X_{t-1}\epsilon_t \xrightarrow{D} \mathcal{N}\big(0, \frac{\sigma^4}{1-\alpha^2}\big). \tag{396}$$

Proof. We give the proof for the first result. First, recall the AR(1) model:

$$X_t = \alpha X_{t-1} + \epsilon_t$$

Then take sums:

$$\sum_{t=1}^{T} X_t = \alpha\sum_{t=1}^{T} X_{t-1} + \sum_{t=1}^{T}\epsilon_t.$$

Use our trick to rewrite the term as:

$$\sum_{t=1}^{T} X_t = \alpha\sum_{t=1}^{T} X_t + \alpha X_0 - \alpha X_T + \sum_{t=1}^{T}\epsilon_t.$$

Re-arrange:

$$(1-\alpha)\sum_{t=1}^{T} X_t = \alpha X_0 - \alpha X_T + \sum_{t=1}^{T}\epsilon_t.$$

and divide through:

$$\sum_{t=1}^{T} X_t = \frac{\alpha}{(1-\alpha)}X_0 - \frac{\alpha}{(1-\alpha)}X_T + \frac{1}{(1-\alpha)}\sum_{t=1}^{T}\epsilon_t.$$

Now scale by $T^{-1/2}$ to get:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t = \frac{1}{\sqrt{T}}\frac{\alpha}{(1-\alpha)}X_0 - \frac{1}{\sqrt{T}}\frac{\alpha}{(1-\alpha)}X_T + \frac{1}{\sqrt{T}}\frac{1}{(1-\alpha)}\sum_{t=1}^{T}\epsilon_t.$$

We now analyse this term by term. For the first term on the right, we have that:

$$\frac{1}{\sqrt{T}}\frac{\alpha}{(1-\alpha)}X_0 = o_p(1).$$

For the second term, we have that:

$$\frac{1}{\sqrt{T}}\frac{\alpha}{(1-\alpha)}X_T = o(1)O_p(1) = o_p(1).$$

For the final term, we can use the regular central limit theorem for i.i.d sequence:

$$\frac{1}{\sqrt{T}}\frac{1}{(1-\alpha)}\sum_{t=1}^{T}\epsilon_t \xrightarrow{D} \frac{1}{1-\alpha}\mathcal{N}(0,\sigma^2) = \mathcal{N}\big(0, \frac{\sigma^2}{(1-\alpha)^2}\big).$$

Putting all this together, we conclude that:

$$T^{-1/2}\sum_{t=1}^{T}X_{t-j} \xrightarrow{D} \mathcal{N}\big(0, \frac{\sigma^2}{(1-\alpha)^2}\big).$$

□

We now go back to our original AR(1) t-test. Let our AR(1) model with an intercept be:

$$Y_t = \alpha Y_{t-1} + \mu + \epsilon_t.$$

We can rewrite this as an AR(1) model without an intercept:

$$Y_t = \mu_Y + X_t$$

where

$$X_t = \alpha X_{t-1} + \epsilon_t$$

is an AR(1) model without an intercept. We now want to look at the test statistic to test the intercept:

$$H_0 : \mu_Y = 0.$$

The t-test statistic for this is:

$$t = \frac{\overline{Y} - 0}{\sqrt{\frac{s_Y^2}{T}}}$$

where the variance can be written as:

$$s_Y^2 = \frac{1}{T-1}\sum_{t=1}^{T}\big(Y_t - \overline{Y}\big)^2$$

whereby

$$\overline{Y} = \frac{1}{T}\sum_{t=1}^{T}Y_t = \frac{1}{T}\sum_{t=1}^{T}(\mu_Y + X_t) = \mu_Y + \overline{X}.$$

Now, also note that:

$$\begin{aligned} T^{1/2}\Big(\overline{Y} - \mu_Y\Big) &= T^{1/2}\overline{X} \\ &= T^{-1/2}\sum_{t=1}^{T}X_t \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2}{(1-\alpha)^2}\right) \end{aligned}$$

where we used the CLT for AR(1) model. We therefore have that the numerator converges to a normal distribution. Now, if we look the sample variance:

$$Y_t - \overline{Y} \quad \leftrightarrow \quad (\mu_Y + X_t) - (\mu_Y + \overline{X}) = X_t - \overline{X}.$$

Now, we look at the limiting result of s_Y^2:

$$\begin{aligned} s_Y^2 &= \frac{1}{T-1}\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2 \\ &= \frac{T}{T-1}\left[\frac{1}{T}\sum_{t=1}^{T}X_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T}X_t\right)^2\right] \xrightarrow{P} \frac{\sigma^2}{1-\alpha^2}. \end{aligned}$$

where the first term inside the brackets uses the LLN for AR(1), and the second term uses the LLN for AR(1) plus the continuous mapping theorem.

Under $H_0 : \mu_Y = 0$, we therefore rewrite the t-statistic as:

$$\begin{aligned} t = \frac{T^{1/2}\overline{Y}}{s_Y} &\xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2/(1-\alpha)^2}{\sigma^2/(1-\alpha^2)}\right) \\ &= \mathcal{N}\left(0, \frac{1+\alpha}{1-\alpha}\right). \end{aligned}$$

where we again use Slutsky's theorem.

12.13 Higher Order Autoregression

12.13.1 AR(2) Model

So far, we have only studied the $AR(1)$ model:

$$Y_t = \alpha Y_{t-1} + \mu + \epsilon_t.$$

However, this is quite a restrictive assumption. We now aim to relax this assumption. We first look at the AR(2) model.

Definition 484 (AR(2) With Intercept)
The AR(2) with intercept model is given by:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \mu + \epsilon_t \tag{397}$$

where μ is an intercept term, $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ and independent of $Y_{t-1}, ..., Y_0$, and we also require two initial observations Y_{-1}, Y_0.

Note that the AR(2) model can also be written in vector notation:

$$Y_t = \boldsymbol{\alpha}^T \boldsymbol{Y}_{t-1} + \mu + \epsilon_t$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$ and $\boldsymbol{Y}_{t-1} = (Y_{t-1}, Y_{t-2})^T$. We can estimate such a model through the least squares estimator and in fact, coincides with the conditional MLE.

Now, notice that our AR(2) model has an intercept term:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \mu + \epsilon_t.$$

We want to rewrite this as an AR(2) model without an intercept. The next proposition tells us how to do this.

Proposition 485 (AR(2) Model without an Intercept)
The AR(2) model with an intercept can be written as an AR(2) model without an intercept whereby Y_t is now the sum of a mean-zero process and its long-run mean μ_Y:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \epsilon_t \tag{398}$$

$$Y_t = X_t + \mu_Y \tag{399}$$

$$\mu_Y = \frac{\mu}{1 - \alpha_1 - \alpha_2} \tag{400}$$

Proof. First, subtract μ_Y from both sides:

$$Y_t - \mu_Y = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \mu + \epsilon_t - \mu_Y$$

Now add and subtract $\alpha_1 \mu_Y$ and $\alpha_2 \mu_Y$ from the right hand side:

$$Y_t - \mu_Y = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \mu + \epsilon_t - \mu_Y + \alpha_1 \mu_Y - \alpha_1 \mu_Y + \alpha_2 \mu_Y - \alpha_2 \mu_Y.$$

Now collect terms:

$$Y_t - \mu_Y = \alpha_1(Y_{t-1} - \mu_Y) + \alpha_2(Y_{t-2} - \mu_Y) + \Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y\Big] + \epsilon_t$$

Now, note that $Y_t - \mu_Y$ will have mean zero. We define the mean zero process:

$$X_t \equiv Y_t - \mu_Y$$

to that our expression now becomes:

$$Y_t - \mu_Y = \alpha_1(Y_{t-1} - \mu_Y) + \alpha_2(Y_{t-2} - \mu_Y) + \Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y\Big] + \epsilon_t$$
$$\leftrightarrow X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y\Big] + \epsilon_t$$

Hence, we have the expression now:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y\Big] + \epsilon_t$$

If we look carefully, nearly all the terms have mean zero:

$$\underbrace{X_t}_{\text{Mean Zero}} = \alpha_1 \underbrace{X_{t-1}}_{\text{Mean Zero}} + \alpha_2 \underbrace{X_{t-2}}_{\text{Mean Zero}} + \underbrace{\Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y\Big]}_{(*)} + \underbrace{\epsilon_t}_{\text{Mean Zero}}.$$

We now look at the remaining (*) term. Now, as the left-hand side has mean zero, and all the terms on the right-hand side has mean zero, since (*) is a constant, we therefore require that it be equal to the zero in order for the expectation of (*) to be zero. That is, we impose that:

$$\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y = 0.$$

We then re-arrange to solve for the long-run mean μ_Y to get:

$$\mu_Y = \frac{\mu}{1 - \alpha_1 - \alpha_2}.$$

Now, if we go back to:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \underbrace{\Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y\Big]}_{(*)} + \epsilon_t$$

since we have imposed that (*) = 0, we get:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \epsilon_t$$

which is an AR(2) process without an intercept!

Hence, to summarise what we have done, we took an AR(2) model with an intercept:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \mu + \epsilon_t$$

and rewrote it as an AR(2) model without an intercept whereby Y_t is now the sum of a mean-zero process and its

long-run mean:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \epsilon_t$$
$$Y_t = X_t + \mu_Y$$
$$\mu_Y = \frac{\mu}{1 - \alpha_1 - \alpha_2}$$

□

We have just shown how to write an AR(2) model with an intercept Y_t into an AR(2) model without an intercept X_t. Next, we want to be able to write this AR(2) model without an intercept as an AR(1) model without an intercept. Writing an AR(2) model in its **companion form** allows us to do this.

Definition 486 (Companion Form of AR(2))
Let X_t be an AR(2) model:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \epsilon_t$$

Then, the **companion form** of the AR(2) model can be written as:

$$\begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \epsilon_t \tag{401}$$

or in matrix form:

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \tau\epsilon_t \tag{402}$$

whereby $\tau = (1, 0)^T$.

The companion form $\boldsymbol{X}_t$ now allows us to use similar techniques that we used for AR(1) processes. Recall that for an AR(1) model:

$$X_t = \alpha X_{t-1} + \epsilon_t$$

we could iteratively back-substitute to get the recursive form:

$$X_t = \sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0.$$

We can do something similar for the companion form $\boldsymbol{X}_t$ of the AR(2) model.

Proposition 487 (Recursive form of AR(2) Model)
Let $\boldsymbol{X}_t$ be the companion form of the AR(2) process X_t. Then, the recursive solution to $\boldsymbol{X}_t$ is:

$$\boldsymbol{X}_t = \sum_{j=0}^{t-1} \boldsymbol{A}^j \tau \epsilon_{t-j} + \boldsymbol{A}^t \boldsymbol{X}_0. \tag{403}$$

Jumping back to the AR(1) model:

$$X_t = \sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t X_0,$$

we showed that if the AR(1) possessed a stationarity solution ($|\alpha| < 1$) we can then take limits and under Kolmogorov 2 series criterion, write this as:

$$X_t = \sum_{j=0}^{\infty} \alpha^j \epsilon_{t-j}.$$

The question we are interested now in asking is that for our AR(2) companion form, can we write:

$$\boldsymbol{X}_t = \sum_{j=0}^{t-1} \boldsymbol{A}^j \tau \epsilon_{t-j} + \boldsymbol{A}^t \boldsymbol{X}_0$$

in terms of an infinite moving average representation:

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \epsilon_{t-j}.$$

The answer is yes whereby we now need to check the condition that $\boldsymbol{A}$ is *smaller than 1*. We state what this means for matrices.

Proposition 488 (Existence of Stationary Solution for AR(2) Model)
Let $\boldsymbol{X}_t$ be the AR(2) in companion form with the companion matrix:

$$\boldsymbol{A} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix}.$$

Then, if all eigenvalues of $\boldsymbol{A}$ have modulus strictly less than 1 in absolute value, then the AR(2) model possesses a stationary solution.

Remark 489. *Technically speaking, this is the condition for stability which then implies stationarity under some additional assumptions.*

Remark 490. *We recall some facts from linear algebra. A matrix $\boldsymbol{A}$ is a linear function. We are interested in knowing which vectors are only stretched by the transformation $\boldsymbol{v}$. This stretching is determined by the eigenvalues λ. Hence, we want to know the vectors $\boldsymbol{v}$ such that:*

$$\boldsymbol{A}v = \lambda \boldsymbol{v}$$

so that the vectors are only stretched. We can then rewrite this as:

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{v} = \boldsymbol{0}$$

This equation has a non-zero solution vector $\boldsymbol{v}$ if and only if

$$det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0.$$

That is, the eigenvalues of $\boldsymbol{A}$ are the values λ that satisfy the equation:

$$det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0.$$

By the fundamental theorem of algebra, we can write this as the product:

$$det(\boldsymbol{A} - \lambda \boldsymbol{I}) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)...(\lambda_n - \lambda) = 0$$

where n is the degree of the characteristic polynomial.

Recall that to compute the eigenvalues for the matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix}$$

you first compute the **characteristic polynomial** (recall that the roots to the characteristic polynomial are the eigenvalues) from computing the determinant of $\boldsymbol{A}$:

$$0 = det(\boldsymbol{A} - \lambda \boldsymbol{I}) = (\alpha_1 - \lambda)(-\lambda) - \alpha_2 = \lambda^2 - \alpha_1\lambda - \alpha_2.$$

So we then check the roots for:

$$\lambda^2 - \alpha_1\lambda - \alpha_2 = \lambda^2[1 - \alpha_1\frac{1}{\lambda} - \alpha_2\frac{1}{\lambda^2}] = \lambda^2\alpha(1/\lambda) = 0$$

where $\alpha(z) = 1 - \alpha_1 z - \alpha_2 z^2$ is the autoregressive polynomial. As

$$det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \lambda^2\alpha(1/\lambda) = 0$$

then this implies that the eigenvalues satisfy:

$$\alpha(1/\lambda) = 1 - \alpha_1\frac{1}{\lambda} - \alpha_2\frac{1}{\lambda^2} = 0.$$

If all roots λ that solves the characteristic polynomial has a modulus strictly less than one, then all eigenvalues of $\boldsymbol{A}$ are strictly less than one, and this means that the all entries of the matrix

$$\boldsymbol{A}^t \to 0$$

at an exponential rate.

Remark 491. *Note that we can also equivalently express the condition for stationarity as:*

$$1 - \alpha_1 z - \alpha_2 z^2 = 0$$

and check if all roots has modulus strictly **greater** *than one.*

Example 492

(Root Calculation). Suppose we had the autoregressive polynomial:

$$a(L) = (1 + 0.3L - 0.1L^2).$$

Now, we can factorise this:

$$(1 + 0.3L - 0.1L^2) = (1 + 0.5L)(1 - 0.2L) = 0.$$

Then, solving for roots, we have that:

$$1 + 0.5L = 0 \quad \leftrightarrow \quad L = -2$$

$$1 - 0.2L = 0 \quad \leftrightarrow \quad L = 5$$

Both roots have modulus greater than 1 so they are outside the unit circle and hence we have stability in the system.

Hence, going back to our AR(2) model:

$$\boldsymbol{X}_t = \sum_{j=0}^{t-1} \boldsymbol{A}^j \tau \epsilon_{t-j} + \boldsymbol{A}^t \boldsymbol{X}_0$$

if all eigenvalues of $\boldsymbol{A}$ are strictly less than one, we will have that $\boldsymbol{A}^t \boldsymbol{X}_0 \to 0$ and in combination with Kolmogorov's 2 series criterion, we have that:

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \epsilon_{t-j}.$$

Proposition 493 (Moving Average Representation of AR(2))
Let $\boldsymbol{X}_t$ be the AR(2) model in companion form with the companion matrix $\boldsymbol{A}$. Suppose that all the eigenvalues of $\boldsymbol{A}$ has modulus strictly less than 1. Then, $\boldsymbol{X}_t$ can be represented in the infinite moving average form:

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \epsilon_{t-j} \tag{404}$$

where $\tau = (1, 0)^T$.

Now that we have this moving average representation of the AR(2) model, we can now compute the statistical properties of this model just like we did for the AR(1) model.

Proposition 494 (Statistical Properties of AR(2) Model)
Let $\boldsymbol{X}_t$ be the AR(2) model in companion form. Suppose that $\boldsymbol{X}_t$ admits a stationary solution so that it has the infinite order representation:

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \epsilon_{t-j} \tag{405}$$

where $\tau = (1, 0)^T$. Then, we have the following statistical properties of $\boldsymbol{X}_t$:

$$\mathbb{E}[\boldsymbol{X}_t] = 0 \tag{406}$$

$$\Sigma_X \equiv Var[\boldsymbol{X}_t] = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \sigma^2 \tau^T (\boldsymbol{A}^j)^T \tag{407}$$

$$Cov[\boldsymbol{X}_t, \boldsymbol{X}_{t-s}] = \boldsymbol{A}^s \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \sigma^2 \tau^T (\boldsymbol{A}^j)^T \tag{408}$$

We can in fact get a distribution for the AR(2) model.

Proposition 495 (Distribution of AR(2) Model)

Let $\boldsymbol{X}_t$ be the AR(2) model in companion form. Suppose that $\boldsymbol{X}_t$ admits a stationary solution so that it has the infinite order representation:

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \epsilon_{t-j} \tag{409}$$

where $\tau = (1,0)^T$ and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. Then, the distribution of X_t:

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \epsilon_{t-j} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_X) \tag{410}$$

whereby $\Sigma_X \equiv Var[\boldsymbol{X}_t] = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \sigma^2 \tau^T (\boldsymbol{A}^j)^T$.

12.13.2 AR(p) Model

We now generalise from the AR(2) model to the AR(p) model. Many of the results we derived for the AR(2) model hold for the AR(p) model.

Definition 496 (AR(p) Model with Intercept)

The AR(p) model with an intercept is given by:

$$Y_t = \sum_{j=1}^{p} \alpha_j Y_{t-j} + \mu + \epsilon_t \tag{411}$$

where μ is an intercept term, $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ and independent of $Y_{t-1}, ..., Y_0$, and we also require p initial observations.

We show that we can write the AR(p) model with an intercept as an AR(p) model without an intercept just like we did for the AR(2) model.

Proposition 497 (AR(p) Model without an Intercept)

The AR(p) model with an intercept can be written as an AR(p) model without an intercept whereby Y_t is now the sum of a mean-zero process and its long-run mean μ_Y:

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \epsilon_t \tag{412}$$

$$Y_t = X_t + \mu_Y \tag{413}$$

$$\mu_Y = \frac{\mu}{1 - \alpha_1 - ... - \alpha_p} \tag{414}$$

Proof. The proof is the same for AR(2) model except we subtract off more terms. First, subtract μ_Y from both sides:

$$Y_t - \mu_Y = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \mu + \epsilon_t - \mu_Y$$

Now add and subtract $\alpha_j \mu_Y$ for $j = 1, ..., p$ from the right hand side:

$$Y_t - \mu_Y = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \mu + \epsilon_t - \mu_Y + \alpha_1 \mu_Y - \alpha_1 \mu_Y + \alpha_2 \mu_Y - \alpha_2 \mu_Y + ... + \alpha_p \mu_Y - \alpha_p \mu_Y$$

Now collect terms:

$$Y_t - \mu_Y = \alpha_1(Y_{t-1} - \mu_Y) + \alpha_2(Y_{t-2} - \mu_Y) + ... + \alpha_p(Y_{t-p} - \mu_Y) + \Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y + ... + \alpha_p\mu_Y\Big] + \epsilon_t$$

Now, note that $Y_t - \mu_Y$ will have mean zero. We define the mean zero process:

$$X_t \equiv Y_t - \mu_Y$$

to that our expression now becomes:

$$Y_t - \mu_Y = \alpha_1(Y_{t-1} - \mu_Y) + \alpha_2(Y_{t-2} - \mu_Y) + ... + \alpha_p(Y_{t-p} - \mu_Y) + \Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y + ... + \alpha_p\mu_Y\Big] + \epsilon_t$$

$$\leftrightarrow X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y + ... + \alpha_p\mu_Y\Big] + \epsilon_t$$

Hence, we have the expression now:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y + ... + \alpha_p\mu_Y\Big] + \epsilon_t$$

If we look carefully, nearly all the terms have mean zero:

$$\underbrace{X_t}_{\text{Mean Zero}} = \alpha_1 \underbrace{X_{t-1}}_{\text{Mean Zero}} + \alpha_2 \underbrace{X_{t-2}}_{\text{Mean Zero}} + ... + \alpha_p \underbrace{X_{t-p}}_{\text{Mean Zero}} + \underbrace{\Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y + ... + \alpha_p\mu_Y\Big]}_{(*)} + \underbrace{\epsilon_t}_{\text{Mean Zero}}.$$

We now look at the remaining (*) term. Now, as the left-hand side has mean zero, and all the terms on the right-hand side has mean zero, since (*) is a constant, we therefore require that it be equal to the zero in order for the expectation of (*) to be zero. That is, we impose that:

$$\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y + ... + \alpha_p\mu_Y = 0.$$

We then re-arrange to solve for the long-run mean μ_Y to get:

$$\mu_Y = \frac{\mu}{1 - \alpha_1 - \alpha_2 - ... - \alpha_p}.$$

Now, if we go back to:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \underbrace{\Big[\mu - \mu_Y + \alpha_1\mu_Y + \alpha_2\mu_Y + ... + \alpha_p\mu_Y\Big]}_{(*)} + \epsilon_t$$

since we have imposed that (*) = 0, we get:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \epsilon_t$$

which is an AR(p) process without an intercept!

Hence, to summarise what we have done, we took an AR(p) model with an intercept:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \mu + \epsilon_t$$

and rewrote it as an AR(p) model without an intercept whereby Y_t is now the sum of a mean-zero process and its

long-run mean:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \epsilon_t$$
$$Y_t = X_t + \mu_Y$$
$$\mu_Y = \frac{\mu}{1 - \alpha_1 - \alpha_2 - ... - \alpha_p}$$

□

We have just shown how to write an AR(p) model with an intercept Y_t into an AR(p) model without an intercept X_t. Next, we want to be able to write this AR(p) model without an intercept as a vector AR(1) model without an intercept. Writing an AR(p) model in its **companion form** allows us to do this.

Definition 498 (Companion Form of AR(p))
Let X_t be an AR(p) model:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \epsilon_t$$

Then, the **companion form** of the AR(p) model can be written as:

$$\begin{bmatrix} X_t \\ X_{t-1} \\ \cdot \\ \cdot \\ \cdot \\ X_{t-p+2} \\ X_{t-p+1} \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 & ... & \alpha_{p-1} & \alpha_p \\ 1 & 0 & ... & 0 & 0 \\ 0 & 1 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 1 & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_{t-2} \\ \cdot \\ \cdot \\ \cdot \\ X_{t-p+1} \\ X_{t-p} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \epsilon_t \tag{415}$$

or in matrix form:

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \tau\epsilon_t \tag{416}$$

whereby $\tau = (1, 0, ..., 0)^T$ is a $p \times 1$ vector.

We now want to write this matrix in a recursive form just like for the AR(1) model.

Proposition 499 (Recursive form of AR(p) Model)
Let $\boldsymbol{X}_t$ be the companion form of the AR(p) process X_t. Then, the recursive solution to $\boldsymbol{X}_t$ is:

$$\boldsymbol{X}_t = \sum_{j=0}^{t-1} \boldsymbol{A}^j \tau \epsilon_{t-j} + \boldsymbol{A}^t \boldsymbol{X}_0. \tag{417}$$

We have written out the recursive form of the AR(p) model:

$$\boldsymbol{X}_t = \sum_{j=0}^{t-1} \boldsymbol{A}^j \tau \epsilon_{t-j} + \boldsymbol{A}^t \boldsymbol{X}_0$$

. We want to now check if we can write it in terms of an infinite moving average representation:

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \epsilon_{t-j}.$$

To do so, we check the eigenvalues of $\boldsymbol{A}$.

Proposition 500 (Stability Condition for AR(p) Companion Matrix)
Let λ_j for $j = 1, ..., p$ be the **eigenvalues** associated to the AR(p) companion matrix $\boldsymbol{A}$. Then, if:

$$\max_{j=1,...,p} |\lambda_j| < 1 \tag{418}$$

then the AR(p) model admits a stationarity solution.

Remark 501. *Note that the eigenvalues of $\boldsymbol{A}$ solve the equation:*

$$det(\boldsymbol{A} - \lambda \boldsymbol{I}_p) = 0.$$

This is characterised by the characteristic polynomial:

$$\lambda^p - \alpha_1 \lambda^{p-1} - ... - \alpha_p \lambda^0 = 0.$$

Then if all the roots to this characteristic equation lie **inside** *the unit circle, then the stability condition holds.*

Alternatively, we can set $\gamma = \frac{1}{\lambda}$ and write the characteristic polynomial as:

$$\gamma^0 - \alpha_1 \gamma^1 - ... - \alpha_p \gamma^p = 0.$$

Then, if all the roots to this characteristic equation lie **outside** *the unit circle, then the stability condition holds.*

We now show how we could have derived this using lag polynomials. First, write the lag operator:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \epsilon_t$$

to get:

$$X_t = \alpha_1 L X_t + \alpha_2 L^2 X_t + ... + \alpha_p L^p X_t + \epsilon_t$$

and re-arrange:

$$(1 - \alpha_1 L - \alpha_2 L^2 - ... - \alpha_p L^p) X_t = \epsilon_t.$$

Define the autoregressive polynomial:

$$\alpha(L) = (1 - \alpha_1 L - \alpha_2 L^2 - ... - \alpha_p L^p).$$

Now, we have that invertibility holds if the roots to:

$$(1 - \alpha_1 L - \alpha_2 L^2 - ... - \alpha_p L^p)$$

are all greater than 1. This is equivalent to the eigenvalues of the companion matrix $\boldsymbol{A}$ being less than one. This allows us to invert the autoregressive polynomial to get the moving average representation.

We can now state the moving average representation of the AR(p) model.

Proposition 502 (Moving Average Representation of AR(p))

Let $\boldsymbol{X}_t$ be the AR(p) model in companion form with the companion matrix $\boldsymbol{A}$. Suppose that all the eigenvalues of $\boldsymbol{A}$ has modulus strictly less than 1. Then, $\boldsymbol{X}_t$ can be represented in the infinite moving average form:

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \epsilon_{t-j} \tag{419}$$

where $\tau = (1, 0, ..., 0)^T$ is a $p \times 1$ vector.

We can get the distribution for the AR(p) model as well.

Proposition 503 (Distribution of AR(p) Model)

Let $\boldsymbol{X}_t$ be the AR(p) model in companion form. Suppose that $\boldsymbol{X}_t$ admits a stationary solution so that it has the infinite order representation:

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \epsilon_{t-j} \tag{420}$$

where $\tau = (1, 0, ..., 0)^T$ is a $p \times 1$ vector and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. Then, the distribution of X_t is given by:

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \epsilon_{t-j} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_X) \tag{421}$$

whereby $\Sigma_X \equiv Var[\boldsymbol{X}_t] = \sum_{j=0}^{\infty} \boldsymbol{A}^j \tau \sigma^2 \tau^T (\boldsymbol{A}^j)^T$.

12.14 ADL Models

So far, we have looked at models whereby the variables of interested Y_t only depended on lag values of itself such as the AR(p) model:

$$Y_t = \sum_{j=1}^{p} \alpha_j Y_{t-j} + \epsilon_t.$$

We can now extend this case by now incorporating exogenous variables Z_t in the system. We now define such a model.

Definition 504 (Autoregressive Distributed Lag Model)

An **Autoregressive Distributed Lag** model (ADL(p,r)) model is given by:

$$Y_t = \sum_{j=1}^{p} \alpha_j Y_{t-j} + \sum_{j=0}^{r} \beta_j Z_{t-j} + \mu + \epsilon_t \tag{422}$$

where Z_t is an exogenous variable and μ is a deterministic term.

Remark 505. *Notice that this is similar to an AR(p) model except we now have an exogenous variable Z_t as well. We get back an AR(p) model from an ADL(p,r) model when we set $\beta_j = 0$ for all $j = 0, ..., r$.*

We can estimate the parameters of the ADL(p,r) model via OLS. To interpret these ADL models, it is informative to write them in **equilibrium correct form**.

We now motivate equilibrium correction forms. First, consider the ADL(1,1) model:

$$Y_t = \omega Z_t + \alpha Y_{t-1} + \beta Z_{t-1} + \mu + \epsilon_t \tag{423}$$

whereby the series $\{Y_t\}_{t\in\mathbb{N}}$ and $\{Z_t\}_{t\in\mathbb{N}}$ are **stationary** with long-run means $\mathbb{E}[Y_t] = \mu_Y$ and $\mathbb{E}[Z_t] = \mu_Z$ and $\mathbb{E}[\epsilon_t] = 0$. If we shut down the noise process ϵ_t, then there is no noise in the system and in the **long-run** the system will converge to its long run equilibrium, which we denote by:

$$Y = \omega Z + \alpha Y + \beta Z + \mu$$

which we can then rewrite whereby we use the fact that Y_t and Z_t will converge to their long run mean due to stationarity of the series:

$$\mu_Y = (\omega + \beta)\mu_Z + \alpha\mu_Y + \mu. \tag{424}$$

Eq. (424) is the long equilibrium which our system Eq. (423) will converge to. Now, let us go back to Eq. (423):

$$Y_t = \omega Z_t + \alpha Y_{t-1} + \beta Z_{t-1} + \mu + \epsilon_t.$$

We now want to write this in equilibrium correction form. First, let us center Y_t and Z_t around their long-run means μ_Y and μ_Z by adding and subtracting μ_Y and $\omega\mu_Z$ from both sides:

$$\begin{aligned} Y_t - \mu_Y &= \omega Z_t + \alpha Y_{t-1} + \beta Z_{t-1} + \mu + \epsilon_t - \mu_Y + \omega\mu_Z - \omega\mu_Z \\ \leftrightarrow (Y_t - \mu_Y) &= \omega(Z_t - \mu_Z) + \alpha Y_{t-1} + \beta Z_{t-1} + \mu + \epsilon_t - \mu_Y + \omega\mu_Z \end{aligned}$$

Now add and subtract $\alpha\mu_Y$ to the right hand side:

$$\begin{aligned} \leftrightarrow (Y_t - \mu_Y) &= \omega(Z_t - \mu_Z) + \alpha Y_{t-1} + \beta Z_{t-1} + \mu + \epsilon_t - \mu_Y + \omega\mu_Z + \alpha\mu_Y - \alpha\mu_Y \\ \leftrightarrow (Y_t - \mu_Y) &= \omega(Z_t - \mu_Z) + \alpha(Y_{t-1} - \mu_Y) + \beta Z_{t-1} + \mu + \epsilon_t - \mu_Y + \omega\mu_Z + \alpha\mu_Y \end{aligned}$$

Finally, add and subtract $\beta\mu_Z$ to the right hand side:

$$\begin{aligned} \leftrightarrow (Y_t - \mu_Y) &= \omega(Z_t - \mu_Z) + \alpha(Y_{t-1} - \mu_Y) + \beta Z_{t-1} + \mu + \epsilon_t - \mu_Y + \omega\mu_Z + \alpha\mu_Y + \beta\mu_Z - \beta\mu_Z \\ \leftrightarrow (Y_t - \mu_Y) &= \omega(Z_t - \mu_Z) + \alpha(Y_{t-1} - \mu_Y) + \beta(Z_{t-1} - \mu_Z) + \mu + \epsilon_t - \mu_Y + \omega\mu_Z + \alpha\mu_Y + \beta\mu_Z \end{aligned}$$

We now collect all the terms together and see we have:

$$(Y_t - \mu_Y) = \omega(Z_t - \mu_Z) + \alpha(Y_{t-1} - \mu_Y) + \beta(Z_{t-1} - \mu_Z) + \Big[\mu - (1-\alpha)\mu_Y + (\omega+\beta)\mu_Z\Big] + \epsilon_t$$

Now, if you look carefully, we have that nearly all the terms will have mean zero:

$$\underbrace{(Y_t - \mu_Y)}_{\text{Mean 0}} = \omega\underbrace{(Z_t - \mu_Z)}_{\text{Mean 0}} + \alpha\underbrace{(Y_{t-1} - \mu_Y)}_{\text{Mean 0}} + \beta\underbrace{(Z_{t-1} - \mu_Z)}_{\text{Mean 0}} + \underbrace{\Big[\mu - (1-\alpha)\mu_Y + (\omega+\beta)\mu_Z\Big]}_{(*)} + \underbrace{\epsilon_t}_{\text{Mean 0}}.$$

We now want to argue that the term (*) will also have mean zero. Notice that (*) is exactly the long-run equilibrium term Eq. (424) that we derived earlier but re-arranged:

$$\mu_Y = (\omega+\beta)\mu_Z + \alpha\mu_Y + \mu \quad \leftrightarrow \quad 0 = (\omega+\beta)\mu_Z - (1-\alpha)\mu_Y + \mu.$$

Hence, we can conclude that for the expression:

$$(Y_t - \mu_Y) = \omega(Z_t - \mu_Z) + \alpha(Y_{t-1} - \mu_Y) + \beta(Z_{t-1} - \mu_Z) + \underbrace{\Big[\mu - (1-\alpha)\mu_Y + (\omega+\beta)\mu_Z\Big]}_{(*)} + \epsilon_t$$

that for the left hand side to equal zero, the term (*) must also equal zero as all other terms on the right hand side equal zero!

Now, go back to long-run relationship term (*):

$$\mu - (1-\alpha)\mu_Y + (\omega+\beta)\mu_Z = 0$$

we can then solve for μ:

$$\mu = (1-\alpha)\mu_Y - (\omega+\beta)\mu_Z \tag{425}$$

We can interpret Eq. (425) to be the long-run equilibrium relationship between the means μ_Y and μ_Z. That is, in the long-run, the means of Y_t and Z_t must satisfy this equilibrium relationship. We again rewrite Eq. (425) as:

$$0 = (1-\alpha)\left(\mu_Y - \frac{\omega+\beta}{1-\alpha}\mu_Z - \frac{\mu}{1-\alpha}\right) \tag{426}$$

which is another way of saying what the long-run relationship between μ_Y and μ_Z is. We will soon see why Eq. (426) is useful.

We can now finally write the equilibrium correction form of:

$$Y_t = \omega Z_t + \alpha Y_{t-1} + \beta Z_{t-1} + \mu + \epsilon_t.$$

First, subtract Y_{t-1} from both sides and add and subtract ωZ_{t-1} on the right hand side to get:

$$\begin{aligned}
&Y_t - Y_{t-1} = \omega Z_t + \alpha Y_{t-1} + \beta Z_{t-1} + \mu + \epsilon_t - Y_{t-1} + \omega Z_{t-1} - \omega Z_{t-1} \\
&\leftrightarrow Y_t - Y_{t-1} = \omega(Z_t - Z_{t-1}) + (\alpha - 1)Y_{t-1} + (\beta+\omega)Z_{t-1} + \mu + \epsilon_t \\
&\leftrightarrow \Delta Y_t = \omega \Delta Z_t - (1-\alpha)Y_{t-1} + (\beta+\omega)Z_{t-1} + \mu + \epsilon_t
\end{aligned}$$

where $\Delta = (1-L)$ is the **difference operator** such that $\Delta Y_t = Y_t - Y_{t-1}$ and $\Delta Z_t = Z_t - Z_{t-1}$. Now, rewrite it as:

$$\begin{aligned}
&\leftrightarrow \Delta Y_t = \omega \Delta Z_t - (1-\alpha)Y_{t-1} + (\beta+\omega)Z_{t-1} + \mu + \epsilon_t \\
&\leftrightarrow \Delta Y_t = \omega \Delta Z_t - (1-\alpha)\left[Y_{t-1} - \frac{(\beta+\omega)}{1-\alpha}Z_{t-1} - \frac{\mu}{1-\alpha}\right] + \epsilon_t
\end{aligned}$$

We now have the new expression for our ADL(1,1) model as:

$$\Delta Y_t = \omega \Delta Z_t - (1-\alpha)\underbrace{\left[Y_{t-1} - \frac{(\beta+\omega)}{1-\alpha}Z_{t-1} - \frac{\mu}{1-\alpha}\right]}_{(*)} + \epsilon_t.$$

But notice the term (*), it is actually Eq. (426) in the short-run! We compare the two expressions to make this

more clear:

$$\left[Y_{t-1} - \frac{(\beta+\omega)}{1-\alpha}Z_{t-1} - \frac{\mu}{1-\alpha}\right]$$
$$\left[\mu_Y - \frac{\omega+\beta}{1-\alpha}\mu_Z - \frac{\mu}{1-\alpha}\right]$$

Therefore, we can conclude that in the long-run, (*) will be equal to 0:

$$(*) = (1-\alpha)\left[Y_{t-1} - \frac{(\beta+\omega)}{1-\alpha}Z_{t-1} - \frac{\mu}{1-\alpha}\right] = 0.$$

This expression is called the **equilibrium correction mechanism**.

Definition 506 (Equilibrium Correction Mechanism)

The **equilibrium correction mechanism** for the ADL(1,1) model is:

$$ecm_{t-1} = \left[Y_{t-1} - \frac{(\beta+\omega)}{1-\alpha}Z_{t-1} - \frac{\mu}{1-\alpha}\right]. \tag{427}$$

We therefore replace this expression into our ADL(1,1) model:

$$\Delta Y_t = \omega\Delta Z_t - (1-\alpha)\left[Y_{t-1} - \frac{(\beta+\omega)}{1-\alpha}Z_{t-1} - \frac{\mu}{1-\alpha}\right] + \epsilon_t$$
$$\Delta Y_t = \omega\Delta Z_t - (1-\alpha)ecm_{t-1} + \epsilon_t$$

If we look now at our ADL(1,1), we can see that the mean of each term is zero:

$$\underbrace{\Delta Y_t}_{\text{Mean zero}} = \omega \underbrace{\Delta Z_t}_{\text{Mean zero}} - (1-\alpha)\underbrace{ecm_{t-1}}_{\text{Mean zero}} + \underbrace{\epsilon_t}_{\text{Mean zero}}$$

We can now officially state the equilibrium correction form.

Definition 507 (Equilibrium Correction Form)

Let the ADL(1,1) model be:

$$Y_t = \omega Z_t + \alpha Y_{t-1} + \beta Z_{t-1} + \mu + \epsilon_t.$$

Then, the equilibrium correction form for the ADL(1,1) model is given by:

$$\Delta Y_t = \omega\Delta Z_t - (1-\alpha)ecm_{t-1} + \epsilon_t \tag{428}$$
$$ecm_{t-1} = \left[Y_{t-1} - \frac{(\beta+\omega)}{1-\alpha}Z_{t-1} - \frac{\mu}{1-\alpha}\right] \tag{429}$$

where $|\alpha| < 1$.

We now show the nice interpretation that comes with equilibrium correction form. The general intuition is that the model will adjust the process so that it tends to the long-run equilibrium. We look at two cases to illustrate this.

First, suppose that Y_{t-1} is large **relative** to Z_{t-1}. As a result, we see that the equilibrium correction mechanism term Eq. (429) will be positive: $ecm_{t-1} > 0$. This then feeds into the equilibrium correction form equation Eq. (428) via $-(1-\alpha)ecm_{t-1}$ so that ΔY_t will be adjusted downwards. This means that Y_t is smaller as $\Delta Y_t = Y_t - Y_{t-1}$. Hence,

we can see that if Y_{t-1} is too big, then the equilibrium correction mechanism ecm_{t-1} will push next period Y_t down in order to bring it back in line with the long-run equilibrium.

Second, we do a similar exercise whereby we suppose Y_{t-1} is small **relative** to Z_{t-1}. Then, the equilibrium correction mechanism term Eq. (429) will be negative: $ecm_{t-1} < 0$. This then feeds into the equilibrium correction form equation Eq. (428) via $-(1-\alpha)ecm_{t-1}$ so that ΔY_t will be adjusted upwards. Therefore, if Y_{t-1} is too small, then the equilibrium correction mechanism ecm_{t-1} will push Y_t up in order to bring it back in line with the long-run equilibrium.

12.15 VAR Models

Recall that we have previously seen the AR(p) model:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t. \tag{430}$$

We now wish to generalise this through the **vector autoregressive model** (VAR model).

Definition 508 (Vector Autoregressive Model)
A VAR(p) of dimension k can be written as:

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + \sum_{j=1}^{p} \Phi_j \boldsymbol{Y}_{t-j} + \boldsymbol{\epsilon}_t \tag{431}$$

where:

1. $\boldsymbol{Y}_t = (Y_{t,1}, ..., Y_{t,k})^T$
2. $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)^T$
3. $\boldsymbol{\epsilon}_t = (\epsilon_{t,1}, ..., \epsilon_{t,k})^T$
4. Φ_j is a $k \times k$ matrix of coefficients.

Furthermore, $\boldsymbol{\epsilon}_t \sim \mathcal{N}_k(\boldsymbol{0}, \Omega_k)$.

Remark 509. *Notice that we have k variables and p lags.*

Proposition 510 (Number of Conditional-Mean Parameters in VAR(p) Model)
The number of conditional-mean parameters in the VAR(p) model is $k + pk^2$.

Proof. We have k parameters in the intercept term $\boldsymbol{\mu}$. Furthermore, we have k^2 terms in the $k \times k$ matrix Φ_j. Then, we have p of these matrices so we need $p \times k^2$. □

Remark 511. *If the innovation covariance matrix is also counted, there are an additional $k(k+1)/2$ distinct parameters.*

We have some interesting cases of the VAR(p) model.

Proposition 512 (Special Cases of VAR)

The special cases of the VAR(p) model are:

1. VAR(1) is when $p = 1$
2. AR(1) is when $p = 1$ and $k = 1$
3. AR(p) is when $k = 1$.

We now want to be able to write the VAR(p) model in a VAR(1) form by writing it in its companion form.

Proposition 513 (Companion Form of VAR(p) Model)

A VAR(p) model can be written in its **companion form** as a VAR(1) by writing it as:

$$\boldsymbol{X}_t = \boldsymbol{c} + \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{u}_t \tag{432}$$

whereby the **companion matrix** $\boldsymbol{A}$ is a $(kp \times kp)$ matrix and is given by:

$$\boldsymbol{A} = \begin{bmatrix} \Phi_1 & \Phi_2 & \dots & \Phi_{p-1} & \Phi_p \\ \boldsymbol{I}_k & \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_k & \dots & \boldsymbol{0} & \boldsymbol{0} \\ . & . & \dots & . & . \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{I}_k & \boldsymbol{0} \end{bmatrix} \tag{433}$$

and the vectors are:

$$\boldsymbol{X}_t = \left(\boldsymbol{Y}_t^T, ..., \boldsymbol{Y}_{t-p+1}^T\right)^T \qquad (kp \times 1) \tag{434}$$

$$\boldsymbol{u}_t = \left(\boldsymbol{\epsilon}_t^T, \boldsymbol{0}, ..., \boldsymbol{0}\right)^T \qquad (kp \times 1) \tag{435}$$

$$\boldsymbol{c} = \left(\boldsymbol{\mu}^T, \boldsymbol{0}, ..., \boldsymbol{0}\right)^T \qquad (kp \times 1) \tag{436}$$

$$\tag{437}$$

Remark 514. *Notice that in $\boldsymbol{X}_t$, each block $\boldsymbol{Y}_t^T$ is a $1 \times k$ row vector before the blocks are stacked and transposed.*

We can state a similar condition for stationarity for the VAR(p) model as we did for the AR(p) model.

Proposition 515 (Stationarity for VAR(p) Model)

The VAR(p) possesses a stationary solution if all roots of the lag polynomial are **outside** the unit circle:

$$|\boldsymbol{I}_k - \sum_{j=1}^{p} \Phi_j z^j| = 0 \tag{438}$$

which only occurs for $|z| > 1$.

Equivalently, the VAR(p) possesses a stationary solution if all the eigenvalues of the companion matrix $\boldsymbol{A}$ lie **inside** the unit circle.

We can estimate a VAR(p) model's parameters by running ordinary least squares equation by equation on:

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + \sum_{j=1}^{p} \Phi_j \boldsymbol{Y}_{t-j} + \boldsymbol{\epsilon}_t.$$

Alternatively, if we assume normality of the error terms, this is equivalent to maximum likelihood conditional on the initial values.

We can rewrite a VAR(1) with an intercept as a VAR(1) model without an intercept just like what we did for a AR(p) model. Note that notation wise, we have switched back to $\boldsymbol{Y}_t$ so don't think that this $\boldsymbol{Y}_t$ is the VAR(p) model without writing it in companion form, we just want to follow a similar notation to what we did for the AR(2) and AR(p) model.

Proposition 516 (VAR(1) Model Without an Intercept)
The k-dimensional VAR(1) model with an intercept:

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{\epsilon}_t \tag{439}$$

can be written as a VAR(1) model without an intercept:

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t \tag{440}$$
$$\boldsymbol{Y}_t = \boldsymbol{X}_t + \boldsymbol{\mu}_Y \tag{441}$$
$$\boldsymbol{\mu}_Y = (\boldsymbol{I}_k - \boldsymbol{A})^{-1}\boldsymbol{\mu} \tag{442}$$

where $\boldsymbol{\mu}_Y$ is the long-run mean.

Proof. The proof is the same as the AR(p) model. First, subtract the vector $\boldsymbol{\mu}_Y$ from both sides:

$$\boldsymbol{Y}_t - \boldsymbol{\mu}_Y = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{\epsilon}_t - \boldsymbol{\mu}_Y.$$

Now, add and subtract $\boldsymbol{A}\boldsymbol{\mu}_Y$ on the right-hand side:

$$\boldsymbol{Y}_t - \boldsymbol{\mu}_Y = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{\epsilon}_t - \boldsymbol{\mu}_Y + \boldsymbol{A}\boldsymbol{\mu}_Y - \boldsymbol{A}\boldsymbol{\mu}_Y$$

and group terms to get:

$$\boldsymbol{Y}_t - \boldsymbol{\mu}_Y = \boldsymbol{A}(\boldsymbol{Y}_{t-1} - \boldsymbol{\mu}_Y) + \left[\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{\mu}_Y - \boldsymbol{\mu}_Y\right] + \boldsymbol{\epsilon}_t$$
$$\leftrightarrow \boldsymbol{Y}_t - \boldsymbol{\mu}_Y = \boldsymbol{A}(\boldsymbol{Y}_{t-1} - \boldsymbol{\mu}_Y) + \left[\boldsymbol{\mu} - [\boldsymbol{I}_k - \boldsymbol{A}]\boldsymbol{\mu}_Y\right] + \boldsymbol{\epsilon}_t$$

So we have:

$$\boldsymbol{Y}_t - \boldsymbol{\mu}_Y = \boldsymbol{A}(\boldsymbol{Y}_{t-1} - \boldsymbol{\mu}_Y) + \left[\boldsymbol{\mu} - [\boldsymbol{I}_k - \boldsymbol{A}]\boldsymbol{\mu}_Y\right] + \boldsymbol{\epsilon}_t$$

We then define our mean-zero process:

$$\boldsymbol{X}_t \equiv \boldsymbol{Y}_t - \boldsymbol{\mu}_Y$$

so that we have:

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \underbrace{\left[\boldsymbol{\mu} - [\boldsymbol{I}_k - \boldsymbol{A}]\boldsymbol{\mu}_Y\right]}_{(*)} + \boldsymbol{\epsilon}_t$$

Finally, we set (*) equal to zero:

$$\Big[\boldsymbol{\mu} - [\boldsymbol{I}_k - \boldsymbol{A}]\boldsymbol{\mu}_Y\Big] = 0 \quad \leftrightarrow \quad \boldsymbol{\mu}_Y = [\boldsymbol{I}_k - \boldsymbol{A}]^{-1}\boldsymbol{\mu}.$$

Hence, we end up with:

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t$$

which is a VAR(1) model without an intercept. □

We can then get the limiting distribution of the VAR(1) model.

Proposition 517 (Limiting Distribution of VAR(1) Model Without Intercept)
Let $\boldsymbol{X}_t$ be the k-dimensional VAR(1) model without an intercept:

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t \tag{443}$$
$$\boldsymbol{Y}_t = \boldsymbol{X}_t + \boldsymbol{\mu}_Y \tag{444}$$
$$\boldsymbol{\mu}_Y = (\boldsymbol{I}_k - \boldsymbol{A})^{-1}\boldsymbol{\mu} \tag{445}$$
$$\boldsymbol{\epsilon}_t \sim \mathcal{N}_k(\boldsymbol{0}, \Omega_k) \tag{446}$$

where $\boldsymbol{\mu}_Y$ is the long-run mean. Then, if $\boldsymbol{X}_t$ admits a stationary solution, then $\boldsymbol{X}_t$ has the limiting distribution:

$$\boldsymbol{X}_t | \boldsymbol{X}_0 \xrightarrow{D} \boldsymbol{X}_t^* = \sum_{j=0}^{\infty} \boldsymbol{A}^j \boldsymbol{\epsilon}_{t-j} \sim \mathcal{N}_k(\boldsymbol{0}, \Sigma_{\boldsymbol{X}}) \tag{447}$$

whereby the moments are:

$$\mathbb{E}[\boldsymbol{X}_t^*] = \boldsymbol{0}$$
$$\Sigma_{\boldsymbol{X}} = \sum_{j=0}^{\infty} \boldsymbol{A}^j \Omega_k (\boldsymbol{A}^j)^T.$$

Correspondingly, the limiting distribution of $\boldsymbol{Y}_t$ is $\mathcal{N}_k(\boldsymbol{\mu}_Y, \Sigma_{\boldsymbol{X}})$.

Proposition 518 (Stationarity from Stability)
Suppose that y_t is a stable AR(p) process. Furthermore, suppose that the mean of the AR process does not change over time, the error term u_t has time-invariant variance σ_u^2, and it's first and second moments are bounded. Then, the AR process y_t is stationary.

Remark 519. *Do not mix up stability and stationarity!*

Definition 520 (I(d) Process)

A stationary process is a I(0) process. For $d \in \mathbb{N}$, a stochastic process y_t is called $I(d)$ if $\Delta^d y_t \equiv z_t$ is a stationary process with infinite-order moving average representation:

$$z_t = \sum_{j=0}^{\infty} \theta_j u_{t-j} = \theta(L) u_t \tag{448}$$

where the MA coefficients satisfy the conditions:

1. $\sum_{j=0}^{\infty} j|\theta_j| < \infty$
2. $\theta(1) = \sum_{j=0}^{\infty} \theta_j \neq 0$
3. u_t is a white noise process.

Definition 521 (Wold Representation Theorem)

Every K-dimensional non-deterministic zero mean stationary process $\boldsymbol{y}_t$ has an moving average representation:

$$\boldsymbol{y}_t = \sum_{i=0}^{\infty} \boldsymbol{\Phi}_i \boldsymbol{u}_{t-i} \tag{449}$$

where $\boldsymbol{\Phi}_0 = \boldsymbol{I}_K$ and the coefficient matrices are square summable.

Let the K time series variable $\boldsymbol{y}_t = (y_{1t}, ..., y_{Kt})^T$ be decomposed as:

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{x}_t \tag{450}$$

where $\boldsymbol{\mu}_t$ is a deterministic part and $\boldsymbol{x}_t$ is a stochastic process with mean zero.

$$\mathbb{E}[\boldsymbol{y}_t] = \boldsymbol{\mu}_t.$$

Definition 522 (VAR(p) Model)

The stochastic process $\boldsymbol{x}_t$ is a linear VAR process of order p if it is of the form:

$$\boldsymbol{x}_t = \boldsymbol{A}_1 \boldsymbol{x}_{t-1} + ... + \boldsymbol{A}_p \boldsymbol{x}_{t-p} + \boldsymbol{u}_t \tag{451}$$

where $\boldsymbol{x}_t = (x_{1t}, ..., x_{Kt})^T$ is a $(K \times 1)$ random vector, $\boldsymbol{A}_i$ are $K \times K$ coefficient matrices, and $\boldsymbol{u}_t = (u_{1t}, ..., u_{Kt})^T$ is a K-dimensional zero-mean white noise process with covariance matrix $\mathbb{E}[\boldsymbol{u}_t \boldsymbol{u}_t^T] = \Sigma_u$ such that $\boldsymbol{u}_t \sim (\boldsymbol{0}, \Sigma_u)$.

Remark 523. *We can relax the assumption of i.i.d errors $\boldsymbol{u}_t$ by requiring them to form a martingale difference sequence instead.*

We can define the matrix polynomial in the lag operator:

$$\boldsymbol{A}(L) = I_k - \boldsymbol{A}_1 L - ... - \boldsymbol{A}_p L^p$$

This allows us to rewrite the VAR(p) model:

$$\begin{aligned} & \boldsymbol{x}_t = \boldsymbol{A}_1 \boldsymbol{x}_{t-1} + ... + \boldsymbol{A}_p \boldsymbol{x}_{t-p} + \boldsymbol{u}_t \\ \leftrightarrow & \ \boldsymbol{x}_t - \boldsymbol{A}_1 \boldsymbol{x}_{t-1} - ... - \boldsymbol{A}_p \boldsymbol{x}_{t-p} = \boldsymbol{u}_t \\ \leftrightarrow & \left[I_k - \boldsymbol{A}_1 L - ... - \boldsymbol{A}_p L^p \right] \boldsymbol{x}_t = \boldsymbol{u}_t \end{aligned}$$

Therefore, we have the expression:

$$\boldsymbol{A}(L)\boldsymbol{x}_t = \boldsymbol{u}_t \tag{452}$$

whereby $\boldsymbol{A}(L) = I_k - \boldsymbol{A}_1 L - ... - \boldsymbol{A}_p L^p$.

Recall that we were looking at the decomposition:

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{x}_t$$

Therefore, $\boldsymbol{y}_t$ inherits the stochastic properties of $\boldsymbol{x}_t$ after accounting for the deterministic component. In particular, stationarity of $\boldsymbol{y}_t$ requires $\boldsymbol{\mu}_t$ to be constant. To see the algebra more formally, let us multiply this expression by $\boldsymbol{A}(L)$:

$$\begin{aligned} & \boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{x}_t \\ \leftrightarrow & \ \boldsymbol{A}(L)\boldsymbol{y}_t = \boldsymbol{A}(L)\boldsymbol{\mu}_t + \boldsymbol{A}(L)\boldsymbol{x}_t \\ \leftrightarrow & \ \boldsymbol{A}(L)\boldsymbol{y}_t = \boldsymbol{A}(L)\boldsymbol{\mu}_t + \boldsymbol{u}_t \\ \leftrightarrow & \left[I_k - \boldsymbol{A}_1 L - ... - \boldsymbol{A}_p L^p \right] \boldsymbol{y}_t = \boldsymbol{A}(L)\boldsymbol{\mu}_t + \boldsymbol{u}_t \\ \leftrightarrow & \ \boldsymbol{y}_t = \boldsymbol{A}(L)\boldsymbol{\mu}_t + \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + ... + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t \end{aligned}$$

This is called a **reduced form** as all the variables on the right hand side are lagged.

Definition 524 (Multivariate Stability Condition)

Let $\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{x}_t$. Define the lag-operator associated to $\boldsymbol{x}_t$ as $\boldsymbol{A}(L) = I_k - \boldsymbol{A}_1 L - ... - \boldsymbol{A}_p L^p$. Then, the process $\boldsymbol{y}_t$ is stable if the stochastic process $\boldsymbol{x}_t$ is stable, which holds when all the roots of the characteristic polynomial are outside the complex unit circle:

$$det(\boldsymbol{A}(z)) = det(\boldsymbol{I}_k - \boldsymbol{A}_1 z - ... - \boldsymbol{A}_p z^p) \neq 0 \tag{453}$$

for all $|z| \leq 1$.

Proposition 525 (Stationarity of Multivariate Process)

Let $\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{x}_t$, where $\boldsymbol{\mu}$ is constant. If $\boldsymbol{x}_t$ is stable and has white-noise innovations with finite variance, then $\boldsymbol{y}_t$ is covariance stationary.

Definition 526 (Jordan Canonical Form)

Let $\boldsymbol{A}$ be a $m \times m$ matrix with eigenvalues $\lambda_1, ..., \lambda_n$. Then, there exists a non-singular matrix $\boldsymbol{P}$ such that:

$$\boldsymbol{A} = \boldsymbol{P} \begin{bmatrix} \boldsymbol{\Lambda}_1 & ... & 0 \\ ... & ... & ... \\ 0 & ... & \boldsymbol{\Lambda}_n \end{bmatrix} \boldsymbol{P}^{-1} \equiv \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1} \tag{454}$$

where:

$$\boldsymbol{\Lambda}_i = \begin{bmatrix} \lambda_i & 1 & 0 & ... & 0 \\ 0 & \lambda_i & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \\ 0 & 0 & 0 & ... & \lambda_i \end{bmatrix} \tag{455}$$

For any matrix $\boldsymbol{A}$, we can decompose it into its Jordan canonical form. The following result is why we care about it.

Proposition 527 (Convergence of a Matrix with Eigenvalues Inside the Unit Circle)

Suppose that $\boldsymbol{A}$ is a $m \times m$ matrix with eigenvalues $\lambda_1, ..., \lambda_n$ which all have modulus less than 1, $|\lambda_i| < 1$ for $i = 1, ..., n$. Then

$$\boldsymbol{A}^j \to 0 \tag{456}$$

as $j \to \infty$.

Proof. First, we can write $\boldsymbol{A}$ into its Jordan canonical form:

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}.$$

This then implies that:

$$\boldsymbol{A}^j = (\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1})^j = \boldsymbol{P}\boldsymbol{\Lambda}^j\boldsymbol{P}^{-1}.$$

Write each Jordan block as $\boldsymbol{\Lambda}_i = \lambda_i \boldsymbol{I} + \boldsymbol{N}_i$, where $\boldsymbol{N}_i$ is nilpotent of order r_i. Then

$$\boldsymbol{\Lambda}_i^j = \sum_{q=0}^{r_i - 1} \binom{j}{q} \lambda_i^{j-q} \boldsymbol{N}_i^q.$$

For every fixed q, the polynomial term $\binom{j}{q}$ is dominated by the geometric decay of $|\lambda_i|^{j-q}$. Thus every Jordan block converges to zero, so $\boldsymbol{\Lambda}^j \to \boldsymbol{0}$ and hence $\boldsymbol{A}^j \to \boldsymbol{0}$.

□

12.15.1 Granger Causality

A variable y_{2t} is called Granger causal for y_{1t} if the information in the past and present values of y_{2t} helps reduce the expectation of the squared prediction error for y_{1t}.

Definition 528 (Granger Causality)

Let Ω_t be the information set containing all information relevant to predicting y_1. Denote the optimal h-step prediction of y_{1t} at origin t based on the information set in Ω_t by $y_{1,t+h|\Omega}$ and the corresponding mean-squared prediction error by $\sigma^2_{y_1}(h|\Omega_t)$. Then the process y_{2t} is said to **Granger cause** y_{1t} if:

$$\sigma^2_{y_1}(h|\Omega_t) < \sigma^2_{y_1}(h|\Omega_t \setminus \{y_{2s}|s \leq t\}) \tag{457}$$

for at least one $h \in \{1, 2, ...\}$ where $\Omega_t \setminus \{y_{2s}|s \leq t\}$ denotes the set of all relevant information in the universe apart from the past and present information about the y_{2t} process.

y_{2t} Granger causes y_{1t} if y_{1t} can be predicted with a lower mean squared error by taking into the account of the information in y_{2s} for $s \leq t$.

We describe how to test Granger causality in a VAR(p) process.

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} + \sum_{i=1}^{p} \begin{bmatrix} a_{11,i} & a_{12,i} \\ a_{21,i} & a_{22,i} \end{bmatrix} \begin{bmatrix} y_{1,t-i} \\ y_{2,t-i} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}. \tag{458}$$

Then, y_{2t} does not Granger cause y_{1t} if and only if $a_{12,i} = 0$ for $i = 1, 2, ..., p$. We can test this using Wald tests.

There are a few issues with Granger causality. First, y_{2t} may Granger cause y_{1t} if y_{3t} is present but not Granger cause if y_{3t} is not present. Another issue of Granger causality is that it does not look at the instantaneous relationships of the variables.

We can also decide how many lags to include in our VAR model.

Definition 529 (Information Criterion)

The information criterion has the form:

$$C(m) = \log(\det(\tilde{\Sigma}_u(m))) + c_T \phi(m) \tag{459}$$

where m is the candidate lag order, $\phi(m)$ is a function of order m that penalizes large larger orders, c_T is a sequence of weights, and $\tilde{\Sigma}_u(m) = T^{-1} \sum_{t=1}^{T} \hat{\boldsymbol{u}}_t \hat{\boldsymbol{u}}_t^T$ is the residual covariance matrix estimator for a reduced-form VAR model of order m.

We are choosing the number of lags m that the VAR(m) model should have. Since we have mK lagged regressor in each equation and K (variables) equations in the VAR model, the total number of regressors (and hence parameters) is: $\phi(m) = mK^2$ If we include an intercept, this then becomes: $\phi(m) = mK^2 + K$.

AIC is not a consistent estimator of the lag length, whereas BIC and HQC (Hannan–Quinn) are consistent.

12.16 Misspecification Testing in Time Series

12.17 Unit Root Testing

So far, we have looked at models where the coefficient $|\alpha| < 1$. We now are interested in the case to see what happens when $\alpha = 1$. We shall soon see that under such a situation, we require new asymptotic results.

Definition 530 (Unit Root)
We say that the process:

$$Y_t = \alpha Y_{t-1} + \epsilon_t \tag{460}$$

possesses a **unit root** if $\alpha = 1$.

Recall that previously, through back-substitution, we could write:

$$Y_t = \alpha Y_{t-1} + \epsilon_t$$

as the sum of error terms:

$$Y_t = \sum_{j=0}^{t-1} \alpha^j \epsilon_{t-j} + \alpha^t Y_0$$

where Y_0 is the initial value of the series. Now, if we have that $\alpha = 1$, this then becomes:

$$Y_t = \sum_{j=0}^{t-1} \epsilon_{t-j} + Y_0. \tag{461}$$

The reason why this is significant is stated in the next proposition.

Proposition 531 (Unit Roots Processes are Not Weakly Stationary)
An AR(1) series that possesses a unit root is **not weakly stationary**.

Proof. Write the AR(1) model as:

$$Y_t = \sum_{j=0}^{t-1} \epsilon_{t-j} + Y_0.$$

Now compute the variance:

$$Var[Y_t] = Var[\sum_{j=0}^{t-1} \epsilon_{t-j}] + Var[Y_0] = \sigma^2 t + Var[Y_0]$$

We can see that the variance now depends on the time index t. Therefore, it is not weakly stationary. □

This result is in fact more general whereby any process with a unit root is **not weakly stationary**. An issue is that when we construct OLS estimates, we usually require the time series process to be stationary, but now with a nonstationary process, this may produce invalid OLS estimates. We can also show that we have issues with doing inference and hypothesis testing.

Suppose we had the AR(1) model:

$$Y_t = \alpha Y_{t-1} + \epsilon_t$$

where ϵ_t is i.i.d with $\mathbb{E}[\epsilon_t] = 0$ and $Var[\epsilon_t] = \sigma^2$. We then recall that by the law of large numbers for AR(1):

$$T^{-1} \sum_{t=1}^{T} Y_{t-1}^2 \xrightarrow{P} \frac{\sigma^2}{1 - \alpha^2}$$

and by the central limit theorem for AR(1):

$$T^{-1/2}\sum_{t=1}^{T} Y_{t-1}\epsilon_t \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^4}{1-\alpha^2}\right).$$

Now, if we were interested in testing $\hat{\alpha}$, we can construct the test statistic:

$$\sqrt{T}\left(\hat{\alpha}-\alpha\right) = \frac{T^{-1/2}\sum_{t=1}^{T} Y_{t-1}\epsilon_t}{T^{-1}\sum_{t=1}^{T} Y_{t-1}^2} \xrightarrow{D} \frac{1}{\frac{\sigma^2}{1-\alpha^2}}\mathcal{N}\left(0, \frac{\sigma^4}{1-\alpha^2}\right) = \mathcal{N}(0, 1-\alpha^2).$$

This derivation assumes $|\alpha| < 1$, so we cannot obtain the unit-root limit by substituting $\alpha = 1$ into the stationary formula. At the boundary, the numerator and denominator have different stochastic orders: $\hat{\alpha} - 1$ converges at rate T, rather than $\sqrt{T}$, and has a non-Gaussian limiting distribution. We derive that distribution below.

We now will need to introduce some new terms in order to show what our new limiting distribution will be in the presence of unit roots.

Definition 532 (Random Walk)

A random walk is defined as:

$$Y_t = \sum_{s=1}^{t} \epsilon_s. \tag{462}$$

where ϵ_t is i.i.d with $\mathbb{E}[\epsilon_t] = 0$ and $\mathbb{E}[\epsilon_t^2] = \sigma^2$.

Remark 533. *Another name for* $\sum_{s=1}^{t} \epsilon_s$ *is a* **stochastic trend**.

To see why this is called a random walk, notice that if we difference this:

$$Y_t - Y_{t-1} = \sum_{s=1}^{t} \epsilon_s - \sum_{s=1}^{t-1} \epsilon_s = \epsilon_t$$

which is an independent random variable of previous terms. Hence, where the name random walk comes from.

We also introduce another tool that will be quite useful for us.

Definition 534 (Brownian Motion)

A Brownian motion $(B_u)_{u\in[0,1]}$ is characterized by:

1. $B_0 = 0$
2. it has independent increments
3. $B_t - B_s \sim \mathcal{N}(0, t-s)$ for $0 \leq s < t \leq 1$
4. $u \to B_u$ is continuous (continuous sample paths).

With this, we can state some new asymptotic results that will be quite useful.

Proposition 535 (AR(1) with Unit Root Asymptotic Results)
Suppose that $Y_t = Y_{t-1} + \epsilon_t$ with i.i.d ϵ_t with $\mathbb{E}[\epsilon_t] = 0$ and $\mathbb{E}[\epsilon_t^2] = \sigma^2$. Then,

$$\frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}^2 \xrightarrow{D} \sigma^2 \int_0^1 B_u^2 du \tag{463}$$

$$\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\epsilon_t \xrightarrow{D} \sigma^2 \int_0^1 B_u dB_u = \frac{1}{2}(B_1^2 - 1)\sigma^2 \tag{464}$$

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 \xrightarrow{P} \sigma^2 \tag{465}$$

where $(B_u)_{u\in[0,1]}$ is a Brownian motion.

We will use these results to help us find the limiting distribution of test statistics under the null hypothesis of a unit root. In particular, we want to find the limiting distribution of:

$$T(\hat{\alpha} - 1). \tag{466}$$

Theorem 536 (Limiting Distribution of Test Statistic Under Unit Root)
The limiting distribution of the test statistic under an unit root is:

$$T(\hat{\alpha} - 1) \xrightarrow{D} \frac{\int_0^1 B_u dB_u}{\int_0^1 B_u^2 du}. \tag{467}$$

Proof. We have that:

$$\begin{aligned} T(\hat{\alpha} - 1) &= T\hat{\pi} \\ &=_{(1)} \frac{\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\epsilon_t}{\frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}^2} \xrightarrow{D} \frac{\sigma^2 \int_0^1 B_u dB_u}{\sigma^2 \int_0^1 B_u^2 du} = \frac{\int_0^1 B_u dB_u}{\int_0^1 B_u^2 du}. \end{aligned}$$

where $=_{(1)}$ uses the formula of the least squares estimator. Furthermore, the numerator uses Eq. (464) and the denominator uses Eq. (463). □

Remark 537. *Notice that the scaling factor is now T rather than $\sqrt{T}$. As a result, we say that $\hat{\alpha}$ is* **super-consistent** *in the presence of a unit root.*

We now discuss how we test for a *unit root* $H_0 : \alpha = 1$

$$Y_t = \alpha Y_{t-1} + \epsilon_t.$$

Notice that is equivalent to testing: $H_0 : \pi = 0$ in:

$$\Delta Y_t = \pi Y_{t-1} + \epsilon_t$$

as $\pi = \alpha - 1$.

We now state an important distribution that will come up in our testing.

Definition 538 (Dickey-Fuller Distribution)

The Dickey Fuller distribution is given by:

$$DF = \frac{\int_0^1 B_u dB_u}{(\int_0^1 B_u^2 du)^{1/2}} \tag{468}$$

where B_u is Brownian motion.

We can now state the distribution of our test.

Proposition 539 (Distribution of Test Statistic for Unit Root Testing)

Let Y_t be an AR(1) process. Suppose we are testing $H_0 : \pi = 0$ in:

$$\Delta Y_t = \pi Y_{t-1} + \epsilon_t$$

Then, the limiting distribution of the test statistic is:

$$Z_{\pi=0} = \frac{\hat{\pi}}{\hat{\sigma}_{\hat{\pi}}} \xrightarrow{D} DF \tag{469}$$

where DF is the Dickey-Fuller distribution.

Proof. First, write out:

$$\begin{aligned} Z_{\pi=0} &= \frac{\hat{\pi}}{\hat{\sigma}_{\hat{\pi}}} \\ &= \frac{\sum_{t=1}^T Y_{t-1}\epsilon_t}{\sum_{t=1}^T Y_{t-1}^2 \cdot \hat{\sigma}(\sum_{t=1}^T Y_{t-1}^2)^{-1/2}} \\ &= \frac{\sum_{t=1}^T Y_{t-1}\epsilon_t}{(\sum_{t=1}^T Y_{t-1}^2)^{1/2} \cdot \hat{\sigma}} \\ &= \frac{\frac{1}{T}\sum_{t=1}^T Y_{t-1}\epsilon_t}{\frac{1}{T}(\sum_{t=1}^T Y_{t-1}^2)^{1/2} \cdot \hat{\sigma}} \xrightarrow{D} \frac{\sigma^2 \int_0^1 B_u dB_u}{\sigma^2(\int_0^1 B_u^2 du)^{1/2}} = \frac{\int_0^1 B_u dB_u}{(\int_0^1 B_u^2 du)^{1/2}} \end{aligned}$$

that is $\hat{\sigma} \to \sigma$. □

We are testing:

$$H_0 : \pi = 0 \quad \text{vs} \quad H_1 : \pi \in (-2, 0)$$

whereby we reject if $Z_{\pi=0}$ falls below the 5% quantile of the DF distribution. That is, we reject for small values of the test statistic:

$$Z_{\pi=0} = \frac{\hat{\pi}}{\hat{\sigma}_{\hat{\pi}}} = \frac{(\hat{\pi} - \pi)}{\hat{\sigma}_{\hat{\pi}}} \quad \text{under } H_0 : \pi = 0.$$

This is a one-sided test whereby we only reject for small values of the test statistic. That is, it is a left hand side test.

12.17.1 Deterministic Terms

We have derived the distribution under the null hypothesis of a unit root in the absence of deterministic terms. We now want to see the effects including a constant term and a trend term.

First suppose that we suspect that an intercept is relevant (usually in the case when we think the mean of Y_t is non-zero). That is, the model is:

$$Y_t = \alpha Y_{t-1} + \mu_c + \epsilon_t$$

with the unit-root null hypothesis:

$$H_0 : \alpha = 1,$$

where μ_c is included as an unrestricted nuisance parameter.

Under H_0, we end up with a random walk:

$$Y_t = Y_{t-1} + \mu_c + \epsilon_t.$$

If $\mu_c \neq 0$, this is a random walk with drift and its recursive formulation contains a deterministic linear trend:

$$Y_t = Y_0 + \sum_{s=1}^{t} \epsilon_s + \mu_c t$$

where $\mu_c t$ is a deterministic trend.

We can represent this model with an intercept in a different manner:

$$\begin{aligned} & Y_t = \alpha Y_{t-1} + \mu_c + \epsilon_t \\ & \leftrightarrow Y_t - Y_{t-1} = \alpha Y_{t-1} - Y_{t-1} + \mu_c + \epsilon_t \\ & \leftrightarrow \Delta Y_t = (\alpha - 1) Y_{t-1} + \mu_c + \epsilon_t \end{aligned}$$

Then, the unit root hypothesis becomes:

$$H_0 : \alpha = 1 \quad \Leftrightarrow \quad H_0 : \pi = 0,$$

with μ_c unrestricted. Under this null, we get:

$$\Delta Y_t = \mu_c + \epsilon_t$$

and therefore solving for the level process gives:

$$Y_t = Y_0 + \mu_c t + \sum_{s=1}^{t} \epsilon_s.$$

The limiting distribution of the t-statistic is different from the no-intercept case:

$$Z_{\pi=0} \xrightarrow{D} DF_C$$

where DF_C is a different distribution to the ordinary Dickey Fuller distribution we saw before.

We could separately test the joint hypothesis $H_0 : \pi = \mu_c = 0$ using a likelihood-ratio test statistic:

$$LR_{\pi=\mu_c=0} \xrightarrow{D} DF_C^{(2)}$$

where $DF_C^{(2)}$ is another different distribution.

We now look at a different case where we now include a trend in our model:

$$\Delta Y_t = \pi Y_{t-1} + \mu_c + \mu_l t + \epsilon_t$$

with the null hypothesis:

$$H_0 : \pi = 0,$$

where μ_c and μ_l are included as unrestricted nuisance parameters for the Dickey–Fuller t-test.

The limiting distribution of the t-statistic based on the least squares estimator is:

$$Z_{\pi=0} \xrightarrow{D} DF_{C+l}$$

where DF_{C+l} is the Dickey–Fuller distribution appropriate when a constant and linear trend are included. We could also test the joint hypothesis $H_0 : \pi = \mu_l = 0$, with μ_c unrestricted, using a likelihood-ratio test statistic:

$$LR_{\pi=\mu_l=0} \xrightarrow{D} DF^{(2)}_{C+l}$$

where again $DF^{(2)}_{C+l}$ is a different distribution.

12.18 Cointegration

12.18.1 Beveridge-Nelson Decomposition

Suppose that we have an AR(p) process:

$$y_t = \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \epsilon_t.$$

We can rewrite this using lag operators:

$$\begin{aligned} y_t &= \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \epsilon_t \\ \leftrightarrow y_t - \alpha_1 y_{t-1} - ... - \alpha_p y_{t-p} &= \epsilon_t \\ \leftrightarrow y_t(1 - \alpha_1 L - ... - \alpha_p L^p) &= \epsilon_t \end{aligned}$$

Define the autoregressive polynomial:

$$\alpha(L) = (1 - \alpha_1 L - \alpha_2 L^2 - ... - \alpha_p L^p).$$

Now, we have that invertibility holds if the roots to:

$$(1 - \alpha_1 L - \alpha_2 L^2 - ... - \alpha_p L^p)$$

are all greater than 1. We can rewrite this as:

$$(1 - \alpha_1 L - \alpha_2 L^2 - ... - \alpha_p L^p) = (1 - \lambda_1 L)...(1 - \lambda_p L)$$

where λ_i are the reciprocals of the roots to the polynomial. A **unit root** is when a root $\frac{1}{\lambda_i}$ equals one.

Definition 540 (Integrated of Order d)
A univariate process with d unit roots is called **integrated of order d I(d).**

If a process has a single unit root, then the process is I(1). Then, we can construct a stable process by taking first differences of the original process:

$$\Delta y_t = y_t - y_{t-1}.$$

If a process is I(d), it can be made stable by differencing d times:

$$\Delta^d y_t = (1 - L)^d y_t.$$

Definition 541 (I(1) Process)
A process y_t is an I(1) process if

$$\Delta y_t = w_t$$

where w_t is a stationary process with infinite MA representation:

$$w_t = \sum_{j=0}^{\infty} \theta_j u_{t-j}.$$

A unit root process is an autoregression with a single root at unity. This means that the differenced process ΔY_t

is serially correlated but stationary. The Beveridge Nelson decomposition is a way to decompose a unit process into a permanent (random walk) component and a transitory (stationary) component.

If ΔY_t is stationary, then by the **Wold representation** theorem, it has the form:

$$\Delta Y_t = b(L)\epsilon_t \tag{470}$$

where

$$b(L) = \sum_{j=0}^{\infty} b_j L^j.$$

The lag polynomial can be generalised as:

$$b(z) = \sum_{j=0}^{\infty} b_j z^j.$$

Note that:

$$b(1) = \sum_{j=0}^{\infty} b_j.$$

To proceed, we need to show a useful property of lag polynomials.

Proposition 542 (Factorisation of Lag Polynomial)

The lag polynomial $b(z)$ can be factorised as:

$$b(z) = b(1) + (1-z)b^*(z) \tag{471}$$

where $b(1) = \sum_{j=0}^{\infty} b_j$ and $b^*(z)$ is the lag polynomial:

$$b^*(z) = \sum_{j=0}^{\infty} b_j^* z^j$$

$$b_j^* = -\sum_{l=j+1}^{\infty} b_l$$

Proof. To show this, we work from the right hand side and show that it equals the left hand side:

$$b(z) = b(1) + (1-z)b^*(z).$$

Therefore, we have that:

$$\begin{aligned}
b(1)+(1-z)b^*(z) &= \sum_{j=0}^{\infty} b_j - \sum_{j=0}^{\infty}\sum_{l=j+1}^{\infty} b_l z^j (1-z) \\
&= \sum_{j=0}^{\infty} b_j - \sum_{j=0}^{\infty}\sum_{l=j+1}^{\infty} b_l z^j + \sum_{j=0}^{\infty}\sum_{l=j+1}^{\infty} b_l z^{j+1} \\
&= \sum_{j=0}^{\infty} b_j - \sum_{j=1}^{\infty} b_j - \sum_{j=1}^{\infty}\sum_{l=j+1}^{\infty} b_l z^j + \sum_{j=0}^{\infty}\sum_{l=j+1}^{\infty} b_l z^{j+1} \\
&= b_0 - \sum_{j=1}^{\infty}\sum_{l=j+1}^{\infty} b_l z^j + \sum_{j=1}^{\infty}\sum_{l=j}^{\infty} b_l z^j \\
&= b_0 + \sum_{j=1}^{\infty} b_j z^j \\
&= \sum_{j=0}^{\infty} b_j z^j \\
&= b(z).
\end{aligned}$$

□

Now going back to our original problem, we had the stationary ΔY_t in a Wold representation:

$$\Delta Y_t = b(L)\epsilon_t.$$

We can apply our factorisation to $b(L)$ and this gives us the Beveridge–Nelson decomposition.

Theorem 543 (Beveridge–Nelson Decomposition)
Let Y_t be a unit root process. Then, the **Beveridge–Nelson decomposition** is given by:

$$Y_t = S_t + U_t + V_0 \tag{472}$$

where $S_t = \sum_{i=1}^{t} \xi_i$ is a random walk, whereby ξ_i is a white noise process, U_t is a strictly stationary process, and V_0 is an initial condition.

Proof.

$$\Delta Y_t = b(L)\epsilon_t.$$

We then apply our factorisation to $b(L) = b(1) + (1-L)b^*(L)$ to get:

$$\begin{aligned}
\Delta Y_t &= \Big(b(1) + (1-L)b^*(L)\Big)\epsilon_t \\
&= b(1)\epsilon_t + (1-L)b^*(L)\epsilon_t \\
&= b(1)\epsilon_t + b^*(L)\epsilon_t - b^*(L)\epsilon_{t-1} \\
&\equiv \xi_t + U_t - U_{t-1}
\end{aligned}$$

where ξ_t is the permanent innovation:

$$\xi_t = b(1)\epsilon_t$$

and U_t, U_{t-1} is strictly stationary under assumptions of **square-summability** of b_j^*. We then sum both sides and use telescoping sums:

$$\sum_{i=1}^{t}(Y_i - Y_{i-1}) = \sum_{i=1}^{t}\xi_i + \sum_{i=1}^{t}(U_i - U_{i-1})$$
$$\leftrightarrow Y_t - Y_0 = \sum_{i=1}^{t}\xi_i + U_t - U_0$$
$$\leftrightarrow Y_t = \sum_{i=1}^{t}\xi_i + U_t + Y_0 - U_0$$
$$\equiv Y_t = S_t + U_t + V_0$$

where we defined:

$$S_t = \sum_{i=1}^{t}\xi_i$$
$$U_t = b^*(L)\epsilon_t$$
$$V_0 = Y_0 - U_0$$

□

If we look at:

$$Y_t = S_t + U_t + V_0$$

we have that S_t is the **permanent** (trend) component of Y_t as it determines the long-run behaviour of Y_t and U_t is the **transitory component**.

12.18.2 Cointegration

We may suspect that there are equilibrium relationships between economic variables. Let us collect k variables into:

$$\boldsymbol{y}_t = \begin{bmatrix} y_{1t} \\ \dots \\ \dots \\ \dots \\ y_{Kt} \end{bmatrix}.$$

Suppose that the K variables have that their long-run equilibrium relation is:

$$\boldsymbol{\beta}^T \boldsymbol{y}_t = \beta_1 y_{1t} + \dots + \beta_K y_{Kt} = 0$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^T$. However, in any given period, they may instead be at the deviation from the long-run level:

$$\boldsymbol{\beta}^T \boldsymbol{y}_t = \boldsymbol{z}_t$$

where $\boldsymbol{z}_t$ measures the deviation from equilibrium. However, it may be the case that all the variables are driven by a **common stochastic trend** so that all the variables in $\boldsymbol{y}_t$ are I(1). However, it may be the case that there exists a

linear combination:

$$z_t \equiv \boldsymbol{\beta}^T \boldsymbol{y}_t$$

such that z_t is I(0). Then we say that $\boldsymbol{y}_t \sim CI(1,1)$ whereby the vector $\boldsymbol{\beta}$ is called a **cointegrating/cointegration vector** and $\boldsymbol{y}_t$ is called a **cointegrated process**.

Definition 544 (Cointegrated of Order 1)

A set of I(1) time series variables $\boldsymbol{y}_t$ is **cointegrated** if there exists a linear combination $\boldsymbol{\beta}^T \boldsymbol{y}_t$ of these variables such that:

$$\boldsymbol{\beta}^T \boldsymbol{y}_t \tag{473}$$

is I(0). $\boldsymbol{\beta}$ is called the cointegrated vector.

Remark 545. *Another different definition of cointegration is that a K-dimensional process $\boldsymbol{y}_t$ is integrated of order d, $\boldsymbol{y}_t \sim I(d)$ if $\Delta^d \boldsymbol{y}_t$ is stable and $\Delta^{d-1} \boldsymbol{y}_t$ is not stable.*

Example 546

Suppose that $y_t \sim I(1)$ and $z_t \sim I(1)$ and

$$2y_t - 3z_t \sim I(0).$$

Then (y_t, z_t) is cointegrated with cointegration vector $\boldsymbol{\beta} = (2, -3)$.

We now give a motivation of error correction models and see their relationship to cointegration. In error correction models, changes in a variable depends on the deviations from some equilibrium relationship. We now show an example of how does cointegration and ECM relate to each other.

Example 547

(Basic ECM Cointegration). Suppose that y_{1t} and y_{2t} is the price of the same good in two different markets. Suppose that the equilibrium relationship between the two variables are $y_{1t} = \beta_1 y_{2t}$. By the error correction model, the change in the level of y_{1t} and y_{2t} depends on the deviations from this equilibrium in the period $t-1$:

$$\Delta y_{1t} = \alpha_1(y_{1,t-1} - \beta_1 y_{2,t-1}) + u_{1t}$$
$$\Delta y_{2t} = \alpha_2(y_{1,t-1} - \beta_1 y_{2,t-1}) + u_{2t}$$

To further extend upon this, suppose that Δy_{it} also depends on previous changes in both variables too:

$$\Delta y_{1t} = \alpha_1(y_{1,t-1} - \beta_1 y_{2,t-1}) + \gamma_{11,1}\Delta y_{1,t-1} + \gamma_{12,1}\Delta y_{2,t-1} + u_{1t}$$
$$\Delta y_{2t} = \alpha_2(y_{1,t-1} - \beta_1 y_{2,t-1}) + \gamma_{21,1}\Delta y_{1,t-1} + \gamma_{22,1}\Delta y_{2,t-1} + u_{2t}$$

Now, suppose that y_{1t} and y_{2t} are I(1) variables and u_{it} are white noise errors (therefore they are stable). We have that every term except for the first term on the right hand side is a stable process. We re-arrange for this term:

$$\alpha_1(y_{1,t-1} - \beta_1 y_{2,t-1}) = \Delta y_{1t} - \gamma_{11,1}\Delta y_{1,t-1} - \gamma_{12,1}\Delta y_{2,t-1} - u_{1t}$$
$$\alpha_2(y_{1,t-1} - \beta_1 y_{2,t-1}) = \Delta y_{2t} - \gamma_{21,1}\Delta y_{1,t-1} - \gamma_{22,1}\Delta y_{2,t-1} - u_{2t}$$

For the right-hand side to be stable, it must be the case that the left-hand side $y_{1,t-1} - \beta_1 y_{2,t-1}$ is also stable, assuming that $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$. However, by definition, this is cointegration! Therefore, we see that the error correction mechanism is also the source of the cointegration.

12.18.3 Cointegration for VAR(2)

We now generalise our example to a VAR(2) model.

Example 548

(ECMT Cointegration with VAR(2)). We take the same price example from before:

$$\Delta y_{1t} = \alpha_1(y_{1,t-1} - \beta_1 y_{2,t-1}) + \gamma_{11,1}\Delta y_{1,t-1} + \gamma_{12,1}\Delta y_{2,t-1} + u_{1t}$$
$$\Delta y_{2t} = \alpha_2(y_{1,t-1} - \beta_1 y_{2,t-1}) + \gamma_{21,1}\Delta y_{1,t-1} + \gamma_{22,1}\Delta y_{2,t-1} + u_{2t}$$

Now, we rewrite this in terms of vectors and matrices:

$$\boldsymbol{y}_t - \boldsymbol{y}_{t-1} = \boldsymbol{\alpha}\boldsymbol{\beta}^T \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1(\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}) + \boldsymbol{u}_t$$

whereby

$$\boldsymbol{\alpha} \equiv \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \quad \boldsymbol{\beta}^T \equiv \begin{bmatrix} 1 & -\beta_1 \end{bmatrix} \quad \boldsymbol{\Gamma}_1 \equiv \begin{bmatrix} \gamma_{11,1} & \gamma_{12,1} \\ \gamma_{21,1} & \gamma_{22,1} \end{bmatrix}.$$

We then re-arrange to get:

$$\boldsymbol{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}^T \boldsymbol{y}_{t-1} + \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1 \boldsymbol{y}_{t-1} - \boldsymbol{\Gamma}_1 \boldsymbol{y}_{t-2} + \boldsymbol{u}_t$$

Now, we notice that we now have a VAR(2) representation:

$$\boldsymbol{y}_t = (\boldsymbol{I}_K + \boldsymbol{\Gamma}_1 + \boldsymbol{\alpha}\boldsymbol{\beta}^T)\boldsymbol{y}_{t-1} - \boldsymbol{\Gamma}_1 \boldsymbol{y}_{t-2} + \boldsymbol{u}_t$$

Hence, as we have shown that our cointegrated system can be written as a VAR(2) model.

We have shown how to go from cointegration to VAR(2). We now want to do the reverse.

Let us generalise the coefficient matrices in order to tie this to unit roots. Suppose we had the VAR(2) process:

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \boldsymbol{A}_2 \boldsymbol{y}_{t-2} + \boldsymbol{u}_t$$

where $\boldsymbol{y}_t$ is a K-dimensional vector now. Rewrite this using lag operator:

$$(\boldsymbol{I}_K - \boldsymbol{A}_1 L - \boldsymbol{A}_2 L^2)\boldsymbol{y}_t = \boldsymbol{u}_t$$

Suppose that the process is unstable whereby the autoregressive polynomial:

$$det(\boldsymbol{I}_K - \boldsymbol{A}_1 L - \boldsymbol{A}_2 L^2) = (1 - \lambda_1 L)...(1 - \lambda_n L)$$

has at least one root being equal to one. Elaborating on this, recall that to find a given root j, we use:

$$1 - \lambda_j L = 0 \quad \leftrightarrow \quad L = \frac{1}{\lambda_j}.$$

Meanwhile, all other roots are outside the unit circle. Since we have a unit root, this implies that:

$$det(\boldsymbol{I}_K - \boldsymbol{A}_1 L - \boldsymbol{A}_2 L^2) = (1 - \lambda_1 L)...(1 - \lambda_n L) = 0$$

and therefore the matrix:

$$\boldsymbol{\Pi} \equiv \boldsymbol{A}_1 + \boldsymbol{A}_2 - \boldsymbol{I}_K$$

is singular as it has a determinant of 0. Suppose that the $rank(\boldsymbol{\Pi}) = r < K$. Then we can decompose this matrix:

$$\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $(K \times r)$ matrices. As we assumed I(1) variables, we can difference the VAR(2):

$$\begin{aligned}
&\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \boldsymbol{A}_2 \boldsymbol{y}_{t-2} + \boldsymbol{u}_t \\
&\leftrightarrow \boldsymbol{y}_t - \boldsymbol{y}_{t-1} = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \boldsymbol{A}_2 \boldsymbol{y}_{t-2} + \boldsymbol{u}_t - \boldsymbol{y}_{t-1} \\
&\leftrightarrow \boldsymbol{y}_t - \boldsymbol{y}_{t-1} = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \boldsymbol{A}_2 \boldsymbol{y}_{t-2} + \boldsymbol{u}_t - \boldsymbol{y}_{t-1} + \boldsymbol{A}_2 \boldsymbol{y}_{t-1} - \boldsymbol{A}_2 \boldsymbol{y}_{t-1} \\
&\leftrightarrow \boldsymbol{y}_t - \boldsymbol{y}_{t-1} = (\boldsymbol{A}_1 + \boldsymbol{A}_2 - \boldsymbol{I}_K)\boldsymbol{y}_{t-1} - \boldsymbol{A}_2(\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}) + \boldsymbol{u}_t \\
&\leftrightarrow \boldsymbol{y}_t - \boldsymbol{y}_{t-1} = (\boldsymbol{A}_1 + \boldsymbol{A}_2 - \boldsymbol{I}_K)\boldsymbol{y}_{t-1} - \boldsymbol{A}_2 \Delta \boldsymbol{y}_{t-1} + \boldsymbol{u}_t \\
&\leftrightarrow \Delta \boldsymbol{y}_t = \boldsymbol{\Pi} \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + \boldsymbol{u}_t
\end{aligned}$$

where we defined $\boldsymbol{\Gamma}_1 \equiv -\boldsymbol{A}_2$ and $\boldsymbol{\Pi} \equiv \boldsymbol{A}_1 + \boldsymbol{A}_2 - \boldsymbol{I}_K$. We re-arrange to get:

$$\Delta \boldsymbol{y}_t - \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} - \boldsymbol{u}_t = \boldsymbol{\Pi} \boldsymbol{y}_{t-1}.$$

As all the terms on the left-hand side are stationary, this implies that $\boldsymbol{\Pi} \boldsymbol{y}_{t-1} = \boldsymbol{\alpha} \boldsymbol{\beta}^T \boldsymbol{y}_{t-1}$ is also stationary. Furthermore, stationarity holds when you multiply by $(\boldsymbol{\alpha}^T \boldsymbol{\alpha})^{-1} \boldsymbol{\alpha}^T$:

$$(\boldsymbol{\alpha}^T \boldsymbol{\alpha})^{-1} \boldsymbol{\alpha}^T \boldsymbol{\alpha} \boldsymbol{\beta}^T \boldsymbol{y}_{t-1} = \boldsymbol{\beta}^T \boldsymbol{y}_{t-1}$$

so then this implies that $\boldsymbol{\beta}^T \boldsymbol{y}_{t-1}$ is also stationary. But this is the definition of a cointegrating relationship again!

12.18.4 Cointegration for VAR(p)

We now look at the K-dimensional case. Suppose that we have the K-dimensional VAR(p) process:

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + ... + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t. \tag{474}$$

where each variable is either I(1) or I(0).

Definition 549 (Cointegration of VAR(p))
Let $\boldsymbol{y}_t$ be a K-dimensional VAR(p) process:

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + ... + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t. \tag{475}$$

Then, $\boldsymbol{y}_t$ is called **cointegrated of rank r** if

$$\boldsymbol{\Pi} \equiv -(\boldsymbol{I}_K - \boldsymbol{A}_1 - ... - \boldsymbol{A}_p) \tag{476}$$

has rank r. Then, the matrix $\boldsymbol{\Pi}$ can be written as:

$$\boldsymbol{\Pi} = \boldsymbol{\alpha} \boldsymbol{\beta}^T \tag{477}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $(K \times r)$ matrices with rank r, so $\boldsymbol{\beta}^T$ is $(r \times K)$. The columns of $\boldsymbol{\beta}$ are called the **cointegrating vectors** and $\boldsymbol{\alpha}$ is called the **loading matrix**.

Remark 550. *Notice the two special cases where $r = 0$ and $r = K$. When $r = 0$, then $\Delta \boldsymbol{y}_t$ has a stable VAR(p-1) representation. When $r = K$, then $det(-\boldsymbol{\Pi}) \neq 0$ and hence the VAR(p) has no unit roots so that $\boldsymbol{y}_t$ is a stable VAR(p) process.*

Given our VAR(p) model:

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + ... + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t$$

we can write this VAR into a vector error correction form.

Definition 551 (Vector Error Correction Model)
Let $\boldsymbol{y}_t$ be a VAR(p) process. Then, the associated **vector error correction model** for $\boldsymbol{y}_t$ is:

$$\Delta \boldsymbol{y}_t = \boldsymbol{\Pi} \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + ... + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t \tag{478}$$

where

$$\boldsymbol{\Pi} = -(\boldsymbol{I}_k - \boldsymbol{A}_1 - ... - \boldsymbol{A}_p)$$
$$\boldsymbol{\Gamma}_i = -(\boldsymbol{A}_{i+1} + ... + \boldsymbol{A}_p)$$

for $i = 1, ..., p-1$.

Remark 552. *Notice that due to the decomposition of* $\boldsymbol{\Pi}$*, we could have written the VECM as:*

$$\Delta \boldsymbol{y}_t = \boldsymbol{\alpha} \boldsymbol{\beta}^T \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + ... + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t$$

We have shown how to go from a VAR(p) to a VECM. We can go in reverse too. Suppose we are given a VECM:

$$\Delta \boldsymbol{y}_t = \boldsymbol{\Pi} \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + ... + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t$$
$$\boldsymbol{\Pi} = -(\boldsymbol{I}_k - \boldsymbol{A}_1 - ... - \boldsymbol{A}_p)$$
$$\boldsymbol{\Gamma}_i = -(\boldsymbol{A}_{i+1} + ... + \boldsymbol{A}_p)$$

we can also retrieve the VAR(p) form by setting:

$$\boldsymbol{A}_1 = \boldsymbol{\Pi} + \boldsymbol{I}_K + \boldsymbol{\Gamma}_1$$
$$\boldsymbol{A}_i = \boldsymbol{\Gamma}_i - \boldsymbol{\Gamma}_{i-1} \quad \text{for } i = 2, ..., p-1$$
$$\boldsymbol{A}_p = -\boldsymbol{\Gamma}_{p-1}$$

which gives us back the VAR(p) model:

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + ... + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t$$

Linear Algebra Review

We collect some results from linear algebra.

Definition 553 (Orthogonal Complement)
If U is a subset of vector space V, then the **orthogonal complement** of U, denoted $U_\perp$, is the set of all vectors in V that are orthogonal to every vector in U:

$$U_\perp = \{v \in V : v \cdot u = 0 \text{ for every } u \in U\}. \tag{479}$$

Proposition 554 (Direct Sum)

Suppose U is a finite-dimensional subspace of the vector space V. Then

$$V = U \bigoplus U_{\perp} \tag{480}$$

where $U_{\perp}$ is the orthogonal complement to U.

This result is important as this means that for any vector $v \in V$, we can decompose it into the sum of a vector in U and in a vector in $U_{\perp}$.

Proposition 555 (Dimension of the Orthogonal Complement)

Suppose V is finite-dimensional and U is a subspace of V. Then

$$dim\, U_{\perp} = dim\, V - dim\, U. \tag{481}$$

Example 556

Let the vector space be $V = \mathbb{R}^5$. Then, suppose we have the matrix:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 6 \\ 3 & 9 \\ 5 & 3 \end{bmatrix}.$$

We can then define the subset $W \subset V$ as:

$$W = \{u \in V : \boldsymbol{A}x = u, x \in \mathbb{R}^2\}.$$

Notice that $\boldsymbol{A}$ has a column rank of 2. Now, we can define the orthogonal complement as:

$$W_{\perp} = \{v \in V : u \cdot v = 0, u \in W\}.$$

Alternatively, we can construct an orthogonal matrix $\tilde{\boldsymbol{A}}$:

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} -2 & -3 & -5 \\ -6 & -9 & -3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Here, we have that:

$$W_{\perp} = \{v \in V : \tilde{\boldsymbol{A}}y = v, y \in \mathbb{R}^3\}.$$

Notice that every column in $\boldsymbol{A}$ is orthogonal to every column in $\tilde{\boldsymbol{A}}$.

We now summarise this section of linear algebra. Let $\boldsymbol{X}$ be a $m \times n$ matrix whereby $m \geq n$. Then, we denote $\boldsymbol{X}_{\perp}$ as the **orthogonal complement** of the $m \times n$ matrix with $rank(\boldsymbol{X}) = n$. Then, $\boldsymbol{X}_{\perp}$ is a $(m \times (m - n))$ matrix with

$rank(\boldsymbol{X}_\perp) = m - n$. Furthermore, this matrix is orthogonal:

$$\boldsymbol{X}^T \boldsymbol{X}_\perp = 0.$$

12.18.5 Granger Representation Theorem

We now give another representation of a cointegrated system $\boldsymbol{y}_t$. To do so, we need two assumptions.

Assumption 557 (Granger Representation)

Let $\boldsymbol{y}_t$ be a cointegrated vector autoregressive model.

1. The characteristic polynomial of $\boldsymbol{y}_t$ has stationary and unit roots only.
2. The number of unit roots equal $K - r$.

Definition 558 (Cointegrating Rank)

Let $\boldsymbol{\Pi}$ be the matrix associated to the VECM of $\boldsymbol{y}_t$. Then $\boldsymbol{y}_t$ has cointegrating rank r if:

$$\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T \tag{482}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{K \times r}$.

Theorem 559 (Granger VECM Representation Theorem)

Let $\boldsymbol{y}_t$ be a K-dimensional cointegrated I(1) process with cointegrating rank r where $0 \leq r < K$. Furthermore, assume that $\boldsymbol{y}_t$ has the vector error correction model representation:

$$\Delta \boldsymbol{y}_t = \boldsymbol{\Pi} \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + ... + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t \tag{483}$$

Then, the **Granger representation** is given by:

$$\boldsymbol{y}_t = \boldsymbol{C} \sum_{i=1}^{t} \boldsymbol{u}_i + \boldsymbol{C}^*(L)\boldsymbol{u}_t + \boldsymbol{y}_0^* \tag{484}$$

where

$$\boldsymbol{C} = \boldsymbol{\beta}_\perp \left[\boldsymbol{\alpha}_\perp^T \left(\boldsymbol{I}_K - \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \right) \boldsymbol{\beta}_\perp \right]^{-1} \boldsymbol{\alpha}_\perp^T$$

$\boldsymbol{C} \in \mathbb{R}^{K \times K}$, has rank $K - r$, and

$$\boldsymbol{C}^*(L)\boldsymbol{u}_t = \sum_{j=0}^{\infty} \boldsymbol{C}_j^* \boldsymbol{u}_{t-j}$$

is an I(0) process with $\boldsymbol{C}_j^*$ is a coefficient matrix and $\boldsymbol{y}_0^*$ contains the initial values. Furthermore, it follows that $\boldsymbol{y}_t$ is a cointegrated I(1) process with cointegrating vectors $\boldsymbol{\beta}$. Finally, we have that:

$$\boldsymbol{\beta}^T \boldsymbol{C} = 0 \quad \boldsymbol{C}\boldsymbol{\alpha} = 0$$

where $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ has dimensions $K \times r$ and the orthogonal complements $(\boldsymbol{\alpha}_\perp, \boldsymbol{\beta}_\perp)$ has dimensions $K \times (K - r)$.

Proof. (Sketch). The matrix:

$$\boldsymbol{\alpha}_{\perp}^{T}\left(\boldsymbol{I}_{K}-\sum_{i=1}^{p-1}\boldsymbol{\Gamma}_{i}\right)\boldsymbol{\beta}_{\perp}$$

is a $(K-r)\times(K-r)$ matrix and hence for this matrix to be invertible, it must be full rank K - r. This then implies that $\boldsymbol{C}$ has rank $K-r$. The number of unit roots is exactly $K-r$ as a result of this too. □

Remark 560. *The Granger representation theorem is a multivariate version of the Beveridge-Nelson decomposition. Here, $\boldsymbol{y}_t$ is decomposed into I(1) and I(0) components. In particular, the process $\boldsymbol{y}_t$ is driven by $K-r$ I(1) components and r I(0) components.*

The first term:

$$\boldsymbol{C}\sum_{i=1}^{t}\boldsymbol{u}_{i}$$

consists of K random walks $\sum_{i=1}^{t}\boldsymbol{u}_i$ being multiplied by a matrix with rank $K-r$, denoted by $\boldsymbol{C}$. This means that there are in fact $K-r$ stochastic trends driving the system.

The second term:

$$\boldsymbol{C}^{*}(L)\boldsymbol{u}_{t}=\sum_{j=0}^{\infty}\boldsymbol{C}_{j}^{*}\boldsymbol{u}_{t-j}$$

is an I(0) component where $\boldsymbol{C}_j^*$ is a coefficient matrix similar to that seen in the Beveridge–Nelson decomposition. The expression for this coefficient matrix is complicated and not needed.

This representation has some interesting properties. First, if we multiply the matrix by the cointegrating vector $\boldsymbol{\beta}$, we eliminate the random walk and we get:

$$\begin{aligned}
\boldsymbol{y}_{t}&=\boldsymbol{C}\sum_{i=1}^{t}\boldsymbol{u}_{i}+\boldsymbol{C}^{*}(L)\boldsymbol{u}_{t}+\boldsymbol{y}_{0}^{*}\\
\boldsymbol{\beta}^{T}\boldsymbol{y}_{t}&=\boldsymbol{\beta}^{T}\boldsymbol{C}\sum_{i=1}^{t}\boldsymbol{u}_{i}+\boldsymbol{\beta}^{T}\boldsymbol{C}^{*}(L)\boldsymbol{u}_{t}+\boldsymbol{\beta}^{T}\boldsymbol{y}_{0}^{*}\\
\boldsymbol{\beta}^{T}\boldsymbol{y}_{t}&=\boldsymbol{\beta}^{T}\left[\boldsymbol{\beta}_{\perp}\left[\boldsymbol{\alpha}_{\perp}^{T}\left(\boldsymbol{I}_{K}-\sum_{i=1}^{p-1}\boldsymbol{\Gamma}_{i}\right)\boldsymbol{\beta}_{\perp}\right]^{-1}\boldsymbol{\alpha}_{\perp}^{T}\right]\sum_{i=1}^{t}\boldsymbol{u}_{i}+\boldsymbol{\beta}^{T}\boldsymbol{C}^{*}(L)\boldsymbol{u}_{t}+\boldsymbol{\beta}^{T}\boldsymbol{y}_{0}^{*}\\
\boldsymbol{\beta}^{T}\boldsymbol{y}_{t}&=\boldsymbol{\beta}^{T}\boldsymbol{C}^{*}(L)\boldsymbol{u}_{t}+\boldsymbol{\beta}^{T}\boldsymbol{y}_{0}^{*}
\end{aligned}$$

and since $\boldsymbol{C}^*(L)\boldsymbol{u}_t$ is an I(0) process, we are left with an I(0) process once we premultiply $\boldsymbol{y}_t$ by $\boldsymbol{\beta}^T$.

13 Panel Data

We now look at panel data, which comprises both of a cross-sectional and time-series variation. Panel data is quite useful because if we had omitted **time-invariant** variables (i.e. omitted variables that do not change over time), then we can overcome this omitted variable problem by allowing for individual-specific intercepts which will control for these omitted variables. That is, the intercept will absorb the omitted variable error. We will be looking small T panel datasets and taking number of cross-sectional observations $n \to \infty$ when we do asymptotics.

13.1 Two-Way Error Components Model

We will test and impose restrictions on our panel dataset that require some of the estimated parameters to be common to all the observation units. These restrictions are known as **pooling restrictions**, and estimators which impose these restrictions are known as **pooled estimators**.

We will have three different types of variables.

1. Observations that vary both across individual observation unit i and over time t. An example of this could be someone's wage over their lifetime. Each person will have a different wage, and for a given person, their wage will vary over the course of their lifetime.
2. Observations that vary across individual observations i but not over time. An example of this could be someone's name or gender whereby different people will have different names but for that person, their name or gender will remain the same throughout the course of their life.
3. Observations that vary across time t but not individual observations. An example of this could be the interest rate on government bonds that individuals can save in. Everyone has access to the same interest rate, so this does not vary across individuals. However, the interest rate on government bonds does vary over time.

As we have three different types of variables, we can also define three different types of errors associated to these three different variables.

Definition 561 (Panel Data Explanatory Variables and Errors) 1. $\boldsymbol{x}_{it}$: Explanatory variables that vary over individuals and time.

2. $\boldsymbol{w}_i$: Explanatory variables that vary over individuals.
3. $\boldsymbol{s}_t$: Explanatory variables that vary over time.

The associated error terms are:

1. v_{it} for variables that vary over individuals and time,
2. η_i for variables that vary over individuals,
3. f_t for variables that vary over time.

Remark 562. *The best way to memorise this is to look at the subscript of the variable. If the variable has both i for individual and t for time, so in this $\boldsymbol{x}_{it}$, then it is a variable that is varying across time and individuals.*

If the variable only has the t subscript, such as $\boldsymbol{s}_t$, then we automatically know it must be the explanatory variable that is varying over time but is the same across individuals.

We can now state the linear model involving these variables.

Definition 563 (Two-Way Error Components Model)

The two-way error components model is:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{\gamma} + \boldsymbol{s}_t\boldsymbol{\delta} + (v_{it} + \eta_i + f_t). \tag{485}$$

Remark 564. *It is called two-way error even though we have three error terms!*

Remark 565. *As we will later see, we have not said anything about η which is known as the unobserved heterogeneity between individuals. In particular, we have not said anything about its correlation with $\boldsymbol{X}_i$. This is quite important and will lead to assumptions known as fixed and random effects later on.*

13.1.1 Time Dummies

Before we proceed, we will now show how to deal with the time variable $\boldsymbol{s}_t$ and associated error term f_t in the two-way error components model quite easily. The idea is that if we include T dummy variables, which acts as an intercept, we can absorb the effect of these time-varying variables into these T different intercepts. We now formally show this.

First, start off with the two-way error components model:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{\gamma} + \boldsymbol{s}_t\boldsymbol{\delta} + (v_{it} + \eta_i + f_t).$$

Now, we want to combine the time-varying explanatory variable $\boldsymbol{s}_t$ and the time-varying error component f_t as:

$$\phi_t = \boldsymbol{s}_t\boldsymbol{\delta} + f_t. \tag{486}$$

We can then substitute this into the two-way error components model to get:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{\gamma} + (\boldsymbol{s}_t\boldsymbol{\delta} + f_t) + (v_{it} + \eta_i).$$

which then becomes:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{\gamma} + \phi_t + (v_{it} + \eta_i).$$

Therefore, we can absorb this into the error term and we now have the new equation:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{\gamma} + (v_{it} + \eta_i + \phi_t). \tag{487}$$

We now define time dummy variables.

Definition 566 (Time Dummies)

Time dummy variables are dummy variables which indicate which period of time we are in:

$$D_t^j = \begin{cases} 1 & \text{for observations in period j} \\ 0 & \text{otherwise.} \end{cases} \tag{488}$$

for $j = 1, ..., T$.

Therefore, we can take our equation:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{\gamma} + (v_{it} + \eta_i + \phi_t).$$

and notice that

$$D_t^s \phi_t = \phi_t \quad \text{if } s = t.$$

Therefore, we can therefore rewrite our earlier equation as:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{\gamma} + \phi_1 D_t^1 + \phi_2 D_t^2 + ... + \phi_T D_t^T + (v_{it} + \eta_i).$$

We do not have an intercept in $\boldsymbol{x}_{it}$ and therefore we do not suffer from a **dummy variable trap** when we include these T dummy variables. This equation has period-specific intercepts whereby for each period s, we will have a different intercept. This intercept will then absorb the error term f_t. This equation has a special name.

Definition 567 (One-way Error Components Model)
The one-way error components model is given by:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{\gamma} + \sum_{s=1}^{T} \phi_s D_t^s + (v_{it} + \eta_i) \tag{489}$$

where $\boldsymbol{x}_{it}$ is a $1 \times K$ vector, $\boldsymbol{\beta}$ is a $K \times 1$ vector, $\boldsymbol{w}_i$ is a $1 \times G$ vector and $\boldsymbol{\gamma}$ is a $G \times 1$ vector.

With the one-way error components model, we can estimate the T parameters $(\phi_1, ..., \phi_T)$ alongside the original $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ vectors. With small T panels, losing identification of the $\boldsymbol{\delta}$ parameter associated to the time varying variable $\boldsymbol{s}_t$ is not a concern. We also say that this model is **saturated** in the time dimension.

Writing the one-way error components model is quite a useful tool as this means we can now not worry about time-varying $\boldsymbol{s}_t\boldsymbol{\delta} + f_t$ terms as we can simply include these time dummy variables into our model and take care of them.

13.2 Pooled OLS in Linear Panel Model

From our discussion in the last section, we can focus on the one-way error components model whereby we take care of the time varying variables by simply including a dummy variable. Hence, we are interested in the model:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{\gamma} + (v_{it} + \eta_i)$$

for $i = 1, ..., N$ and $t = 1, ..., T$. Now, we want to stack the T observations for each individual observation i to get:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{W}_i\boldsymbol{\gamma} + (\boldsymbol{v}_i + \eta_i \boldsymbol{j}_T)$$

whereby the dimensions are:

1. $\boldsymbol{y}_i$ is a $T \times 1$ vector which represents all T time observations for individual i.
2. $\boldsymbol{X}_i$ is a $T \times K$ matrix where each row t is the K explanatory variables for observation i at time t and each column j represents all T time observations for individual i for explanatory variable j.
3. $\boldsymbol{W}_i$ is a $T \times G$ matrix of time-invariant variables. Each row contains the G time-invariant variables. Each row is identical as these variables do not vary over time.
4. $\boldsymbol{v}_i$ is a $T \times 1$ vector where each row t contains the error term for observation i at time t.
5. $\eta_i \boldsymbol{j}_T$ is a $T \times 1$ vector where η_i is a scalar and $\boldsymbol{j}_T$ is a $T \times 1$ vector full of ones.

Now, we can stack over the N individuals to get:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{\gamma} + (\boldsymbol{\eta} + \boldsymbol{v}) \tag{490}$$

where the dimensions are:

1. $\boldsymbol{y}$ is a $NT \times 1$ vector.
2. $\boldsymbol{X}$ is a $NT \times K$ matrix.
3. $\boldsymbol{W}$ is a $NT \times G$ matrix.
4. $\boldsymbol{v}$ is a $NT \times 1$ vector.
5. $\boldsymbol{\eta}$ is a $NT \times 1$ vector.

Before, we proceed, we make an assumption.

Assumption 568 (No Time-Invariant Variables)
We assume that there are no time-invariant explanatory variables.

$$\boldsymbol{W} = \boldsymbol{0}. \tag{491}$$

However, note that even if $\boldsymbol{W} = \boldsymbol{0}$, we can still have time invariant errors $\boldsymbol{\eta}_i$. Therefore, our model is now:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{\eta} + \boldsymbol{v}).$$

Now, if we want to carry out estimation, we need **two** more assumptions over:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + (\boldsymbol{\eta}_i\boldsymbol{j}_T + \boldsymbol{v}_i).$$

Assumption 569 (Cross-Sectional Independence)
Observations on $(y_i, \boldsymbol{X}_i)$ are independent over $i = 1, ..., N$.

Remark 570. *This says that if you know $(\boldsymbol{X}_i, \boldsymbol{y}_i)$ for one person, it tells you nothing about person j's $(\boldsymbol{X}_j, \boldsymbol{y}_j)$.*

Assumption 571 (Slope Parameter Homogeneity)
The parameters in $\boldsymbol{\beta}$ are common to all observations $i = 1, ..., N$.

We can now estimate $\boldsymbol{\beta}$ under all these assumptions.

Definition 572 (Pooled OLS Estimator)
Suppose we had no time-invariant variables $\boldsymbol{w}_i$. We then have the structural linear panel model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{492}$$

where $\boldsymbol{u} = \boldsymbol{\eta} + \boldsymbol{v}$. Furthermore, assume that we had cross-sectional independence and slope parameter homogeneity. Then, the **pooled ordinary least squares** estimator is:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{493}$$

Remark 573. *The pooled OLS is also known as the* **levels OLS**.

The pooled OLS corresponds to running OLS on the observations pooled across individuals i over time t and then computing the OLS.

We now want to establish asymptotic properties of the pooled OLS estimator. To do so, we require more assumptions. We now go through them.

Assumption 574 (Errors Have Expected Value of Zero)
The error components have expected value of zero:

$$\mathbb{E}[u_{it}] = \mathbb{E}[\eta_i] + \mathbb{E}[v_{it}] = 0. \tag{494}$$

Remark 575. *This assumption is not actually not needed but this makes notation easier since we do not need to include intercept terms in our analysis.*

Recall that when we looked at consistency in the setting of OLS, we require assumptions on orthogonality between the regressor $\boldsymbol{X}$ and the error term $\boldsymbol{u}$. In panel data, we need to do a similar thing whereby we need to impose orthogonality between the explanatory variables $\boldsymbol{x}_{it}$ and the error terms v_{it} and η_i. In fact, for this entire section on panel data, the aim is to know which orthogonality conditions we will need between $\boldsymbol{x}_{it}$ and the error terms. Hence, pay close attention to each time we mention what is the orthogonality condition that we are imposing.

We now discuss what assumptions we shall impose on the correlation between the error term $u_{it} = \eta_i + v_{it}$ and $\boldsymbol{x}_{it}$.

Assumption 576 (Contemporaneous Exogeneity)
The K explanatory variables are uncorrelated with the time varying individual error term in the same time period:

$$\mathbb{E}[\boldsymbol{x}_{it} v_{it}] = 0 \tag{495}$$

for $t = 1, 2, ..., T$.

Remark 577. *This says that $\boldsymbol{x}_{it}$ is uncorrelated with the individual time-varying error term v_{it} in every period t.*

Note that under contemporaneous exogeneity, it may still be the case that:

$$\mathbb{E}[\boldsymbol{x}_{it} v_{i,t-k}] \neq 0$$

for $k > 0$ so that $\boldsymbol{x}_{it}$ may be correlated with lagged errors $v_{i,t-k}$.

We now need to discuss the correlation between $\boldsymbol{x}_{it}$ and the time-invariant (individual specific) error term η_i.

Assumption 578 (Uncorrelated Individual Effects/Random Effects)
All K explanatory variables $\boldsymbol{x}_{it}$ is uncorrelated with all individual-specific effects η_i:

$$\mathbb{E}[\boldsymbol{x}_{it} \eta_i] = \boldsymbol{0} \tag{496}$$

for $t = 1, 2, ..., T$.

We have stated the two assumptions we need for orthogonality between $\boldsymbol{x}_{it}$ and the two error terms. Before we proceed, we need one more assumption that we also needed in OLS.

Assumption 579 (Full Rank)
The matrix $\boldsymbol{X}$ is full rank.

We now collect all the assumptions together to recall what we have assumed so far.

Assumption 580 (Assumptions for Consistency of Pooled OLS)
1. $(y_i, \boldsymbol{X}_i)$ are independent over $i = 1, ..., N$.
2. $\boldsymbol{\beta}$ are common to all observations $i = 1, ..., N$.
3. The matrix $\boldsymbol{X}$ is full rank.
4. $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = \boldsymbol{0}$ for $t = 1, 2, ..., T$.
5. $\mathbb{E}[\boldsymbol{x}_{it}v_{it}] = \boldsymbol{0}$ for $t = 1, 2, ..., T$.
6. $\mathbb{E}[u_{it}] = \mathbb{E}[\eta_i] + \mathbb{E}[v_{it}] = 0$.

Recall that the linear panel model is:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

where $\boldsymbol{u} = \boldsymbol{\eta} + \boldsymbol{v}$. Under these assumptions for consistency of pooled OLS, we have that:

$$\mathbb{E}[\boldsymbol{x}_{it}u_{it}] = \mathbb{E}[\boldsymbol{x}_{it}\eta_i] + \mathbb{E}[\boldsymbol{x}_{it}v_{it}] = 0.$$

This is the **orthogonality condition** that we have previously seen in linear regression that we needed in order to establish consistency.

Proposition 581 (Consistency of Pooled OLS Estimator)
Suppose we have the linear panel model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{497}$$

where $\boldsymbol{u} = \boldsymbol{\eta} + \boldsymbol{v}$. Then, the assumptions for consistency of pooled OLS, we have that the pooled OLS estimator is a consistent estimator:

$$\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{P} \boldsymbol{\beta} \tag{498}$$

as $N \to \infty$ with T fixed. Joint (N, T) asymptotics require additional time-series dependence conditions.

We have shown consistency and now want to look at inference. If we wanted to compute standard errors for inference for the pooled OLS, we notice that we have some minor issues as the error terms $u_{it} = \eta_i + v_{it}$ are **serially correlated**.

Proposition 582 (Serial Correlation of Error Terms in Pooled OLS)
Under the standard error-components assumptions with $Var(\eta_i) > 0$ and serially uncorrelated v_{it}, the composite errors $u_{it} = \eta_i + v_{it}$ are **serially correlated**. As a result, ordinary pooled OLS is generally **not efficient.**

Proof. If v_{it} is serially uncorrelated and independent of η_i, the errors u_{it} are correlated through the time-invariant η_i

term as:

$$u_{i,t} = \eta_i + v_{i,t}$$
$$u_{i,t-1} = \eta_i + v_{i,t-1}$$

η_i appears in both terms, and $Cov(u_{it}, u_{is}) = Var(\eta_i) > 0$ for $t \neq s$. This violates the spherical-error condition used for OLS efficiency, so pooled OLS is generally not efficient. □

As a result of this serial correlation, we need to use **cluster-robust standard errors and test statistics** whereby we treat all T observations for individual i as a cluster to help us rectify this source of serial correlation.

13.3 Fixed Effects Model

First, we recall from last section the uncorrelated individual effects assumption we made:

$$\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = \mathbf{0}$$

for $t = 1, 2, ..., T$. This is in fact a **strong and restrictive assumption**. We can in fact relax this assumption whereby we now allow some explanatory variables $\boldsymbol{x}_{it}$ to be correlated with the individual-specific effects η_i.

We first describe the setting we are in.

Assumption 583 (Short T Panel)
The panel set has a small number of time observations T.

Now, we state the new relaxed assumption we will make.

Assumption 584 (Correlated Individual Effects/Fixed Effects)
Some of the K explanatory variables $\boldsymbol{x}_{it}$ are correlated with the individual-specific effects η_i:

$$\mathbb{E}[\boldsymbol{x}_{it}\eta_i] \neq \mathbf{0} \tag{499}$$

for $t = 1, 2, ..., T$.

For the other error term v_{it}, we keep the same assumption we made in pooled OLS.

Assumption 585 (Contemporaneous Exogeneity)
The K explanatory variables are uncorrelated with the time varying individual error term in the same time period:

$$\mathbb{E}[\boldsymbol{x}_{it}v_{it}] = \mathbf{0} \tag{500}$$

for $t = 1, 2, ..., T$.

Remark 586. *This rules out contemporaneous correlation between $\boldsymbol{x}_{it}$ and v_{it}.*

Recalling the section on instrumental variables, this η_i can be thought of as containing omitted variables and hence capturing the omitted variables in our model. As a result of this condition, if we go back to the linear panel model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

where $\boldsymbol{u} = \boldsymbol{\eta} + \boldsymbol{v}$, we can think of $\boldsymbol{\eta}$ as capturing omitted variables. To not have to worry about adding intercepts, we assume that the error components have expected values of zero:

$$\mathbb{E}[\eta_i] = \mathbb{E}[v_{it}] = 0.$$

Under this new assumption of correlated individual effects, we no longer have orthogonality assumption between $\boldsymbol{x}_{it}$ and the error terms:

$$\mathbb{E}[\boldsymbol{x}_{it}u_{it}] = \mathbb{E}[\boldsymbol{x}_{it}\eta_i] + \mathbb{E}[\boldsymbol{x}_{it}v_{it}] = \mathbb{E}[\boldsymbol{x}_{it}\eta_i] + 0 \neq \mathbf{0}.$$

Therefore, we no longer have the orthogonality condition $\mathbb{E}[\boldsymbol{x}_{it}u_{it}] = 0$ needed for pooled OLS to be consistent. As a result of this, we need to think of other ways to estimate $\boldsymbol{\beta}$. What we shall do is to transform the original model

using the fact that we now have repeated observations over time. This will eliminate the time-invariant components from the error term of the transformed equation. This will then allow us to derive consistent estimators of $\boldsymbol{\beta}$ by using OLS on the transformed model. The issue though from this approach is that even though we eliminated η_i from the error term, we now require new exogeneity assumptions instead of the predetermined assumption which is what we used in the uncorrelated individual effects model, which mathematically was $\mathbb{E}[\boldsymbol{x}_{it} v_{it}] = \mathbf{0}$.

We will now look at a few different estimators for $\boldsymbol{\beta}$ under the assumption of correlated individual effects.

13.3.1 Within Groups Estimator

We look at the first transformation we can use. The idea is to take the average of each variable over its time component and demean the original variable by its time average. Intuitively, as the nuisance time-invariant error term η_i does not vary over time, this transformation will eliminate it. We now formally state this transformation.

Definition 587 (Within Transformation)
The within transformation is given by:

$$\tilde{y}_{it} = y_{it} - \overline{y}_i \tag{501}$$

$$\tilde{\boldsymbol{x}}_{it} = \boldsymbol{x}_{it} - \overline{\boldsymbol{x}}_i \tag{502}$$

$$\tilde{\eta}_i = \eta_i - \overline{\eta}_i \tag{503}$$

$$\tilde{v}_{it} = v_{it} - \overline{v}_i \tag{504}$$

where the means are defined as:

$$\overline{y}_i = \frac{1}{T}\sum_{s=1}^{T} y_{is}$$

$$\overline{\boldsymbol{x}}_i = \frac{1}{T}\sum_{s=1}^{T} \boldsymbol{x}_{is}$$

$$\overline{\eta}_i = \frac{1}{T}\sum_{s=1}^{T} \eta_i$$

$$\overline{v}_i = \frac{1}{T}\sum_{s=1}^{T} v_{is}.$$

Under this within transformation, something interesting happens to the error terms.

Proposition 588 (Within Transformation of Error Terms)
The within transformation of the time-invariant error term η_i is zero:

$$\tilde{\eta}_i = 0. \tag{505}$$

Furthermore, for individual i's time-varying error term v_{it}, its within transformation depends on all of its time-varying error terms:

$$\tilde{v}_{it} = v_{it} - \frac{1}{T}\left(v_{i1} + v_{i2} + \ldots + v_{iT}\right) \tag{506}$$

whereby $v_{i1}, \ldots, v_{iT}$ are all the time varying error terms for individual i.

Proof. For the time-invariant error term's within transformation:

$$\tilde{\eta}_i = \eta_i - \frac{1}{T}\sum_{t=1}^{T}\eta_i = \eta_i - \frac{1}{T}\big(T\eta_i\big) = \eta_i - \eta_i = 0.$$

For the time-varying error term's within transformation:

$$\tilde{v}_{it} = v_{it} - \overline{v}_i = v_{it} - \frac{1}{T}\big(v_{i1} + v_{i2} + ... + v_{iT}\big).$$

□

Therefore, we have just shown that the time-invariant error term η_i has been eliminated under the within transformation. This is good news because under our correlated individual effects assumption, we had that:

$$\mathbb{E}[\boldsymbol{x}_{it}\eta_i] \neq \mathbf{0}$$

which prevented us from having consistency in our pooled OLS estimator. We have now successfully removed this error term η_i from our model and therefore don't have to worry about endogeneity between our error term η_i and regressor $\boldsymbol{x}_{it}$.

However, an issue that has arisen now is that $\tilde{v}_{it}$ depends on **all** of its time-varying components $v_{i1}, v_{i2}, ..., v_{iT}$. This matters for our previous assumption of exogeneity:

$$\mathbb{E}[\boldsymbol{x}_{it}v_{it}] = \mathbf{0}$$

will **not** guarantee for us that:

$$\mathbb{E}[\tilde{\boldsymbol{x}}_{it}\tilde{v}_{it}] = \mathbf{0}.$$

To be more explicit on why this is the case, recall that the explicit expressions of the within transformations of the explanatory variable and time-varying error terms are:

$$\tilde{\boldsymbol{x}}_{it} = \boldsymbol{x}_{it} - \frac{1}{T}\big(\boldsymbol{x}_{i1} + \boldsymbol{x}_{i2} + ... + \boldsymbol{x}_{iT}\big)$$
$$\tilde{v}_{it} = v_{it} - \frac{1}{T}\big(v_{i1} + v_{i2} + ... + v_{iT}\big).$$

Therefore, when we try to impose an exogeneity assumption on this:

$$\mathbb{E}[\tilde{x}_{it}\tilde{v}_{it}] = \mathbb{E}\left[\left(\boldsymbol{x}_{it} - \frac{1}{T}\big(\boldsymbol{x}_{i1} + \boldsymbol{x}_{i2} + ... + \boldsymbol{x}_{iT}\big)\right)\left(v_{it} - \frac{1}{T}\big(v_{i1} + v_{i2} + ... + v_{iT}\big)\right)\right]$$

we can see that whilst our previous exogeneity assumption of $\mathbb{E}[\boldsymbol{x}_{it}v_{it}] = 0$ will eliminate some of the expectation of the products, there will be many terms in the above expression where:

$$\mathbb{E}[\boldsymbol{x}_{it}v_{is}] \neq \mathbf{0}$$

for $s \neq t$ which means we do not have exogeneity for OLS. Therefore, we need to impose a new assumption called strict exogeneity.

Assumption 589 (Strict Exogeneity)

All K explanatory variables are uncorrelated with individual time-varying error terms over all periods:

$$\mathbb{E}[\boldsymbol{x}_{it} v_{is}] = \mathbf{0} \tag{507}$$

for all $s, t \in \{1, 2, ..., T\}$.

Strict exogeneity requires that the current values of $\boldsymbol{x}_{it}$ are uncorrelated with past, present, and future values of the time-varying component of the error term v_{is} for the same individual in the original model.

An issue with the assumption of strict exogeneity is that it rules out allowing lagged dependent variables $y_{i,t-1}$ as an explanatory variable in $\boldsymbol{x}_{it}$, which is what we have previously done for autoregressive models in time series. The reason we can't do so is because $y_{i,t-1}$ depends on previous error terms v_{it-1} by definition of the structural equation. Therefore, we will come back later on in the course to learn how to allow for lagged dependent variables as an explanatory variable.

We can now estimate $\boldsymbol{\beta}$.

Definition 590 (Within Groups/Fixed Effects Estimator)

Suppose that our structural model is:

$$\tilde{y}_{it} = \tilde{\boldsymbol{x}}_{it}\boldsymbol{\beta} + \tilde{v}_{it}. \tag{508}$$

where $\tilde{y}_{it}, \tilde{\boldsymbol{x}}_{it}, \tilde{v}_{it}$ are the within-transformations. Then, stack the observations to form the structural equation:

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{v}}. \tag{509}$$

The **within groups estimator** of $\boldsymbol{\beta}$ is given by:

$$\hat{\boldsymbol{\beta}}_{WG} = \left(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}\right)^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{y}}. \tag{510}$$

Remark 591. *The within groups estimator is also known as the* **fixed effects estimator**.

We also have all the necessary orthogonality conditions needed for consistency of the within groups estimator.

Proposition 592 (Consistency of Within Groups Estimator)

Under the assumption of strict exogeneity, the within groups estimator is a consistent estimator:

$$\hat{\boldsymbol{\beta}}_{WG} \xrightarrow{P} \boldsymbol{\beta} \tag{511}$$

as $N \to \infty$ with T fixed.

It is important to note that we have just shown that $\hat{\boldsymbol{\beta}}_{WG}$ is a consistent estimator for $\boldsymbol{\beta}$ even if all explanatory variables $\boldsymbol{x}_{it}$ are correlated with η_i:

$$\mathbb{E}[\boldsymbol{x}_{it}\eta_i] \neq \mathbf{0}$$

which is what we called the correlated individuals effects/fixed effects. We can also derive the asymptotic distribution. First, we need an assumption.

Assumption 593 (Panel Data Homoskedasticity)

Panel data homoskedasticity assumes that the time varying individual error terms v_{it} satisfy:

$$v_{it}|\boldsymbol{X} \sim i.i.d(0, \sigma_v^2) \tag{512}$$

for all observations i and time periods t.

We can now derive the asymptotic distribution of the estimator.

Theorem 594 (Asymptotic Normality of Within Groups Estimator)

Under strict exogeneity and panel data homoskedasticity, the asymptotic distribution of the within groups estimator is:

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{WG} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma_v^2 \boldsymbol{Q}^{-1}), \tag{513}$$

as $N \to \infty$ with T fixed, where $\boldsymbol{Q} = p\lim_{n\to\infty} N^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$. A consistent estimator $\hat{\sigma}_v^2$ can be used in the corresponding estimated covariance matrix.

We will discuss the consistent estimator term $\hat{\sigma}_v^2$ soon.

We note that even if we don't have homoskedasticity, this **does not** affect the consistency of the within groups estimator as consistency only requires strict exogeneity from our earlier arguments.

Furthermore, if panel data homoskedasticity does not hold, we can use heteroskedasticity-consistent and cluster-robust standard errors. As a result, we do not need to worry too much about serial correlation in the error terms.

Recall that in the asymptotic distribution of the within groups estimator, we had a $\hat{\sigma}_v^2$ estimator. We now describe what we can use as a consistent estimator $\hat{\sigma}_v^2$.

Proposition 595 (Consistent Estimator of the Variance)

Under the panel data homoskedasticity assumption, we can estimate the variance of v_{it}, denoted by σ_v^2, by constructing the estimator:

$$\hat{\sigma}_v^2 = \frac{\sum_{i=1}^N \sum_{t=1}^T \hat{\hat{v}}_{it}^2}{N(T-1)-K} = \frac{\hat{\hat{\boldsymbol{v}}}^T\hat{\hat{\boldsymbol{v}}}}{N(T-1)-K} \tag{514}$$

where $\hat{\hat{v}}_{it}$ are the residuals from the within groups estimator:

$$\hat{\hat{v}}_{it} = \tilde{y}_{it} - \tilde{\boldsymbol{x}}_{it}\hat{\boldsymbol{\beta}}_{WG}. \tag{515}$$

Remark 596. *We will later see where the does the N(T - 1) - K degrees of freedom come from when we look at least squares dummy variables estimators.*

13.3.2 Least Squares Dummy Variables Estimator

This is another estimator we can use to estimate $\boldsymbol{\beta}$ under the assumption of fixed effects. If we do not have too many cross-sectional observations n, we can include N dummy variables in order to absorb the error term η_i which we are having issues with. This is a similar idea to what we did for the time dummy variables.

Definition 597 (Dummy Variables)

We define the N dummy variables for the N observations as:

$$I_i^1 = 1 \quad \text{for observations on individual 1 or otherwise } I_i^1 = 0 \tag{516}$$

$$I_i^2 = 1 \quad \text{for observations on individual 2 or otherwise } I_i^2 = 0 \tag{517}$$

$$\ldots \tag{518}$$

$$I_i^N = 1 \quad \text{for observations on individual N or otherwise } I_i^N = 0 \tag{519}$$

With these N dummy variables, we now include them into our structural model and describe our new estimation strategy.

Definition 598 (Least Squares Dummy Variables Estimator)

Suppose that the structural model with dummy variables is now:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \eta_1 I_i^1 + \eta_2 I_i^2 + \ldots + \eta_N I_i^N + v_{it}. \tag{520}$$

Let $\boldsymbol{D}$ collect the individual dummy variables and let $\boldsymbol{M}_D = \boldsymbol{I}_{NT} - \boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{D}^T$. The **Least Squares Dummy Variables Estimator** (LSDV) for the slope coefficients is the OLS estimator in the above structural equation:

$$\hat{\boldsymbol{\beta}}_{LSDV} = (\boldsymbol{X}^T\boldsymbol{M}_D\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}_D\boldsymbol{y}. \tag{521}$$

The LSDV estimator should produce the same coefficients as the fixed effect (FE) estimator.

Theorem 599 (Equivalency of LSDV and WG)

The LSDV estimator and WG estimator are equivalent:

$$\hat{\boldsymbol{\beta}}_{LSDV} = \hat{\boldsymbol{\beta}}_{WG}. \tag{522}$$

When based on the same error-variance estimator and degrees-of-freedom adjustment, the two methods also produce the same conventional standard errors for $\boldsymbol{\beta}$. Their main difference is computational: the within transformation avoids estimating a large set of individual dummy coefficients when N is large.

If we look at the LSDV structural equation, we have our usual K parameters we estimate in $\boldsymbol{\beta}$. We also have an additional N parameters $\eta_1, \ldots, \eta_N$ which we estimate. Therefore, the degrees of freedom is of our estimation is:

$$NT - K - N = N(T-1) - K$$

which is what we saw earlier in Eq. (514) when computing the consistent estimator of the variance under panel data homoskedasticity assumption.

13.3.3 First-Differenced OLS

This is another estimator we can use to estimate $\boldsymbol{\beta}$ under the fixed effects assumptions. We can perform another transformation that is inspired by time series techniques.

Definition 600 (First-Differencing)

The first differencing operator Δ is defined as the operator such that when applied to y_{it} results in:

$$\Delta y_{it} = y_{it} - y_{i,t-1}. \tag{523}$$

Now, we can take our original structural model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\eta} + \boldsymbol{v}$$

and apply our first-difference transformation on it.

Proposition 601 (First-Difference Transformation)

The first-differenced transformed equation of our panel data structural model is:

$$\Delta y_{it} = \Delta \boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta v_{it} \tag{524}$$

for $i = 1, 2, ..., N$ and $t = 2, 3, ..., T$.

Proof. Start off with the panel data structural model:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \eta_i + v_{it}$$

First, we can simply difference y_{it} to get:

$$y_{it} - y_{i,t-1} = \Delta y_{it}$$

and the same applies for $\boldsymbol{x}_{it}, v_{it}$:

$$\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1} = \Delta \boldsymbol{x}_{it}.$$

$$v_{it} - v_{i,t-1} = \Delta v_{it}.$$

However, notice that for the time-invariant parameter η_i, differencing it removes it:

$$\eta_i - \eta_i = 0.$$

□

Notice that as a result of first-differencing, we start from $t = 2$ as we do not observe Δy_{i1} and Δx_{i1}. We can now state our estimator under this transformation.

Definition 602 (First-Differenced OLS)

Suppose that the structural model under first-differencing is:

$$\Delta y_{it} = \Delta \boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta v_{it} \tag{525}$$

for $i = 1, 2, ..., N$ and $t = 2, 3, ..., T$.

Then, the first-differenced OLS estimator is:

$$\hat{\boldsymbol{\beta}}_{FDOLS} = \left[(\Delta \boldsymbol{X})^T \Delta \boldsymbol{X}\right]^{-1} (\Delta \boldsymbol{X})^T (\Delta \boldsymbol{y}). \tag{526}$$

If we look at the first-differenced structural equation:

$$\Delta y_{it} = \Delta \boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta v_{it}$$

we can see that in order for usual orthogonality condition for consistency of the OLS estimator, we require that:

$$\mathbb{E}\left[\Delta \boldsymbol{x}_{it}\Delta v_{it}\right] = \mathbf{0}.$$

Assumption 603 (First Differenced Exogeneity)
The first differenced of $\boldsymbol{x}_{it}$ and the first differenced time varying individual error v_{it} are uncorrelated:

$$\mathbb{E}[\Delta \boldsymbol{x}_{it}\Delta v_{it}] = \mathbf{0}. \tag{527}$$

We can now state our result for consistency of this estimator.

Proposition 604 (Consistency of First-Differenced OLS)
Under the assumption of first differenced exogeneity, the first-differenced OLS estimator is consistent:

$$\hat{\boldsymbol{\beta}}_{FDOLS} \xrightarrow{P} \boldsymbol{\beta} \tag{528}$$

as $N \to \infty$ with T fixed.

First-differenced exogeneity restricts a combination of contemporaneous and adjacent-period cross-moments. By itself, it does not set each of those cross-moments to zero separately. We can nevertheless establish one clear relationship with strict exogeneity.

Proposition 605 (Strict Exogeneity Implies First Differenced Exogeneity)
Strict exogeneity

$$\mathbb{E}[\boldsymbol{x}_{it}v_{is}] = \mathbf{0} \tag{529}$$

for $t, s \in \{1, ..., T\}$ implies first differenced exogeneity:

$$\mathbb{E}[\Delta \boldsymbol{x}_{it}\Delta v_{it}] = \mathbf{0} \tag{530}$$

for $t = 1, ..., T$.

Proof. We directly show this by definition of first differenced exogeneity:

$$\begin{aligned}
\mathbb{E}[\Delta \boldsymbol{x}_{it}\Delta v_{it}] &= \mathbb{E}[(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})(v_{it} - v_{i,t-1})] \\
&= \mathbb{E}[\boldsymbol{x}_{it}v_{it}] - \mathbb{E}[\boldsymbol{x}_{it}v_{i,t-1}] - \mathbb{E}[\boldsymbol{x}_{i,t-1}v_{it}] + \mathbb{E}[\boldsymbol{x}_{i,t-1}v_{i,t-1}] \\
&= \mathbf{0} + \mathbf{0} + \mathbf{0} + \mathbf{0}
\end{aligned}$$

where the last equation holds by strict exogeneity.

□

Remark 606. *First-differenced exogeneity does not generally imply contemporaneous exogeneity $\mathbb{E}[\boldsymbol{x}_{it}v_{it}] = \mathbf{0}$, nor does contemporaneous exogeneity generally imply first-differenced exogeneity. Additional restrictions are needed to order these two assumptions. Strict exogeneity implies both.*

Therefore, this means that under the assumptions of strict exogeneity, then the first differenced OLS estimator is consistent!

We can also state what the standard errors are for FDOLS.

Theorem 607 (Standard Errors for First Differenced OLS Estimator)
Classical standard errors are appropriate for $\hat{\boldsymbol{\beta}}_{FDOLS}$ if the time-varying error component v_{it} is a random walk:

$$v_{it} = v_{i,t-1} + \epsilon_{it} \tag{531}$$

with i.i.d innovations so that the first differenced of the error terms:

$$\Delta v_{it} = \epsilon_{it} \sim i.i.d(0, \sigma_\epsilon^2) \tag{532}$$

for all individuals i and time t.

Again, consistency for $\hat{\boldsymbol{\beta}}_{FDOLS}$ **does not** rely on v_{it} being uncorrelated but instead relies on first differenced exogeneity.

13.3.4 Efficiency Comparison of Fixed Effects Estimators

We restrict our attention to a special case of endogeneity.

Assumption 608 (Correlation with Time-Invariant Error Terms)
All K explanatory variables are correlated with η_i

Recall that efficient estimators are estimators that achieves the minimum possible variance as an estimator. We now discuss the conditions for efficiency for the estimators we have seen so far.

Theorem 609 (Efficiency of Within Groups Estimator)
Suppose that we have the classical assumptions:

$$v_{it}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT} \sim i.i.d(0, \sigma_v^2) \tag{533}$$

over $i = 1, ..., N$ and $t = 1, 2, ..., T$. Then, $\hat{\boldsymbol{\beta}}_{WG}$ is an efficient estimator of $\boldsymbol{\beta}$. Furthermore, under normality assumption:

$$v_{it}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT} \sim i.i.d\mathcal{N}(0, \sigma_v^2) \tag{534}$$

then $\hat{\boldsymbol{\beta}}_{WG}$ coincides with the maximum likelihood estimator of $\boldsymbol{\beta}$ and achieves the Cramer-Rao lower bound.

Theorem 610 (Efficiency of First Differenced Estimator)
Suppose that we have that

$$v_{it} = v_{i,t-1} + \epsilon_{it} \tag{535}$$

is a random walk and the innovations

$$\epsilon_{it}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT} \sim i.i.d(0, \sigma_\epsilon^2) \tag{536}$$

over $i = 1, ..., N$ and $t = 1, 2, ..., T$. Then, $\hat{\boldsymbol{\beta}}_{FDOLS}$ is an efficient estimator of $\boldsymbol{\beta}$.

We now compare these two estimators and describe when is one more efficient than the other.

Proposition 611 (Comparison of the Two Fixed Effects Estimators)
Suppose that v_{it} is serially uncorrelated, then $\hat{\boldsymbol{\beta}}_{WG}$ is more efficient than $\hat{\boldsymbol{\beta}}_{FDOLS}$.
Suppose that v_{it} is a random walk, then $\hat{\boldsymbol{\beta}}_{FDOLS}$ is more efficient than $\hat{\boldsymbol{\beta}}_{WG}$.

If we are not worried about serial correlation in v_{it} and there is smooth continuous time series variation in y_{it} and $\boldsymbol{x}_{it}$, then either estimator is fine.

Remark 612. *These two estimators coincide when we only have $T = 2$ time periods in our panel dataset.*

A big issue to worry about for first-differenced OLS is that it may not do well when you have step functions in your explanatory variables.

13.3.5 Large T Asymptotics of Fixed Effects Estimators

In this section, we explain how strict exogeneity can be relaxed under large-T asymptotics, subject to suitable weak-dependence and moment conditions. Recall strict exogeneity:

$$\mathbb{E}[\boldsymbol{x}_{it} v_{is}] = \boldsymbol{0}$$

for $t, s \in \{1, ..., T\}$ for the WG estimator but not for the FDOLS estimator.

First, recall that for the within group transformation:

$$\begin{aligned}
\tilde{y}_{it} &= y_{it} - \frac{1}{T}\sum_{s=1}^{T} y_{is} \\
\tilde{\boldsymbol{x}}_{it} &= \boldsymbol{x}_{it} - \frac{1}{T}\sum_{s=1}^{T} \boldsymbol{x}_{is} \\
\tilde{v}_{it} &= v_{it} - \frac{1}{T}\sum_{s=1}^{T} v_{is}.
\end{aligned}$$

As $T \to \infty$, the time averages do not literally disappear. Rather, under suitable weak-dependence and ergodicity conditions, the endogeneity induced by demeaning can become asymptotically negligible. A standard weaker condition is **predeterminedness**:

$$\mathbb{E}[\boldsymbol{x}_{it} v_{is}] = \boldsymbol{0} \quad \text{for } s \geq t,$$

which allows $\boldsymbol{x}_{it}$ to be correlated with past errors. Under the additional regularity conditions, this can guarantee consistency of the WG estimator:

$$\hat{\boldsymbol{\beta}}_{WG} \xrightarrow{P} \boldsymbol{\beta}$$

as $T \to \infty$.

However, for the FDOLS estimator, we still require the first differenced exogeneity assumption:

$$\mathbb{E}[\Delta \boldsymbol{x}_{it} \Delta v_{it}] = 0$$

for consistency of the FDOLS estimator $\hat{\boldsymbol{\beta}}_{FDOLS} \xrightarrow{P} \boldsymbol{\beta}$ as $T \to \infty$.

Because the large-T consistency result for WG can use weaker sequential-exogeneity conditions, this comparison favours WG on robustness grounds; efficiency and the serial-correlation structure may still favour FDOLS in particular

applications.

Proposition 613 (WG and FDOLS Estimators in Large T)
Under suitable weak-dependence, moment, and identification conditions, the WG estimator is consistent under predeterminedness in large-T samples.

The FDOLS still requires the first-differenced exogeneity assumption.

13.4 Random Effects Model

We now go back to the scenario of small T panels whilst assuming random effects and strictly exogenous.

Assumption 614 (Strict Exogeneity)

All K explanatory variables are uncorrelated with individual time-varying error terms over all periods:

$$\mathbb{E}[\boldsymbol{x}_{it} v_{is}] = \mathbf{0} \tag{537}$$

for all $s, t \in \{1, 2, ..., T\}$.

Assumption 615 (Uncorrelated Individual Effects/Random Effects)

All K explanatory variables $\boldsymbol{x}_{it}$ is uncorrelated with all individual-specific effects η_i:

$$\mathbb{E}[\boldsymbol{x}_{it} \eta_i] = \mathbf{0} \tag{538}$$

for $t = 1, 2, ..., T$.

Notice that this is a very nice setting to be in! We have the strongest form of exogeneity and we have random effects. So what is the issue then? The issue is that the previous estimators we have seen, the within-groups and the first differenced OLS is **not** efficient if we have random effects:

$$\mathbb{E}[\boldsymbol{x}_{it} \eta_i] = \mathbf{0}.$$

Recall that these two estimators were constructed under the assumption of fixed effects $\mathbb{E}[\boldsymbol{x}_{it} \eta_i] \neq \mathbf{0}$ and as a result, we had to perform transformation operations on the levels of the series $(y_{it}, \boldsymbol{x}_{it})$ to get rid of η_i. This then allowed us in those scenarios to estimate $\boldsymbol{\beta}$ under fixed effects. However, if you think about it, if we are under the assumption of random effects $\mathbb{E}[\boldsymbol{x}_{it} \eta_i] = \mathbf{0}$, then there is no need to perform such a transformation to get rid of η_i and in fact, we are losing information as we could've just used the levels. We do not have **efficient estimators**. Therefore, we figure out how to construct efficient estimators under this assumption of random effects.

We could also use pooled OLS since contemporaneous exogeneity $\mathbb{E}[\boldsymbol{x}_{it} v_{it}] = \mathbf{0}$ and random effects $\mathbb{E}[\boldsymbol{x}_{it} \eta_i] = \mathbf{0}$ both hold (strict exogeneity implies contemporaneous exogeneity). However, pooled OLS is not efficient under the standard error-components covariance structure because the errors $u_{it} = \eta_i + v_{it}$ are serially correlated through the time-invariant error η_i.

To this end, we will now look at another estimator called the between groups estimator. However, we will later see that this estimator is not efficient either but once we established this estimator, we will then be able to find an efficient estimator.

13.4.1 Between Groups Estimator

We continue assuming strict exogeneity and random effects. We now define the between groups transformation.

Definition 616 (Between Transformation)

The between transformation is given by taking the mean of observations:

$$\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} \tag{539}$$

$$\overline{\boldsymbol{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_{it} \tag{540}$$

$$\overline{v}_i = \frac{1}{T}\sum_{t=1}^{T} v_{it} \tag{541}$$

$$\overline{\eta}_i = \frac{1}{T}\sum_{t=1}^{T} \eta_i = \eta_i. \tag{542}$$

The between transformation structural equation is:

$$\overline{y}_i = \overline{\boldsymbol{x}}_i\boldsymbol{\beta} + (\overline{v}_i + \eta_i) \tag{543}$$

for i = 1,2, ..., N.

Remark 617. *Again notice that the time-invariant error term has the mean:*

$$\overline{\eta}_i = \frac{1}{T}\sum_{t=1}^{T} \eta_i = \frac{1}{T} T \eta_i = \eta_i.$$

In contrast, the individual specific time-varying error term has the mean:

$$\overline{v}_i = \frac{1}{T}\sum_{t=1}^{T} v_{it} = \frac{1}{T}\left(v_{i1} + v_{i2} + ... + v_{iT}\right).$$

Notice the similarity to the within groups transformation which demeaned the variables:

$$\tilde{y}_{it} = y_{it} - \overline{y}_i = y_{it} - \frac{1}{T}\sum_{s=1}^{T} y_{is}.$$

These transformations are in fact orthogonal complements to one another. We can now define the between groups estimator.

Definition 618 (Between Groups Estimator)

Suppose that the between groups structural equation is:

$$\overline{y}_i = \overline{x}_i\boldsymbol{\beta} + (\overline{v}_i + \eta_i) \tag{544}$$

for i = 1,2, ..., N.

The **between groups estimator** is the OLS estimator of the between groups structural equation:

$$\hat{\boldsymbol{\beta}}_{BG} = (\overline{\boldsymbol{X}}^T\overline{\boldsymbol{X}})^{-1}(\overline{\boldsymbol{X}})^T\overline{\boldsymbol{y}}. \tag{545}$$

Notice that in our structural equation:

$$\overline{y}_i = \overline{x}_i\boldsymbol{\beta} + (\overline{v}_i + \eta_i)$$

we have $\overline{\boldsymbol{x}}_i = \frac{1}{T}\left[\boldsymbol{x}_{i1} + ... + \boldsymbol{x}_{iT}\right]$ and $\overline{v}_i = \frac{1}{T}\left[v_{i1} + ... v_{iT}\right]$. Therefore, in order for our usual orthogonality assumption in OLS between the regressor and error term, we require strict exogeneity $\mathbb{E}[\boldsymbol{x}_{it} v_{is}] = \boldsymbol{0}$ for all $s, t \in \{1, ..., T\}$. Furthermore, notice that we did not get rid of the time-invariant variable η_i and therefore we need to assume random effects so that $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = \boldsymbol{0}$. With these assumptions, we have consistency of this estimator.

Proposition 619 (Consistency of Between-Groups OLS)
Suppose that we have strict exogeneity and random effects. Then the between-groups OLS estimator is a consistent estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{BG} \xrightarrow{P} \boldsymbol{\beta} \tag{546}$$

as $N \to \infty$ with T fixed.

However, an issue is that even under these assumptions and this transformation, $\hat{\boldsymbol{\beta}}_{BG}$ is **not efficient**. The reason for this is that we are losing information when we average the observations over the T time periods which discards information in the time series variation. However, we know that these time series variations can be useful for estimating $\boldsymbol{\beta}$ since this is what we used as the source of our transformations when we constructed within groups and first-differenced OLS estimators.

So even though it seems like we are back at where we started, we shall soon see that we can use the between groups estimator in our process of constructing an efficient estimator when we have random effects and strict exogeneity.

13.4.2 Prereqs For Random Effects Generalised Least Squares Estimator

We have seen that the between-groups estimator is a consistent estimator but not efficient. The current issue is that the time-invariant error η_i is causing a lot of complications. Recall it was this term that caused serial correlation in the setting of pooled OLS. Also recall that it is because of this term that we applied the WG and FDOLS transformations to remove it; the BG transformation retains it but averages away within-individual variation. Therefore, in this section, we shall estimate the variance of this error term and apply what we saw in generalised least squares to deal with this inefficiency.

We will use both the between-groups estimator and the within-groups estimator to construct the random effect generalised least squares estimator. The big idea is that from the between groups estimator and the within groups estimator, we have the structural equations:

$$\begin{aligned}
\overline{y}_i &= \overline{\boldsymbol{x}}_i\boldsymbol{\beta} + (\overline{v}_i + \eta_i) \\
\tilde{y}_i &= \tilde{\boldsymbol{x}}_i\boldsymbol{\beta} + \tilde{v}_i
\end{aligned}$$

where the first equation is the between groups structural equation and the second equation is the within groups structural equation. Now, we can see that in the between groups equation:

$$\overline{y}_i = \overline{\boldsymbol{x}}_i\boldsymbol{\beta} + (\overline{v}_i + \eta_i)$$

we can estimate the variance of the error term $\overline{v}_i + \eta_i$, which we denote by $\sigma^2_{\overline{u}_i}$ where we define the error term $\overline{u}_i \equiv \overline{v}_i + \eta_i$.

From the within groups equation:

$$\tilde{y}_i = \tilde{\boldsymbol{x}}_i\boldsymbol{\beta} + \tilde{v}_i$$

we can estimate the variance of $\tilde{v}_i$, which we denote by σ^2_v.

From this, if we have $\sigma^2_{\overline{u}_i}$ from the between groups equation and σ^2_v from the within groups equation, we can then back out what the variance of η_i is, which we denote by σ^2_η. We can then construct the random effects GLS estimator from these estimated variances. However, in order to back out the variance of η_i, we require an assumption about independence of the error terms.

Assumption 620 (I.I.D Error Terms)

The two errors components η_i and v_{it} are both i.i.d and uncorrelated with each other:

$$\mathbb{E}[\eta_i v_{it}] = 0 \tag{547}$$

for $t = 1, 2, ..., T$.

Now, our goal for this section is to estimate the variance of η_i, which we denote by σ^2_η. This will take a few steps so don't lose sight of the big picture.

Step One: Using Between Groups Estimator

The first step is to retrieve the residuals from between groups estimator:

$$\hat{\overline{u}}_i = \widehat{(\eta_i + \overline{v}_i)} = \overline{y}_i - \overline{\boldsymbol{x}}_i \hat{\boldsymbol{\beta}}_{BG} \tag{548}$$

where $\hat{\boldsymbol{\beta}}_{BG}$ is the between groups estimator. We can then use these residuals $\hat{\overline{u}}_i$ to construct the estimator

$$\hat{\sigma}^2_{\overline{u}} = \frac{\sum_{i=1}^N \widehat{(\eta_i + \overline{v}_i)}^2}{N - K} = \frac{\hat{\overline{\boldsymbol{u}}}^T \hat{\overline{\boldsymbol{u}}}}{N - K} \tag{549}$$

$\hat{\sigma}^2_{\overline{u}}$ estimates the variance $\eta_i + \overline{v}_i$, whereby the variance has the form

$$\sigma^2_\eta + (\frac{1}{T})\sigma^2_v.$$

See slide 41 of lecture 2 Panel data for why it is this form. This means we can write:

$$\hat{\sigma}^2_{\overline{u}} = \frac{\hat{\overline{\boldsymbol{u}}}^T \hat{\overline{\boldsymbol{u}}}}{N - K} = \hat{\sigma_\eta}^2 + (\frac{1}{T})\hat{\sigma}^2_v$$

or just

$$\hat{\sigma}^2_{\overline{u}} = \hat{\sigma}^2_\eta + \left(\frac{1}{T}\right)\hat{\sigma}^2_v. \tag{550}$$

We will use this estimator $\hat{\sigma}^2_{\overline{u}}$ later on when constructing the random effects GLS estimator.

Step Two: Using Within Groups Estimator

We now want to construct an estimator of the variance of v_{it}, denoted by σ^2_v. To do so, we need the residuals from within groups estimator:

$$\hat{\tilde{v}}_{it} = \tilde{y}_{it} - \tilde{\boldsymbol{x}}_{it} \hat{\boldsymbol{\beta}}_{WG} \tag{551}$$

where $\hat{\boldsymbol{\beta}}_{WG}$ is the within groups estimator. We have already discussed how to then construct an estimator of σ^2_v using the residuals $\hat{\tilde{v}}_{it}$ when discussing the within groups estimator in Eq. (514) which we recall:

$$\hat{\sigma}^2_v = \frac{\sum_{i=1}^N \sum_{t=1}^T \hat{\tilde{v}}^2_{it}}{N(T-1) - K} = \frac{\hat{\tilde{\boldsymbol{v}}}^T \hat{\tilde{\boldsymbol{v}}}}{N(T-1) - K}.$$

Step Three: Combining Two Variance Estimators

To recap, we first used the between groups estimator to construct the estimator $\hat{\sigma}^2_{\overline{u}}$ for $\sigma^2_{\overline{u}}$ where $\overline{u} = \eta_i + \overline{v}_i$. We then used the within groups estimator to construct the estimator $\hat{\sigma}^2_v$ for σ^2_v. Recall that the goal of all this was to estimate σ^2_η. We can now use these two variance estimators to compute the variance estimator for σ^2_η by computing:

$$\hat{\sigma}^2_\eta = \hat{\sigma}^2_{\overline{u}} - \left(\frac{1}{T}\right)\hat{\sigma}^2_v. \tag{552}$$

We also know that this estimator is in fact a consistent estimator.

Proposition 621 (Consistency of Time-Invariant Error Variance)
The estimator $\hat{\sigma}^2_\eta$ is a consistent estimator of the variance σ^2_η.

We have now finally achieved our goal of constructing the variance estimator $\hat{\sigma}^2_\eta$. We can now use this estimator in our ultimate goal of constructing an estimator of $\boldsymbol{\beta}$ that is efficient in the setting of strictly exogenous and random effects (uncorrelated individual effects). To do so, we shall use techniques that we saw in generalised least squares.

13.4.3 Random Effects Generalised Least Squares Estimator

Now, we shall discuss how to construct an efficient estimator in the setting of strictly exogenous variables and uncorrelated individual effects.

First, recall that the matrix $\boldsymbol{X}_i$ is a $T \times K$ matrix where each row is $\boldsymbol{x}_{it}$ (k explanatory variables for observation i at time t). Then, we can stack the $\boldsymbol{X}_i$ matrices vertically over individuals to get $\boldsymbol{X}$ which is a $NT \times K$ matrix.

Now, we state important assumptions that we will need (and have to memorise) for our upcoming estimator to be efficient.

Assumption 622 (Classical Assumptions)
The classical assumptions are:

1. $\eta_i | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT} \sim i.i.d(0, \sigma^2_\eta)$ over $i = 1, ..., N$
2. $v_{it} | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT} \sim i.i.d(0, \sigma^2_v)$ over $i = 1, ..., N$ and $t = 1, \ldots, T$
3. η_i and v_{it} are independent conditional on $\boldsymbol{X}_i$
4. $\mathbb{E}[\eta_i | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}] = 0$
5. $\mathbb{E}[v_{it} | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}] = 0$ for each $t \in \{1, 2, ..., T\}$

Now, if we look carefully at the last two assumptions in particular:

1. $\mathbb{E}[\eta_i | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}] = 0$
2. $\mathbb{E}[v_{it} | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}] = 0$ for each $t \in \{1, 2, ..., T\}$

we can see that they are in fact stricter conditions than uncorrelated individuals effects and strictly exogenous covariates.

Proposition 623 (Strengthened Uncorrelated Individual Effects)

The classical assumption:

$$\mathbb{E}[\eta_i|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}] = 0 \tag{553}$$

implies uncorrelated individual effects (random effects):

$$\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = 0. \tag{554}$$

Proof. Apply LIE:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_{it}\eta_i] &= \mathbb{E}[\mathbb{E}[\boldsymbol{x}_{it}\eta_i|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}]] \\
&= \mathbb{E}[\boldsymbol{x}_{it}\mathbb{E}[\eta_i|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}]] \\
&= \mathbb{E}[\boldsymbol{x}_{it} \cdot 0] \\
&= 0.
\end{aligned}$$

□

Proposition 624 (Strengthened Strictly Exogenous Covariates)

The classical assumption:

$$\mathbb{E}[v_{it}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}] = 0 \tag{555}$$

for each $t \in \{1, 2, ..., T\}$ implies strictly exogenous covariates

$$\mathbb{E}[\boldsymbol{x}_{it}v_{is}] = 0 \tag{556}$$

for each $t, s \in \{1, 2, ..., T\}$.

Proof. Apply LIE:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_{it}v_{is}] &= \mathbb{E}[\mathbb{E}[\boldsymbol{x}_{it}v_{is}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}]] \\
&= \mathbb{E}[\boldsymbol{x}_{it}\mathbb{E}[v_{is}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}]] \\
&= \mathbb{E}[\boldsymbol{x}_{it} \cdot 0] \\
&= 0.
\end{aligned}$$

□

Now, under the classical assumptions, we look at the linear regression model where we stack the T time periods for each individual i to get:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{u}_i$$

for $i = 1, ..., N$ with $\boldsymbol{u}_i = \eta_i\boldsymbol{j}_T + \boldsymbol{v}_i$. Here, we have that just like in the OLS setting:

$$\mathbb{E}[\boldsymbol{y}_i|\boldsymbol{X}_i] = \boldsymbol{X}_i\boldsymbol{\beta}$$

and by cross-sectional independence:

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}.$$

In this linear regression model, the efficient estimator of $\boldsymbol{\beta}$ is known as the **Random Effects Generalised Least Squares (GLS)** estimator, which takes into account for the serial correlation $u_{it} = \eta_i + v_{it}$ present which made our pooled OLS inefficient.

The final step before we need to do before constructing our random effects GLS estimator is to state under all these assumptions, what does the variance covariance matrix of the error terms $\boldsymbol{u}$, which is the sum of η_i and $\boldsymbol{v}_i$ errors, $\Omega = \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^T|\boldsymbol{X}]$ look like.

Proposition 625 (Random Effects GLS Variance-Covariance Matrix of Errors)
The variance-covariance matrix of errors associated to the random effects GLS estimator is given by:

$$\Omega \equiv \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^T|\boldsymbol{X}] = \begin{bmatrix} \Omega_i & 0 & \dots & 0 \\ 0 & \Omega_i & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Omega_i \end{bmatrix} \tag{557}$$

where Ω is a $NT \times NT$ matrix where each entry Ω_i is a $T \times T$ matrix.

Proof. There are a few steps to this proof. We will have to make repeated use of the classical assumptions. First, we can show that:

$$\begin{aligned} \mathbb{E}[u_{it}|\boldsymbol{X}_i] &= \mathbb{E}[\eta_i + v_{it}|\boldsymbol{X}_i] \\ &= \mathbb{E}[\eta_i|\boldsymbol{X}_i] + \mathbb{E}[v_{it}|\boldsymbol{X}_i] \\ &= 0 + 0 = 0. \end{aligned}$$

We can then compute the variance

$$\begin{aligned} Var[u_{it}|\boldsymbol{X}_i] &= \mathbb{E}[u_{it}^2|\boldsymbol{X}_i] - \mathbb{E}[u_{it}|\boldsymbol{X}_i]^2 \\ &= \mathbb{E}[u_{it}^2|\boldsymbol{X}_i] - 0 \\ &= \mathbb{E}[(\eta_i + v_{it})^2|\boldsymbol{X}_i] \\ &= \mathbb{E}[\eta_i^2 + 2\eta v_{it} + v_{it}^2|\boldsymbol{X}_i] \\ &= \mathbb{E}[\eta_i^2|\boldsymbol{X}_i] + 2\mathbb{E}[\eta_i v_{it}|\boldsymbol{X}_i] + \mathbb{E}[v_{it}^2|\boldsymbol{X}_i] \\ &= \sigma_\eta^2 + 0 + \sigma_v^2 \\ &= \sigma_\eta^2 + \sigma_v^2. \end{aligned}$$

We then compute the covariance:

$$\begin{aligned} \mathbb{E}[u_{it}u_{is}|\boldsymbol{X}_i] &= \mathbb{E}[(\eta_i + v_{it})(\eta_i + v_{is})|\boldsymbol{X}_i] \\ &= \mathbb{E}[\eta_i^2 + \eta_i v_{is} + \eta_i v_{it} + v_{it}v_{is}|\boldsymbol{X}_i] \\ &= \mathbb{E}[\eta_i^2|\boldsymbol{X}_i] + \mathbb{E}[\eta_i v_{is}|\boldsymbol{X}_i] + \mathbb{E}[\eta_i v_{it}|\boldsymbol{X}_i] + \mathbb{E}[v_{it}v_{is}|\boldsymbol{X}_i] \\ &= \sigma_\eta^2 + 0 + 0 + 0 \\ &= \sigma_\eta^2. \end{aligned}$$

Now, we can stack the error terms u_{it} into the vector $\boldsymbol{u}_i$ which is a $T \times 1$ vector:

$$\boldsymbol{u}_i = \eta_i \boldsymbol{j}_T + \boldsymbol{v}_i.$$

So then variance-covariance matrix of this error term $\boldsymbol{u}_i$ for an individual is:

$$\mathbb{E}[\boldsymbol{u}_i \boldsymbol{u}_i^T | \boldsymbol{X}_i] = \begin{bmatrix} Var[u_{i1}|\boldsymbol{X}_i] & Cov[u_{i1}, u_{i2}|\boldsymbol{X}_i] & \dots & Cov[u_{i1}, u_{iT}|\boldsymbol{X}_i] \\ Cov[u_{i2}, u_{i1}|\boldsymbol{X}_i] & Var[u_{i2}|\boldsymbol{X}_i] & \dots & Cov[u_{i2}, u_{iT}|\boldsymbol{X}_i] \\ \dots & \dots & \dots & \dots \\ Cov[u_{iT}, u_{i1}|\boldsymbol{X}_i] & Cov[u_{iT}, u_{i2}|\boldsymbol{X}_i] & \dots & Var[u_{iT}|\boldsymbol{X}_i] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_\eta^2 + \sigma_v^2 & \sigma_\eta^2 & \dots & \sigma_\eta^2 \\ \sigma_\eta^2 & \sigma_\eta^2 + \sigma_v^2 & \dots & \sigma_\eta^2 \\ \dots & \dots & \dots & \dots \\ \sigma_\eta^2 & \sigma_\eta^2 & \dots & \sigma_\eta^2 + \sigma_v^2 \end{bmatrix} \equiv \Omega_i$$

where Ω_i is a $T \times T$ matrix.

Finally, we stack this across all individuals to get:

$$\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^T | \boldsymbol{X}] = \begin{bmatrix} \Omega_i & 0 & \dots & 0 \\ 0 & \Omega_i & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Omega_i \end{bmatrix} \equiv \Omega. \tag{558}$$

Note that Ω is a $NT \times NT$ matrix where each entry Ω_i is a $T \times T$ matrix. □

We can now state the random effects GLS estimator.

Definition 626 (Random Effects Generalized Least Squares Estimator)
The Random Effects Generalized Least Squares estimator is:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T \Omega^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \Omega^{-1} \boldsymbol{y} \tag{559}$$

where Ω is a $NT \times NT$ matrix.

We finally get our efficient estimator!

Proposition 627 (Efficiency of Random Effects GLS Estimator)
Under the classical random-effects assumptions and with known Ω, the Random Effects Generalized Least Squares estimator is the efficient linear unbiased estimator.

We can also show that it is a consistent estimator.

Proposition 628 (Consistency of Random Effects GLS Estimator)
Under strict exogeneity, uncorrelated individual effects, cross-sectional independence, and the usual moment and full-rank conditions, $\hat{\boldsymbol{\beta}}_{GLS}$ is a consistent estimator as $N \to \infty$ with T fixed.

The issue we have now is that Ω is a $NT \times NT$ matrix, which is really large. Trying to invert this matrix is computationally intensive. Therefore, we need to use other techniques instead to compute this random effects GLS estimator. We move onto another transformation that we can use that uses an OLS but gives the same result as the random effects GLS estimator.

13.4.4 Theta Differencing

We have argued that it is computationally intensive to compute the random effects generalised least squares. We can instead use theta-differencing transformation and use the standard OLS in this transformed model. First, we define what is this new transformation.

Definition 629 (Theta Differencing Transformation)
The theta differencing transformation is:

$$y_{it}^* = y_{it} - (1-\theta)\overline{y}_i \tag{560}$$

where θ is a parameter that is the ratio of variances such that:

$$\theta^2 = \frac{\sigma_v^2}{\sigma_v^2 + T\sigma_\eta^2} \tag{561}$$

$$\overline{y}_i = \frac{1}{T}\sum_{s=1}^{T} y_{is} \tag{562}$$

where σ_v^2 is the variance of v_{it} error and σ_η^2 is the variance of η_i error.

If we know σ_v^2 and σ_η^2, we can first compute θ^2, then θ and $(1-\theta)$. We can then retrieve the transformed variables to get our theta-differenced structural equation.

We can use this theta differencing to construct our new transformed structural equation.

Definition 630 (Theta Differencing)
Suppose that the theta differenced structural equation is:

$$y_{it}^* = \boldsymbol{x}_{it}^*\boldsymbol{\beta} + u_{it}^* \tag{563}$$

for $i = 1, 2, ..., N$ where $u_{it}^* = \eta_i^* + v_{it}^*$.

The theta-differencing estimator is the OLS estimator of the theta differenced structural equation:

$$\hat{\boldsymbol{\beta}}_\theta = ((\boldsymbol{X}^*)^T\boldsymbol{X}^*)^{-1}(\boldsymbol{X}^*)^T\boldsymbol{y}^*. \tag{564}$$

Notice here that the time-invariant error term

$$\eta_i^* = \eta_i - (1-\theta)\eta_i = \theta\eta_i$$

does **not** eliminate the time-invariant component η_i from the error term of the transformed model. Therefore, we still require the uncorrelated individual effects assumption:

$$\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = 0$$

for consistency of the OLS estimator since the error term η_i is still present. However, η_i is eliminated when $\theta = 0$.

Furthermore, we also have that the theta-differenced individual-specific error term is:

$$v_{it}^* = v_{it} - (1-\theta)\overline{v}_i = v_{it} - (1-\theta)\frac{1}{T}(v_{i1} + ... + v_{iT}).$$

Since our theta differenced explanatory variable $\boldsymbol{x}_{it}^*$ also contains $\frac{1}{T}\sum_{s=1}^{T}\boldsymbol{x}_{is}$, we therefore require strict exogeneity between the explanatory variable and the individual specific error term:

$$\mathbb{E}[\boldsymbol{x}_{it}v_{is}] = 0$$

for $t, s \in \{1, ..., T\}$.

Proposition 631 (Consistency of Theta-Differenced OLS Estimator)
Under strict exogeneity and the random-effects assumption, the OLS estimator after θ-differencing is a consistent estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_\theta \xrightarrow{P} \boldsymbol{\beta} \tag{565}$$

as $N \to \infty$ with T fixed.

Don't forget that the purpose of this new estimator was that it is a more computationally efficient way to compute random effects GLS!

Proposition 632 (Equivalency of Theta-Differenced OLS and Random Effects GLS)
The random effects GLS estimator is equivalent to the theta-differenced OLS estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} = \hat{\boldsymbol{\beta}}_\theta. \tag{566}$$

In practice, we don't actually know the variance of the error terms σ_η^2 and σ_v^2, so therefore we need to use consistent estimators. This leads us to feasible GLS.

Definition 633 (Feasible Random Effects Generalised Least Squares)
Feasible random effects generalised least squares estimators is the same as random effects generalised least squares estimator but with consistent estimators of variance of the error terms σ_η^2 and σ_v^2.

From the classical assumptions, we can obtain these consistent estimators of the error terms using within groups and between groups estimator. We can also show that feasible GLS is the same as GLS and therefore is also consistent.

Proposition 634 (Feasible GLS and GLS)
The feasible random effects GLS estimator is asymptotically equivalent to the random effects GLS estimator.

Therefore, under the assumptions of strict exogeneity and uncorrelated individual effects, the feasible random effects GLS estimator is a consistent estimator of $\boldsymbol{\beta}$.

We can show that the GLS estimator coincides with the OLS estimator and the WG estimator under certain settings.

Proposition 635 (GLS and Pooled OLS Without Individual Effects)
When $\sigma_\eta^2 = 0$, the GLS estimator and the pooled OLS estimator coincide:

$$\hat{\boldsymbol{\beta}}_{GLS} = \hat{\boldsymbol{\beta}}_{OLS}. \tag{567}$$

This has a nice interpretation because when $\sigma_\eta^2 = 0$, this says that there are no individual-specific error terms η_i. This then mean that there is no serial correlation in the error $u_{it} = v_{it} + \eta_i = v_{it}$. But recall that the serial correlation in u_{it} is what caused our pooled OLS estimator to be inefficient! Therefore, since we no longer have serial correlation in u_{it}, this means that our pooled OLS estimator will be efficient and in this case, coincide with the random effects GLS estimator!

We now look at the second case where we see that the random effects GLS estimator coincides with the within groups estimator.

Proposition 636 (GLS and WG Estimator and LSDV Estimator in Large T)
As $T \to \infty$, then the GLS estimator converges to the within groups estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} - \hat{\boldsymbol{\beta}}_{WG} \xrightarrow{P} \mathbf{0}. \tag{568}$$

Furthermore, as the Least Squares Dummy Variables estimator is equivalent to the within groups estimator, this implies that the GLS estimator also converges to the Least Squares Dummy Variables estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} - \hat{\boldsymbol{\beta}}_{LSDV} \xrightarrow{P} \mathbf{0}. \tag{569}$$

Finally, as the random effects GLS estimator is an efficient estimator, this implies that in large T samples, the within groups estimator $\boldsymbol{\beta}_{WG}$ and least squares dummy variables estimator $\boldsymbol{\beta}_{LSDV}$ are **efficient estimators** even under the assumption of uncorrelated individual effects.

Proof. First recall the theta differencing transformation, which also coincides with the random effects GLS estimator:

$$y_{it}^* = y_{it} - (1-\theta)\overline{y}_i$$

where

$$\theta^2 = \frac{\sigma_v^2}{\sigma_v^2 + T\sigma_\eta^2}$$
$$\overline{y}_i = \frac{1}{T}\sum_{s=1}^{T} y_{is}$$

Here, as $T \to \infty$, we have that $\theta^2 \to 0$. This then implies that $\theta \to 0$. This then means our theta differenced transformation becomes:

$$y_{it}^* = y_{it} - (1-\theta)\overline{y}_i = y_{it} - \overline{y}_i = \tilde{y}_{it}$$

where $\tilde{y}_{it}$ is the within groups transformation! Therefore our theta differencing transformation coincides with the within groups transformation!

□

Here, we see that as $T \to \infty$, the theta transformation approaches the within transformation because $1 - \theta \to 1$. This indicates that with large-T samples the efficiency difference between random-effects GLS and the within estimator vanishes under the maintained random-effects assumptions. The time average itself does not generally converge to zero; it is the GLS transformation weight that converges to the within-transformation weight. For small T, the within estimator is generally not efficient under uncorrelated individual effects, whereas correctly specified random-effects GLS is efficient. Since LSDV is algebraically equivalent to WG for the slope coefficients, the same large-T equivalence applies to LSDV.

13.5 Testing for Random Effects

So far, we have made use of the uncorrelated individual effects assumption multiple times:

$$\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = 0.$$

We are now interested in testing whether if **some** of the K explanatory variables $\boldsymbol{x}_{it}$ are correlated with the individual specific error η_i.

We assume strict exogeneity for all the tests in this section.

Assumption 637 (Strict Exogeneity)

All K explanatory variables are uncorrelated with individual time-varying error terms over all periods:

$$\mathbb{E}[\boldsymbol{x}_{it}v_{is}] = \mathbf{0} \tag{570}$$

for all $s, t \in \{1, 2, ..., T\}$.

Our test will be similar to what we did in instrumental variables. In particular, we shall compare the difference between estimators that are consistent under random effects against estimators that are not consistent under random effects. Therefore, if there is a big difference between these estimators, that means that there is correlated individual effects so random effects does not hold:

$$\mathbb{E}[\eta_i \boldsymbol{x}_i] \neq 0.$$

We now recall which estimators are consistent estimators under strict exogeneity and **correlated** individual effects (fixed effects):

1. Within Groups $\hat{\boldsymbol{\beta}}_{WG}$
2. First Differenced OLS $\hat{\boldsymbol{\beta}}_{FDOLS}$.

Remark 638. *Recall that these estimators were consistent but not efficient under random effects.*

We also recall which estimators are consistent estimators under strict exogeneity and **uncorrelated** individual effects (random effects):

1. Pooled OLS $\hat{\boldsymbol{\beta}}_{OLS}$
2. Between Groups $\hat{\boldsymbol{\beta}}_{BG}$
3. Random Effects GLS $\hat{\boldsymbol{\beta}}_{GLS}$.

Therefore, we need to pick and compare an estimator from these two lists of estimators. We now state the hypothesis test that we will need.

Definition 639 (Hypothesis Test for Random Effects)

$$H_0 : \mathbb{E}[\eta_i | \boldsymbol{X}_i] = 0. \tag{571}$$

Remark 640. *Notice that this is a stronger test compared to testing for uncorrelated individual effects* $\mathbb{E}[\boldsymbol{X}_i\eta_i] = 0$.

We impose the classical assumptions previously mentioned.

Assumption 641 (Classical Assumptions)

The classical assumptions are:

1. $\eta_i | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT} \sim i.i.d(0, \sigma_\eta^2)$ over $i = 1, ..., N$
2. $v_{it} | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT} \sim i.i.d(0, \sigma_v^2)$ over $i = 1, ..., N$ and $t = 1, \ldots, T$
3. η_i and v_{it} are independent conditional on $\boldsymbol{X}_i$
4. $\mathbb{E}[\eta_i | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}] = 0$
5. $\mathbb{E}[v_{it} | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}] = 0$ for each $t \in \{1, 2, ..., T\}$

We can now state the Hausman test.

Definition 642 (Hausman Test)

The vector of differences are defined to be:

$$\hat{\boldsymbol{q}} = \hat{\boldsymbol{\beta}}_{WG} - \hat{\boldsymbol{\beta}}_{GLS}. \tag{572}$$

The scalar Hausman test statistic is:

$$h = \hat{\boldsymbol{q}}^T \Big[Var[\hat{\boldsymbol{q}}] \Big]^{-1} \hat{\boldsymbol{q}} \xrightarrow{D} \chi^2(K) \tag{573}$$

under the null hypothesis $H_0 : \mathbb{E}[\eta_i | \boldsymbol{X}_i] = 0$.

Furthermore, the variance of the vector of differences is given by:

$$Var[\hat{\boldsymbol{q}}] = Var[\hat{\boldsymbol{\beta}}_{WG}] - Var[\hat{\boldsymbol{\beta}}_{GLS}]. \tag{574}$$

Remark 643. *The variance term makes use of the fact that GLS estimator is efficient and therefore will have a smaller variance.*

Something to note is that in instrumental variable testing, we required conditional homoskedasticity to use the Hausman test. However, we also require conditional homoskedasticity to compute the GLS estimator, so therefore this assumption is satisfied!

We can also construct another Hausman test where we instead use the between groups estimator.

Definition 644 (Alternative Hausman Test)

The vector of differences are defined to be:

$$\hat{\boldsymbol{q}}^* = \hat{\boldsymbol{\beta}}_{WG} - \hat{\boldsymbol{\beta}}_{BG}. \tag{575}$$

The scalar Hausman test statistic is:

$$h^* = \hat{\boldsymbol{q}}^{*T} \Big[Var[\hat{\boldsymbol{q}}^*] \Big]^{-1} \hat{\boldsymbol{q}}^* \xrightarrow{D} \chi^2(K) \tag{576}$$

under the null hypothesis $H_0 : \mathbb{E}[\eta_i | \boldsymbol{X}_i] = 0$.

Furthermore, the variance of the vector of differences is given by:

$$Var[\hat{\boldsymbol{q}}^*] = Var[\hat{\boldsymbol{\beta}}_{WG}] + Var[\hat{\boldsymbol{\beta}}_{BG}]. \tag{577}$$

Remark 645. *Notice the variance term has a + rather than a -. This is due to the fact of the orthogonality between the within groups and between groups estimator.*

Now, if we look at a single scalar parameter $\beta^k \in \boldsymbol{\beta}$, we therefore have that the vector of differences is now a scalar difference:

$$\hat{q}^{'k} = \hat{\beta}^k_{WG} - \hat{\beta}^k_{GLS}.$$

The test statistic then simplifies to:

$$h^k = \frac{\left(\hat{\beta}^k_{WG} - \hat{\beta}^k_{GLS}\right)^2}{Var[\hat{\beta}^k_{WG}] - Var[\hat{\beta}^k_{GLS}]} \xrightarrow{D} \chi^2(1) \tag{578}$$

under the null hypothesis $H_0 : \mathbb{E}[\eta_i|\boldsymbol{X}_i] = 0$. Equivalently, define the signed statistic:

$$z^k = \frac{\hat{\beta}^k_{WG} - \hat{\beta}^k_{GLS}}{\sqrt{Var[\hat{\beta}^k_{WG}] - Var[\hat{\beta}^k_{GLS}]}} \xrightarrow{D} \mathcal{N}(0,1), \tag{579}$$

under the null hypothesis $H_0 : \mathbb{E}[\eta_i|\boldsymbol{X}_i] = 0$. Its square satisfies $(z^k)^2 = h^k$ asymptotically.

13.6 Time-Invariant Explanatory Variables

We now want to include the time-invariant explanatory variable $\boldsymbol{w}_i$ back into our structural model. The within transformation removes $\boldsymbol{w}_i$ together with the individual effect, so its coefficient cannot be identified by the standard fixed-effects estimator without additional instruments or restrictions. The following is only a schematic first-stage setup; a complete estimator also requires explicit exclusion, exogeneity, and rank conditions.

The first-stage linear projection is:

$$\boldsymbol{w}_i = \overline{\boldsymbol{x}}_i \boldsymbol{\Pi} + \boldsymbol{z}_i \boldsymbol{\theta} + \boldsymbol{r}_i. \tag{580}$$

In the scalar single-instrument case, relevance requires:

$$\theta \neq 0. \tag{581}$$

With vector-valued variables, this is replaced by the corresponding full-rank condition on the excluded-instrument coefficient matrix.

Assumption 646 (Serially Uncorrelated Time-Varying Errors)
The time-varying error terms v_{it} are serially uncorrelated.

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{w}_i\boldsymbol{\gamma} + (\eta_i + v_{it}) \tag{582}$$

for $i = 1, 2, ..., N$ and $t = 1, 2, ..., T$.

13.7 Models with Predetermined Covariates

Assumption 647 (Fixed Effects)

$$\mathbb{E}[\boldsymbol{x}_{it}\eta_i] \neq \mathbf{0}. \tag{583}$$

Definition 648 (Predetermined Explanatory Variable)
Suppose that the time-varying component is serially uncorrelated. Then, the explanatory variable $\boldsymbol{x}_{it}$ is **predetermined** with respect to the time-varying error component v_{it} if:

$$\mathbb{E}[\boldsymbol{x}_{it}v_{is}] = \mathbf{0} \text{ for } s \geq t \tag{584}$$

$$\mathbb{E}[\boldsymbol{x}_{it}v_{is}] \text{ may differ from } \mathbf{0} \text{ for } s \leq t-1. \tag{585}$$

We could also write this as:

$$\mathbb{E}[\boldsymbol{x}_{it}v_{is}] = 0 \quad \text{for } s = t, t+1, ..., T \tag{586}$$

$$\mathbb{E}[\boldsymbol{x}_{it}v_{is}] \text{ may differ from } 0 \quad \text{for } s = 1, 2, ..., t-1. \tag{587}$$

This is a relaxation of the assumption of strict exogeneity $\mathbb{E}[\boldsymbol{x}_{it}v_{is}] = \mathbf{0}$ for $s, t \in \{1, ..., T\}$.

13.7.1 First Differenced 2SLS Estimator

We need:

$$\mathbb{E}[\Delta\boldsymbol{x}_{it}\Delta v_{it}] = \mathbf{0}.$$

However, we have already argued that $\Delta\boldsymbol{x}_{it}$ is correlated with Δv_{it}. What we can do instead is to treat $\Delta\boldsymbol{x}_{it}$ as an endogenous variable like what we saw in instrumental variables. Then, we can try to find an instrument $\boldsymbol{z}_{it}$ for $\Delta\boldsymbol{x}_{it}$ such that:

1. The instrument $\boldsymbol{z}_{it}$ is **valid**: $\mathbb{E}[\boldsymbol{z}_{it}\Delta v_{it}] = \mathbf{0}$
2. The instrument $\boldsymbol{z}_{it}$ is **informative**: $\boldsymbol{z}_{it}$ is correlated with $\Delta\boldsymbol{x}_{it}$.

$$\Delta y_{it} = \Delta\boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta v_{it} \tag{588}$$

for $i = 1, 2, ..., N$ and $t = 3, ..., T$.

$$\mathbb{E}[\boldsymbol{x}_{i,t-1}\Delta v_{it}] = \mathbf{0}.$$

This is an instrumental-variable orthogonality condition; it does not make $\Delta \boldsymbol{x}_{it}$ exogenous and therefore does not justify FDOLS. Stack the valid instruments in $\boldsymbol{Z}$ and let $\boldsymbol{P}_Z = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$. The first-differenced 2SLS estimator is

$$\hat{\boldsymbol{\beta}}_{FD2SLS} = \left[(\Delta \boldsymbol{X})^T \boldsymbol{P}_Z (\Delta \boldsymbol{X})\right]^{-1} (\Delta \boldsymbol{X})^T \boldsymbol{P}_Z \Delta \boldsymbol{y}.$$

Under instrument validity, relevance, cross-sectional independence, and the usual moment and rank conditions,

$$\hat{\boldsymbol{\beta}}_{FD2SLS} \xrightarrow{P} \boldsymbol{\beta} \quad \text{as } N \to \infty \text{ with } T \text{ fixed.}$$