Econometrics Notes - PhD in Economics 2025-2026

Department of Economics

Bocconi Università

Raphaël Boulat

Note to the reader

These notes are based partially on older versions of notes from Nicola Limodio's PhD course at Bocconi, and Taisuke Otsu's EC484 course at LSE. Some probability/statistic notions are borrowed from Sandra Fortini's Intro to Probability course at Bocconi. I also use some textbooks, with the main reference being Hansen, 2022. Please contact me (raphael.boulat@phd.unibocconi.it) if you notice any errors or have comments or suggestions.

Contents

1 Classical Linear Regression Model **4**
- 1.1 Model and Gauss-Markov Assumptions . . . 4
- 1.2 Estimation procedures . . . 5
  - 1.2.1 Minimum Distance Estimation - OLS . . . 5
  - 1.2.2 Method of Moments (MM) . . . 5
  - 1.2.3 Maximum Likelihood Estimation (MLE) . . . 6
- 1.3 Goodness of Fit . . . 7
- 1.4 Geometric interpretation of OLS and Frisch–Waugh–Lovell (FWL) . . . 8
  - 1.4.1 Quick linear algebra recap . . . 8
  - 1.4.2 Geometric interpretation of OLS . . . 9
  - 1.4.3 Frisch–Waugh–Lovell (FWL) . . . 12
- 1.5 Finite-Sample Properties . . . 14
  - 1.5.1 Unbiasedness . . . 14
  - 1.5.2 Variance . . . 15
- 1.6 Asymptotic Properties of OLS . . . 17
  - 1.6.1 Probability Theory foundations . . . 17
  - 1.6.2 Consistency of the least squares estimator . . . 22
  - 1.6.3 Asymptotic Normality of the least squares estimator . . . 23
- 1.7 Quick detour on intuition & collinearity . . . 23
- 1.8 Function of Parameters . . . 24
- 1.9 Asymptotic standard errors . . . 25
- 1.10 On the way to hypothesis testing: object construction . . . 26
  - 1.10.1 T-statistic . . . 26
  - 1.10.2 Confidence Intervals . . . 26
  - 1.10.3 Wald Statistic . . . 27

2 Hypothesis Testing **27**
- 2.1 Concepts . . . 27
  - 2.1.1 Optimality and The Neyman-Pearson Lemma . . . 28
- 2.2 Linear hypotheses and joint significance . . . 29
  - 2.2.1 T-test . . . 29
  - 2.2.2 F-test . . . 29
  - 2.2.3 Restricted Least Squares Derivation . . . 30
- 2.3 The Trinity of Testing . . . 31
  - 2.3.1 The ML Estimator . . . 31
  - 2.3.2 Wald Test . . . 33
  - 2.3.3 The Lagrange Multiplier or Score Test . . . 33
  - 2.3.4 The Likelihood Ratio Test . . . 34

3 Generalized Linear Regression **35**
- 3.1 Finite-Sample Properties . . . 36
- 3.2 Asymptotic Properties . . . 36
- 3.3 Generalized Least Squares (GLS) . . . 39

3.3.1 Theoretical GLS . . . 39
3.3.2 Feasible GLS . . . 41
3.3.3 Application: Heteroskedasticity . . . 43
3.3.4 Heteroskedasticity Tests . . . 45

4 Identification and Instrumental Variables **45**
4.1 Cause of Endogeneity . . . 46
4.1.1 Measurement error . . . 46
4.1.2 Simultaneity . . . 47
4.1.3 Omitted Variable Bias . . . 48
4.2 Instrumental Variables and Identification . . . 49
4.2.1 Set-up . . . 49
4.2.2 Identification . . . 50
4.3 Instrumental Variables and moments conditions . . . 51
4.4 Two-Stage Least Squares (2SLS) . . . 54
4.4.1 2SLS derivation . . . 54
4.4.2 Asymptotic properties of 2SLS . . . 54
4.5 Validity Tests and Weak Instruments . . . 57
4.5.1 Testing for the Validity of Instruments: The J-Test . . . 57
4.5.2 Weak Instruments . . . 58

5 GMM and Extremum Estimation **58**
5.1 Intro to GMM . . . 58
5.2 Extremum Estimators . . . 60
5.2.1 General Consistency . . . 60
5.2.2 General asymptotic normality . . . 62
5.3 Linking the two concepts . . . 63
5.4 Optimal GMM . . . 67
5.5 Testing in GMM . . . 69
5.5.1 Sargan-Hansen Test . . . 69
5.5.2 Hausman Test . . . 70

6 Treatment Effects **71**
6.1 Setup . . . 71
6.2 Objects of interest . . . 71
6.2.1 Building the objects . . . 71
6.2.2 Understanding the objects . . . 72
6.3 The counterfactual and Selection Bias . . . 73
6.3.1 The Problem . . . 73
6.3.2 Some solutions . . . 73
6.4 Estimation . . . 74
6.4.1 Outcome specification . . . 74
6.4.2 Participation Decision . . . 75
6.5 Instrumental Variables . . . 75
6.5.1 The structural setup for IV . . . 76

6.5.2 IV estimation . . . 77
6.5.3 LATE . . . 77
6.6 Some other topics . . . 78
6.6.1 The Two-Step Heckman Estimator . . . 78
6.6.2 Matching . . . 79
6.6.3 Difference-in-differences . . . 80

7 Panel Data **80**
7.1 Fixed Effects Models . . . 81
7.1.1 General case . . . 81
7.1.2 The Within Estimator (Demeaning) . . . 83
7.1.3 Two-way ECM with FE . . . 86
7.1.4 Drawbacks of the fixed-effect model . . . 87
7.2 Random Effects Models . . . 87
7.3 Fixed Effects vs. Random Effects . . . 90

8 Recent DiD paper **91**
8.1 Research Question . . . 91
8.2 Framework . . . 91
8.3 Assumptions . . . 91
8.4 The Underidentification Problem . . . 92
8.5 Negative Weights Problem . . . 93
8.6 The Efficient Estimator . . . 94
8.6.1 The general theorem . . . 94
8.6.2 The Imputation Estimator . . . 95
8.7 Pre-trend Testing . . . 95
8.8 Application: Marginal Propensity to Consume Out of Tax Rebates . . . 95
8.9 Asymptotic Properties . . . 96
8.10 Key Takeaways . . . 97

A Appendix **99**
A.1 Asymptotic Tools for i.n.i.d. Data . . . 99
A.2 Additional derivations . . . 100
A.2.1 Restricted Least Squares Derivation . . . 100

1 Classical Linear Regression Model

1.1 Model and Gauss-Markov Assumptions

Let $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times k}$ and $\beta \in \mathbb{R}^k$ the parameter of interest. The general form of the multiple linear regression model is

$$y = X\beta + \varepsilon. \tag{1.1}$$

Assumption 1.1: Gauss Markov Assumptions + Normality

The Gauss-Markov Linear Regression Assumptions are given below

(i) Linearity: The true model is $y = X\beta + \varepsilon$, with $\mathbb{E}(\varepsilon) = 0$

(ii) Full rank: $\text{rank}(X) = k \leq n \implies X'X$ is invertible

(iii) Exogeneity: $\mathbb{E}[\varepsilon \mid X] = 0$

(iv) Spherical error: $\mathbb{V}(\varepsilon \mid X) = \sigma^2 I_n$

(v) Normality: $\varepsilon | X \sim \mathcal{N}(0, \sigma^2 I_n)$

We quickly discuss the intuition and implications of some of the assumptions below.

The *full rank* assumption just states that it is impossible to estimate a k number of β with n observations if $k > n$. This assumption also prevents from perfect collinearity.

The *exogeneity* assumption deserves a longer treatment. Consider the case where X are treated as random variables. First, notice that full independence, $X \perp\!\!\!\perp \varepsilon$ implies that for any measurable $g(\cdot)$ with finite moments, $\mathbb{E}(g(X)\epsilon) = \mathbb{E}(\varepsilon)\mathbb{E}(g(X))$. Specifically, $\mathbb{E}(X'\varepsilon) = \mathbb{E}(X')\mathbb{E}(\varepsilon) = 0$ in this context. Mean independence only implies that $\mathbb{E}(\varepsilon|X) = \mathbb{E}(\varepsilon) = 0$. For our next argument, we introduce the following proposition.

Proposition 1.1: Law of Iterated Expectations (LIE)

Let Z be an integrable random variable and let $\mathcal{G}$ a sigma-algebra generated by a partition. Then, $\mathbb{E}(\mathbb{E}(Z|\mathcal{G})) = \mathbb{E}(Z)$

Finally, notice that in this context, we can write the covariance $\mathbb{C}\text{ov}(\varepsilon_i, x_i) = \mathbb{E}(\varepsilon_i x_i) - \mathbb{E}(\varepsilon_i)\mathbb{E}(x_i) = \mathbb{E}(\mathbb{E}(\varepsilon_i x_i|x_i)) = \mathbb{E}(x_i \mathbb{E}(\varepsilon_i|x_i)) = 0$, where the second equality is an application of LIE.
An interesting observation is hence that independence $\implies$ mean independence $\implies$ uncorrelatedness.

The *spherical errors* assumption can be rewritten as follows:

$$\mathbb{V}(\varepsilon \mid X) = \sigma^2 I_n \Leftrightarrow \mathbb{V}(\varepsilon_i \mid X) = \sigma^2, \forall i \text{ and } Cov(\varepsilon_i, \varepsilon_j \mid X) = 0, \forall i \neq j$$

Therefore, we can see that this is an assumption about homoskedasticity and autocorrelation. Notice that iid assumption for $\varepsilon_i \mid X$ implies spherical errors, but the other way around is not true.

1.2 Estimation procedures

1.2.1 Minimum Distance Estimation - OLS

The problem is

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 = \arg\min_{\beta} (y - X\beta)'(y - X\beta) = \arg\min_{\beta} S(\beta)$$

With $S(\beta)$ being the sum of squared residuals. The necessary condition for a minimum is a first order condition (FOC):

$$\left.\frac{\partial S(\beta)}{\partial \beta}\right|_{\hat{\beta}} = 0 - 2X'y + 2X'X\hat{\beta} = 0$$

This is equivalent to $X'\hat{\varepsilon} = 0$, since $\hat{\varepsilon} = y - X\hat{\beta}$ are the residuals.

We solve this as follows, and these equations are also called *the normal equation*:

$$-2X'y + 2X'X\hat{\beta} = 0 \Rightarrow \hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

Notice that by assumption (ii), the inverse of exists $(X'X)$, which ensures that $\hat{\beta}_{OLS}$ is unique. This estimator is a minimun, by second order conditions:

$$\left.\frac{\partial^2 S(\beta)}{\partial \beta \partial \beta'}\right|_{\hat{\beta}} = 2X'X$$

This matrix must be positive definite.

1.2.2 Method of Moments (MM)

The MM estimator of β is based on the following moment conditions

$$E(x_{ij}\varepsilon_i) = 0 \quad j = 1 \dots k$$

These are the orthogonality conditions: errors are uncorrelated with the regressors. $\hat{\beta}_{MM}$ ensures that the sample analogue moment conditions are satisfied. To find the estimator, replace the expectation with the sample average

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}\hat{\varepsilon}_i = 0 \quad \forall j \qquad \text{or} \qquad \frac{1}{n}\sum_{i=1}^{n} x_{ij}\left(y_i - x_i'\hat{\beta}_{MM}\right) = 0$$

In matrix notations this implies that

$$\frac{1}{n}X'(y - X\hat{\beta}_{MM}) = \frac{1}{n}X'\hat{\varepsilon} = 0 \Rightarrow \hat{\beta}_{MM} = \hat{\beta}_{OLS}$$

The MM estimator of σ^2 is based on the moment condition $\mathbb{E}(\varepsilon_i^2) = \sigma^2$ using the sample analogue moment

$$\hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

1.2.3 Maximum Likelihood Estimation (MLE)

The MLE requires us to have a sample of N independently drawn observations $(y_1, x_1), \ldots, (y_N, x_N)$ with conditional distribution $f(y|x, \beta)$ where the functional form f is known but parameters β are not. Given assumptions (i) and (v), we have that $y|X \sim N(X\beta, \sigma^2 I_n)$ and that $y_1 \ldots y_n$ are independent. We want to estimate β, and hence we choose $\hat{\beta}_{MLE}$ to maximize $\mathbb{P}(y_1, y_2, ..., y_N|\hat{\beta}_{MLE}, x_1, x_2, ..., x_N)$, that is the probability of observing the data if $\hat{\beta}_{MLE}$ was true.

$$\begin{aligned}
f(y_1 \ldots y_n|X; \beta, \sigma^2) &= \prod_{i=1}^{n} f(y_i|X; \beta, \sigma^2) \quad - \quad \text{independence} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right\} \quad - \quad \text{normality} \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right\} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot \exp\left\{-\frac{1}{2\sigma^2}S(\beta)\right\}
\end{aligned}$$

The log likelihood function is

$$l(\beta, \sigma^2; y_1 \ldots y_n, x_1 \ldots x_n) = -\frac{n}{2}\log(2\pi\sigma) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}S(\beta)$$

We take two FOCs wrt to each unknown paramter (β and σ^2).

$$\left.\frac{\partial l(\beta, \sigma^2)}{\partial \beta}\right|_{\hat{\beta}_{MLE}, \hat{\sigma}^2_{MLE}} = -\frac{1}{2\hat{\sigma}^2_{MLE}}\frac{\partial S(\hat{\beta}_{MLE})}{\partial \beta} = 0 \Leftrightarrow \frac{\partial S(\hat{\beta}_{MLE})}{\partial \beta} = 0 \tag{1.2}$$

This shows that maximizing the likelihood under normality with respect to β leads to $\hat{\beta}_{MLE} = \hat{\beta}_{MM} = \hat{\beta}_{OLS}$.

$$\left.\frac{\partial l(\beta, \sigma^2)}{\partial \sigma^2}\right|_{\hat{\beta}_{MLE}, \hat{\sigma}^2_{MLE}} = -\frac{n}{2\hat{\sigma}^2_{MLE}} + \frac{1}{2\hat{\sigma}^4_{MLE}}S(\hat{\beta}_{MLE}) = 0 \tag{1.3}$$

1.3 Goodness of Fit

We define different objects required to study goodness of fit.

Definition 1.1: Total sum of squares

The Total Sum of Squares is given by

$$TSS = \sum_i (y_i - \bar{y}_i)^2$$

Definition 1.2: Explained sum of squares

The Explained Sum of Squares is given by

$$ESS = \sum_i (\hat{y}_i - \bar{y}_i)^2$$

Definition 1.3: Residual sum of squares

The Residuals Sum of Squares is given by

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

Some intuition:[1] The TSS represents the total variability in the dependent variable, including the portions of variability both explained and unexplained by the regression model. By its nature, the TSS reflects the overall dispersion of individual values of the observed dependent variable from its mean value. The ESS captures the portion of the total variability in the dependent variable explained by the regression model. Finally, the RSS reflects the remaining error, or unexplained variability, which is the portion of the total variability in the dependent variable that can't be explained by the regression model.

With these definitions in mind, we can now define three alternative definitions of the R^2.

Definition 1.4: Three alternative definitions of R^2

(i) $R_1^2 = 1 - \frac{RSS}{TSS}$

(ii) $R_2^2 = \frac{[\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}_i)]^2}{[\sum_i (y_i - \bar{y})]^2 [(\hat{y}_i - \bar{\hat{y}}_i)]^2}$

(iii) $R_3^2 = \frac{ESS}{TSS}$

[1] This website provides cool intuition.

Proposition 1.2: Equivalence between different definitions of R^2

When an intercept is included in the model, all three measures are the same and $R^2 \in [0,1]$.

Proof. First, we can show that $R_1^2 = R_3^2$.

$$\begin{aligned} TSS = \sum (y_i - \bar{y})^2 &= \sum \left[(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \right]^2 \\ &= ESS + RSS + 2\sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \\ &= ESS + RSS + 2\sum (\hat{y}_i - \bar{y})\varepsilon_i \end{aligned}$$

The last therm is 0 since residuals are orthogonal to the fitted values. Therefore, we have that $\frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$ and therefore $R_3^2 = R_1^2$.

$$R_2^2 = \frac{\left[\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right]^2}{\left[\sum (y_i - \bar{y})^2\right] \left[\sum (\hat{y}_i - \bar{\hat{y}})^2\right]} = \frac{\left[\sum (y_i - \bar{y})(\hat{y}_i - \bar{y})\right]^2}{TSS \cdot ESS}$$

The numerator can be re-written as follows:

$$\begin{aligned} \left(\sum (y_i - \bar{y})(\hat{y}_i - \bar{y})\right)^2 &= \left(\sum \left[(\hat{y}_i + e_i - \bar{y})(\hat{y}_i - \bar{y}) \right]\right)^2 \\ &= \left(\sum \left[(\hat{y}_i - \bar{y}) + e_i \right](\hat{y}_i - \bar{y})\right)^2 \\ &= \left(\sum (\hat{y}_i - \bar{y})^2 + \sum e_i (\hat{y}_i - \bar{y})\right)^2 \\ &= ESS^2 \end{aligned}$$

Therefore, we have

$$R_2^2 = \frac{ESS^2}{ESS * TSS} = \frac{ESS}{TSS} = R_3^2$$

□

The main drawback of the R^2 is that it never decreases as we increase the number of regressors, even if the new variables have no explanatory power. The adjusted R^2 allows to correct for the number of explanatory variables:

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)}$$

Suppose that we are in a case where we add an additional regressor with no explanatory power. RSS does not change (specifically it does not decrease, what we desire), but $(n-k)$ goes down and therefore the $\bar{R}^2$ goes down as well.

1.4 Geometric interpretation of OLS and Frisch–Waugh–Lovell (FWL)

1.4.1 Quick linear algebra recap

We provide some basic definitions and one proposition that are useful in our application.

Definition 1.5: Span

Consider the vector space V and a set of vectors $A \subseteq V$. The span of A, denoted $spanA$, is the set of all linear combinations of vectors in A. This is given by

$$spanA = \left\{ \sum_{i=1}^{n} \alpha_i v_i | \forall i \in \mathbb{N}, \alpha_i \in \mathbb{R}, v_i \in A \right\}$$

Definition 1.6: Column Space (Range)

Consider an $n \times K$ matrix X with columns $x_1, x_2, \ldots, x_K$. The column space of X, denoted $\mathcal{R}(X)$ or $\text{col}(X)$, is the set of all possible linear combinations of the column vectors of X. Equivalently, the column space is the span of the columns of X:

$$\mathcal{R}(X) = \text{span}\{x_1, x_2, \ldots, x_K\} = \left\{ \sum_{j=1}^{K} \beta_j x_j \mid \beta_j \in \mathbb{R} \right\}$$

Geometrically, $\mathcal{R}(X)$ is the subspace of $\mathbb{R}^n$ generated by the columns of X.

Definition 1.7: Linear Independence

Given a vector space V, a set of vectors $\{v_1, ..., v_n\}$ is linearly *dependent* if there exists scalars $\{\alpha_1, ..., \alpha_n\}$, not all zero, such that $\sum_i \alpha_i v_i = 0$. A set of vectors $\{v_1, ..., v_n\}$ is linearly *independent* if it is not linearly dependent.

Proposition 1.3: Equivalence

Consider an $n \times n$ matrix M. The following are equivalent.

1. M is invertible
2. The row vectors of M are linearly independent
3. The column vectors of M are linearly independent

1.4.2 Geometric interpretation of OLS

Consider the model in matrix form. The fitted value is $\hat{y} = X\hat{\beta}$, the residual is $\hat{e} = y - X\hat{\beta}$, and the orthogonality condition is given by $X'\hat{e} = 0$. We first need to introduce two important matrices in linear algebra.

Definition 1.8: Projection (with properties)

The projection matrix is defined as $P_X = X(X'X)^{-1}X'$. It satisfies the following properties:

1. $P_X = P_X'$ (symmetry)
2. $P_X^2 = P_X$ (idempotent)
3. $P_X X = X$
4. $P_X y = \hat{y}$ (projection)
5. $P_X \hat{e} = 0$
6. $tr(P_X) = k$

Proof. We prove properties 4 to 6 here by inspection of the definitions, for illustration. For 4,

$$P_X y = X(X'X)^{-1}X'y = X\hat{\beta} = \hat{y}$$

For 5,

$$P_X \hat{e} = X(X'X)^{-1}X'(y - X\hat{\beta}) = X(X'X)^{-1}X'y - X(X'X)^{-1}X'X(X'X)^{-1}X'y$$
$$= X(X'X)^{-1}X'y - X(X'X)^{-1}X'y = 0$$

Now, for 6,

$$tr(P_X) = tr(X(X'X)^{-1}X') = tr((X'X)^{-1}X'X) = tr(I_k) = k$$

where the second equality comes from the fact that $tr(AB) = tr(BA)$. □

The key insight from the 5th property is that the projection on a vector orthogonal to X is equal to 0. OLS chooses the fitted vector $\hat{y}$ as the point in the column space $\text{col}(X)$ that is *closest* to y in Euclidean distance, so the residual $\hat{e} = y - \hat{y}$ is the perpendicular from y to that subspace (hence $X'\hat{e} = 0$). The projection P_X "keeps" components along $\text{col}(X)$ and "kills" components orthogonal to it, which is why $P_X y = \hat{y}$ and $P_X \hat{e} = 0$.

Definition 1.9: Annihilator matrix (with properties)

The anihilator matrix (or orthogonal projection matrix) is defined as $M_X = I_k - P_X$. It satisfies the following properties:

1. $M_X = M_X'$ (symmetry)
2. $M_X^2 = M_X$ (idempotent)
3. $M_X X = 0$ (annihilator matrix)
4. $M_X y = \hat{e}$ (orthogonal projection)

5. $M_X\hat{e} = 0$
6. $tr(M_X) = n - k$

Equivalently, the residual-maker $M_X = I - P_X$ removes everything explainable by X since $M_X y = \hat{e}$ and $M_X X = 0$. Note that by definition, the RSS is just $\hat{e}'\hat{e} = y' M_x' M_x y = y' M_X y$.

Before moving to the Frisch–Waugh–Lovell Theorem, I consruct a simple low-dimensional example to understand better what we just discussed. Figure 1.4.2 helps visualize.

Example 1.1: Geometric OLS in $\mathbb{R}^3$

Consider a case where the data space is $\mathbb{R}^3$ ($n = 3$), and we have two explanatory variables, $\mathbf{x}_1$ and $\mathbf{x}_2$ ($K = 2$). The matrix $\mathbf{X}$ is 3×2:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{pmatrix}$$

The column space $\mathcal{R}(\mathbf{X})$ is the span of the two columns, $\mathbf{x}_1$ and $\mathbf{x}_2$.

$$\mathcal{R}(\mathbf{X}) = \text{span}\{\mathbf{x}_1, \mathbf{x}_2\}$$

Since $K = 2$, $\mathcal{R}(\mathbf{X})$ is a 2-dimensional plane (the shaded gray area) embedded within the 3-dimensional space $\mathbb{R}^3$. Now, the vector $\mathbf{y}$ lies somewhere in $\mathbb{R}^3$, generally off the plane $\mathcal{R}(\mathbf{X})$.

Now, the expression $\mathbf{X}\hat{\beta}$ is, by definition, a linear combination of the columns of $\mathbf{X}$:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \hat{\beta}_1 \mathbf{x}_1 + \hat{\beta}_2 \mathbf{x}_2$$

The OLS estimator finds $\hat{\mathbf{y}}$ on the plane $\mathcal{R}(\mathbf{X})$ that is closest to $\mathbf{y}$. This closest point is found by dropping a perpendicular line from $\mathbf{y}$ to the plane. In other words, we find the vector $\hat{\mathbf{y}}$ on the $\mathbf{x}_1, \mathbf{x}_2$ plane that makes the residual $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$ orthogonal to the plane.

Since $\hat{\mathbf{y}}$ is formed by scaling $\mathbf{x}_1$ by $\hat{\beta}_1$ and $\mathbf{x}_2$ by $\hat{\beta}_2$ and summing them up, $\hat{\mathbf{y}}$ must lie on the plane (the column space $\mathcal{R}(\mathbf{X})$). The key is that the orthogonality condition ($\mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}$) is what forces the solution $\hat{\beta}$. If you choose any arbitrary β, the resulting $\mathbf{X}\beta$ will be a vector on the plane, but the residual $\mathbf{y} - \mathbf{X}\beta$ will not be perpendicular to the plane (it will just be some other vector).

Finally, $\hat{\mathbf{y}} = \mathbf{P}_\mathbf{X}\mathbf{y}$ means $\mathbf{P}_\mathbf{X}$ is the matrix that transforms the raw data vector $\mathbf{y}$ into the best possible linear combination of $\mathbf{X}$ (the predicted vector $\hat{\mathbf{y}}$). It collapses the n-dimensional vector $\mathbf{y}$ onto the K-dimensional subspace $\mathcal{R}(\mathbf{X})$.

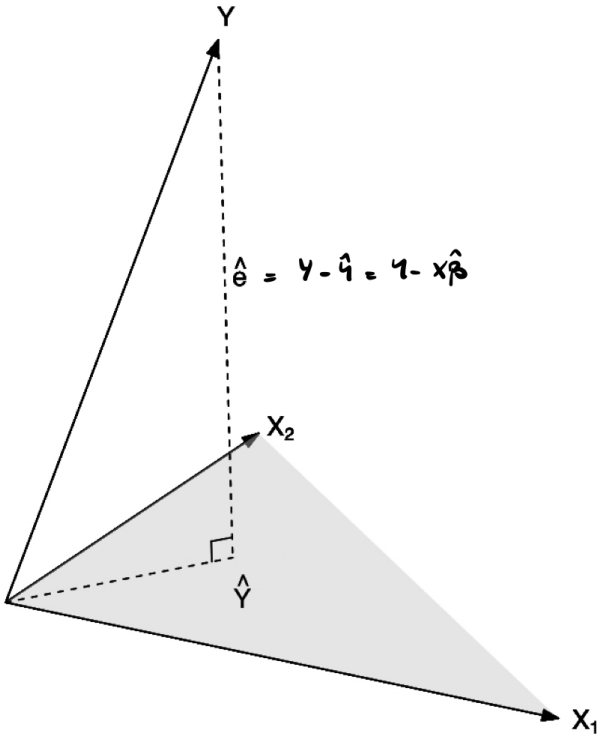

Figure 1: Projection of Y onto X_1 and X_2 (Figure 3.3 from Hansen, 2022)

1.4.3 Frisch–Waugh–Lovell (FWL)

We are now well-equiped to study the Frisch–Waugh–Lovell (FWL) theorem.

Consider the regression $y = X\beta + \varepsilon$, which can be rewritten, partitioning matrix X as $X = [X_1 : X_2]$. We can hence rewrite the regression as $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$. Notice that $dim(X_1) = n \times k_1, dim(X_2) = n \times (k - k_1)$. Additionally, recall that $P_j = X_j(X_j'X_j)^{-1}X_j', M_j = I - P_j, \forall j$

Our OLS estimator is as follows:

$$\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = (X'X)^{-1}X'y = \begin{bmatrix}X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2\end{bmatrix}^{-1}\begin{bmatrix}X_1'y \\ X_2'y\end{bmatrix}$$

Notice that by the inverse formula for partitioned matrix, we can rewrite

$$\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}S_1^{-1} & -S_1^{-1}X_1'X_2\,(X_2'X_2)^{-1} \\ -S_2^{-1}X_2'X_1\,(X_1'X_1)^{-1} & S_2^{-1}\end{bmatrix}\begin{bmatrix}X_1'y \\ X_2'y\end{bmatrix},$$

where $S_1 := X_1'X_1 - X_1'X_2\,(X_2'X_2)^{-1}X_2'X_1, S_2 := X_2'X_2 - X_2'X_1\,(X_1'X_1)^{-1}X_1'X_2$.

We can hence find an expression for $\hat{\beta}_1, \hat{\beta}_2$.

$$\begin{aligned}
\hat{\beta}_1 &= S_1^{-1}\Big(X_1'y - X_1'X_2\,(X_2'X_2)^{-1}X_2'y\Big) \\
&= \big(X_1'(I - X_2(X_2'X_2)^{-1}X_2')X_1\big)^{-1}X_1'(I - X_2(X_2'X_2)^{-1}X_2')y \\
&= (X_1'M_2X_1)^{-1}(X_1'M_2y), \\
\hat{\beta}_2 &= S_2^{-1}\Big(X_2'y - X_2'X_1\,(X_1'X_1)^{-1}X_1'y\Big) \\
&= \big(X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2\big)^{-1}X_2'(I - X_1(X_1'X_1)^{-1}X_1')y \\
&= (X_2'M_1X_2)^{-1}(X_2'M_1y)
\end{aligned}$$

Now, since M_2 is symmetric and idempotent, we have:

$$\hat{\beta}_1 = (X_1'M_2'M_2X_1)^{-1}(X_1'M_2'M_2y) = (\tilde{X}_1'\tilde{X}_1)^{-1}(\tilde{X}_1'\tilde{\varepsilon}_2),$$

where $\tilde{X}_1 = M_2X_1$ and $\tilde{\varepsilon}_2 = M_2y$.

Therefore, the coefficient estimator $\hat{\beta}_1$ is algebraically equal to the least squares regression of $\tilde{\varepsilon}_2$ on $\tilde{X}_1$. The conclusion follows since M_2X_1 are the residuals from a regression of X_2 on X_1, and M_2y are the residuals of a regression of X_2 on y.

Now, for the residuals, in the case of OLS, we have that $\hat{\varepsilon} = y - X\hat{\beta} = y - P_Xy = M_Xy$. Notice that it can be written as $y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2$. Similarly, we have that $\tilde{\varepsilon} = \tilde{\varepsilon}_2 - \tilde{X}_1\hat{\beta}_1 = M_2y - M_2X_1\hat{\beta}_1$. If we pre-multiply $\hat{\varepsilon}$ by M_2, we get

$$M_2\hat{\varepsilon} = M_2y - M_2X_1\hat{\beta}_1 - M_2X_2\hat{\beta}_2$$

Notice that by properties of the annihilator matrix, for any matrix X, $M_XX = 0$. It is easy to see since $M_XX = (I - P_X)X = 0$ as $P_XX = X$. Therefore, we have that $M_2X_2\hat{\beta}_2 = 0$, which implies that $M_2\hat{\varepsilon} = M_2y - M_2X_1\hat{\beta}_1$. The conclusion follows by orthogonality condition $X'\hat{\varepsilon} = 0 \implies X_2'\hat{\varepsilon} = 0$:

$$M_2\hat{\varepsilon} = \hat{\varepsilon} - X_2(X_2'X_2)^{-1}X_2'\hat{\varepsilon} = \hat{\varepsilon}$$

We have hence proven the following theorem.

Theorem 1.1: Frisch–Waugh–Lovell Theorem (FWL)

The OLS estimator of β_1 and the OLS residuals can be computed by either the OLS regression or using the following algorithm:

1. Regress y on X_2, obtain $\tilde{\varepsilon}_2$
2. Regress X_1 on X_2, obtain residuals $\tilde{X}_1$
3. Regress $\tilde{\varepsilon}_2$ on $\tilde{X}_1$, and obtain OLS estimator $\hat{\beta}_1$ and residuals $\hat{\varepsilon}$

Additionally, the residual-based regression of $M_2 y$ on $M_2 X_1$ yields residuals that are numerically equivalent to those obtained from regressing y on $(X_1 : X_2)$.

We are giving some intuition about the theorem below.

1. Regress y on X_2, obtain $\tilde{\varepsilon}_2$. This means that we are calculating the part of y which is not explained by X_2. In other words, we are projecting y onto the space which is orthogonal to X_2, so we are projecting X_2 out of y.
2. Regress X_1 on X_2, obtain residuals $\tilde{X}_1$. Again, we are computing the part of X_1 which is not explained by X_2, or we are projecting X_1 onto the space which is orthogonal to X_2.
3. Regress $\tilde{\varepsilon}_2$ on $\tilde{X}_1$, and obtain OLS estimator $\hat{\beta}_1$ and residuals $\hat{\varepsilon}$. The OLS estimator for the final regression is $\hat{\beta} = (\tilde{X}_1' \tilde{X}_1)^{-1} \tilde{X}_1 \tilde{\varepsilon}_2 = (X_1' M_2' M_2 X_1)^{-1} (X_1' M_2' M_2 y)$

1.5 Finite-Sample Properties

1.5.1 Unbiasedness

Recall our model $y = X\beta + \varepsilon$, which satisfies properties (i) to (v) (Gauss Markov + normality). Our estimator is therefore $\hat{\beta} = (X'X)^{-1} X'y$. It is important to understand that $\hat{\beta}$ is a random variable. We can now compute the conditional mean given X.

$$\mathbb{E}[\hat{\beta} \mid X] = (X'X)^{-1} X' \mathbb{E}[y \mid X] = (X'X)^{-1} X' X \hat{\beta} + (X'X)^{-1} X' \mathbb{E}[\varepsilon | X] = \beta \qquad (1.4)$$

The last equality comes from assumption (iii). I like the interpretation Hansen has of this result. He says that the estimator $\hat{\beta}$ is unbiased for β, conditional on X means that the conditional distribution of $\hat{\beta}$ is centered at β. Conditional on X means that the distribution is unbiased for any realisation on the regressor matrix X. Additionally, notice that conditional unbiasedness implies unconditional unbiasedness by LIE:

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}[\mathbb{E}[\hat{\beta} \mid X]] = \beta$$

Given (i) to (iv), we can define an estimator that is an unbiased estimator of σ^2: $s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$ This estimator gives rise to the standard error of the regression, s, which is the square root of s^2. Recall that $\hat{\varepsilon} = y - \hat{y}$. (i) and (ii) yield

$$\hat{\varepsilon} = M_X y = M_X (X\beta + \varepsilon) = M_X \varepsilon, \text{ hence}$$

$$\hat{\varepsilon}'\hat{\varepsilon} = \varepsilon' M_X \varepsilon \quad \text{with} \quad M_X = I - X(X'X)^{-1} X$$

then: a) first equality, we divide by $n-k$ to adjust biases (see at the end), b) second equality, we introduce the trace operator, c) third equality exploit the linearity of tr and E

$$\mathbb{E}(s^2|X) = \frac{1}{n-k}\mathbb{E}(\varepsilon' M_X \varepsilon|X) = \frac{1}{n-k}E[tr(\varepsilon' M_X \varepsilon)|X] = \frac{1}{n-k}tr[M_X(\mathbb{E}(\varepsilon'\varepsilon|X)]$$

recalling A.3 and A.4

$$\mathbb{E}(s^2|X) = \frac{1}{n-k}tr[M_X\sigma^2 I] = \frac{\sigma^2}{n-k}tr(M_X)$$

We can show that

$$tr(M_X) = tr(I_n - X(X'X)^{-1}X')$$
$$tr(I_n) - tr(X(X'X)^{-1}X') = tr(I_n) - tr(I_k) = n-k$$

then we can simplify

$$E(s^2|X) = \frac{1}{n-k}tr[M_X\sigma^2 I] = \frac{\sigma^2}{n-k}n - k = \sigma^2$$

s^2 is conditionally unbiased. Using the law of iterated expectations, we can show that s^2 is unconditionally unbiased too.

1.5.2 Variance

Given observations (i) to (iii), we have that the conditional variance is given by:

$$\mathbb{V}(\hat{\beta}|X) = \mathbb{E}\left[(\hat{\beta} - \mathbb{E}[\hat{\beta}|X](\hat{\beta} - \mathbb{E}[\hat{\beta}|X])' \mid X\right]$$
$$= \mathbb{E}\left[(\beta + (X'X)^{-1}X'\varepsilon - \beta)((\beta + (X'X)^{-1}X'\varepsilon - \beta)])' \mid X\right]$$
$$= \mathbb{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon' X(X'X)^{-1}\right]$$

This is the general variance-convariance matrix where the diagonal elements are $\mathbb{V}(\hat{\beta}_k|X)$ and the off-diagonal elements are $\mathbb{C}\text{ov}(\hat{\beta}_k, \hat{\beta}_\ell|X)$. If we assume spherical errors (iv), $\varepsilon\varepsilon' = \sigma^2$, and the variance becomes

$$\mathbb{V}(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$$

For X non-stochastic, $\mathbb{V}(\hat{\beta}|X) = \mathbb{V}(\hat{\beta})$. If X is stochastic, we have

$$\mathbb{V}(\hat{\beta}) = \mathbb{E}(\mathbb{V}(\hat{\beta}|X)) + \mathbb{V}(\mathbb{E}(\hat{\beta}|X)) = \sigma^2\mathbb{E}\left[(X'X)^{-1}\right]$$

The last equality comes from the fact that β is just a number, so has variance 0.
The variance can be estimated as follows

$$\widehat{\mathbb{V}}(\hat{\beta}) = s^2(X'X)^{-1}, \text{ where } s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$$

The standard error of the regression is simply the square root of s^2. The standard of $\hat{\beta}_j$ is

$$SE(\hat{\beta}_j) = \sqrt{\widehat{\mathbb{V}}(\hat{\beta})} = \sqrt{s^2(X'X)^{-1}_{jj}}$$

Example 1.2: Variance in simple regression model

Consider the simple regression

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid X] = 0,\ \mathbb{V}(\varepsilon_i \mid X) = \sigma^2.$$

Conditional on X,

$$\mathbb{V}(\hat{\beta} \mid X) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} := \frac{\sigma^2}{(n-1)S_x^2},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $S_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ is the sample variance of x.

The key takeways here are:

1. The $(n-1)$ in the denominator highlights that larger samples reduce the variance.
2. The S_x^2 in the denominator highlights that greater variability of the regressors reduce the variability of our estimators.

We can now state one of the most important theorems so far.

Theorem 1.2: Gauss Markov Theorem

Let assumptions (i) - (iv) to hold. Then,

- The least square estimator $\hat{\beta}$ is the minimum variance linear unbiased estimator.
- For any vector of constants w, the mimimum variance linear unbiased estimator of $w'\beta$ is $w'\hat{\beta}$.

Proof. Let $\tilde{\beta}$ a linear estimator where $\tilde{\beta} = Cy$. Unbiasedness requires

$$\mathbb{E}[\tilde{\beta} \mid X] = CX\beta + \mathbb{E}(\varepsilon|X) = CX\beta =^{!} \beta \implies CX = I_k.$$

Under spherical errors, $\mathbb{V}(y \mid X) = \sigma^2 I_n$, so

$$\mathbb{V}(\tilde{\beta} \mid X) = \sigma^2 CC', \qquad \mathbb{V}(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}.$$

Write the residual-maker $M_X := I_n - P_X$ with $P_X := X(X'X)^{-1}X'$. Note M_X is symmetric, idempotent, and positive semidefinite, and $M_X X = 0$.

Use the decomposition

$$C = (X'X)^{-1}X' + CM_X$$

(which holds because multiplying both sides by X yields I_k). Then

$$\begin{aligned}\mathbb{V}(\tilde{\beta} \mid X) &= \sigma^2\left[(X'X)^{-1}X' + CM_X\right]\left[(X'X)^{-1}X' + CM_X\right]' \\ &= \sigma^2\Big\{(X'X)^{-1}X'X(X'X)^{-1} \;+\; CM_XM_XC'\Big\} \quad \text{(cross terms vanish since } X'M_X = 0\text{)} \\ &= \sigma^2(X'X)^{-1} \;+\; \sigma^2 CM_XC'.\end{aligned}$$

Hence

$$\mathbb{V}(\tilde{\beta} \mid X) - \mathbb{V}(\hat{\beta} \mid X) = \sigma^2\, CM_XC' \geq_{psd} 0,$$

because M_X is positive semidefinite. Therefore OLS has the smallest variance among all linear unbiased estimators.

Finally, for any fixed $w \in \mathbb{R}^k$,

$$\mathbb{V}(w'\tilde{\beta} \mid X) - \mathbb{V}(w'\hat{\beta} \mid X) = w'\left[\mathbb{V}(\tilde{\beta} \mid X) - \mathbb{V}(\hat{\beta} \mid X)\right]w = \sigma^2\,(C'w)'M_X(C'w) \geq_{psd} 0,$$

so every linear combination $w'\hat{\beta}$ is minimum-variance among linear unbiased estimators. □

1.6 Asymptotic Properties of OLS

1.6.1 Probability Theory foundations

We start by defining concepts that will be useful in our treatment of asymptotic theory. First, we need to define a (non-exhaustive) list of modes of convergence for random variables.

Definition 1.10: Convergence of random variables

Consider a sequence X_n of random variables.

- The strongest mode of convergence we mention here is almost sure convergence. X_n is said to convergence a.s. to X (or $X_n \xrightarrow{a.s.} X$) if for every $\epsilon > 0$,

$$\mathbb{P}\left(\limsup_{n\to\infty}(|X_n - X|) > \epsilon\right) = 0$$

- X_n is said to converge in probability to X (or $X_n \xrightarrow{p} X$) if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| > \epsilon\right) = 0$$

- Let $F_{X_n}(x)$ the CDF of X_n and $F_X(x)$ the CDF of X. X_n is said to converge in probability to X (or $X_n \xrightarrow{d} X$) if for every point of continuity of F,

$$F_{X_n}(x) \to F_X(x)$$

Now, we need to introduce two of the most important concepts in probability theory: laws of large numbers (LLN) and central limit theorems (CLT). We start by three LLNs. The first one is the one we will mostly use. It has the weakest assumptions, but turns out to be hard to prove. I will therefore prove Chebyshev's (like) WLLN, which will be useful later once we relax the spherical error assumption.

Theorem 1.3: Khinchin's Weak Law of Large Number

Suppose that $X_i, i \geq 1$ is an i.i.d. sequence of random variables with a finite first moment (i.e., $\mathbb{E}(|X_i|) < \infty$). Let $\mu = \mathbb{E}(X_i)$.Then, the sample mean converges in probability to the expected value:

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

Now, we can state a stronger statement.

Theorem 1.4: Kolmogorov's Law of Large Numbers

Suppose that $X_i, i \geq 1$ is an i.i.d. sequence of random variables with common mean $\mu = \mathbb{E}(X_1)$.The sample mean converges almost surely to the expected value,

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

if and only if the expected value is finite:

$$\mathbb{E}(|X_1|) < \infty$$

Finally, we state the Chebyshev's WLLN, which uses assumption on the second moment.

Theorem 1.5: Chebyshev's WLLN

Let $\{X_n\}$ be a sequence of uncorrelated random variables with finite means $\mu_i = \mathbb{E}[X_i]$ and such that

$$\lim_{n\to\infty} \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(X_i) = 0.$$

Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i \xrightarrow{p} 0.$$

Equivalently,

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

Proof. Let $\bar{X}_n = \frac{1}{n}\sum_i X_i$. First, notice that

$$\mathbb{E}\left(\bar{X}_n\right) = \frac{1}{n}\sum_i \mathbb{E}(X_i)$$

Additionally, we can write the variance as

$$\mathbb{V}(\bar{X}_n) = \mathbb{V}\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{C}\text{ov}(X_i, X_j) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i) \to 0$$

where the last equality comes from uncorrelatedness and convergence comes by assumptions. Now, by Chebyshev's Inequality, we know that for some $\varepsilon > 0$

$$\mathbb{P}\left(\mid \bar{X}_n - \frac{1}{n}\sum_i \mathbb{E}(X_i) \mid > \varepsilon\right) \leq \frac{\mathbb{V}(\bar{X}_n)}{\varepsilon^2} \to 0$$

□

Corollary 1.1: Chebyshev's WLLN for Identical Means

Under the assumptions of Theorem 1.5, suppose in addition that

$$\mathbb{E}[X_i] = \mu \qquad \text{for all } i.$$

Then

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

Proof. By Theorem 1.5,

$$\frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] + o_p(1).$$

If $\mathbb{E}[X_i] = \mu$ for all i, then

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \mu,$$

so the conclusion follows.

□

The condition on the second moments stated in 1.5 is actually much stronger than what we need. We can introduce the concept of uniform integrability to find weaker sufficient conditions. This is not directly relevant for the course, so I show that in Appendix A.1.

We can now state the most standard CLT, which applies only to iid data.

Theorem 1.6: Lindeberg-Levy's Central Limit Theorem

Suppose that $X_i, i \geq 1$ is an i.i.d. sequence of random variables with common finite mean $\mu = \mathbb{E}(X_1)$ and common finite, positive variance $\sigma^2 = \text{Var}(X_1) < \infty$. Then,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Now, if we face a case where we don't have iid data (for instance if we have heteroskedasticity), we need to use another CLT.

Theorem 1.7: Lindeberg-Feller's Central Limit Theorem

Let $\{X_n\}$ be an integrable sequence of independent random variables with $0 < \mathbb{V}(X_i) < \infty$ for all n. Let $\mathbb{E}(X_i) = \mu_i$ and $\mathbb{V}(X_i) = \sigma_i^2$. Now, let

$$C_n := \sqrt{\sum_i \sigma_i^2}, \text{ where } \lim_{n \to \infty} \max_{i \in [1,n]} \frac{\sigma_i^2}{C_n^2} = 0$$

Then, for any $\varepsilon > 0$,

$$\frac{1}{C_n} \sum_i (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0,1) \iff \lim_{n \to \infty} \frac{1}{C_n^2} \sum_i \mathbb{E}\left[(X_i - \mu_i)^2 \mathbb{1}\{|X_i - \mu_i| \geq \varepsilon C_n\} \right] = 0$$

The condition on the right-hand side of the equivalence operator is known as the **Lindeberg Condition**. This condition requires that the sum of the extreme-value variances (the numerator) becomes negligible compared to the total variance C_n^2 (the denominator) as $n \to \infty$.

Now, we can state two other important results that will come handy in our applications.

Theorem 1.8: Continuous Mapping Theorem

Let $\{X_n\}$ be a sequence of random variables (or vectors) and X be a random variable. Let g be a function that is continuous
Then the following hold:

- If $X_n \xrightarrow{a.s.} X$ then $g(X_n) \xrightarrow{a.s.} g(X)$
- If $X_n \xrightarrow{p} X$ then $g(X_n) \xrightarrow{p} g(X)$
- If $X_n \xrightarrow{d} X$ then $g(X_n) \xrightarrow{d} g(X)$

Theorem 1.9: Slutsky's Theorem

Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables. Suppose X_n converges in distribution to a random variable X, and Y_n converges in probability to a constant c:

$$X_n \xrightarrow{d} X \quad \text{and} \quad Y_n \xrightarrow{p} c$$

Then the following hold:

- Sum: $X_n + Y_n \xrightarrow{d} X + c$
- Product: $X_n Y_n \xrightarrow{d} cX$
- Quotient (if $c \neq 0$): $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$

Finally, we introduce objects called (stochastic) orders of magnitude, which can be useful in some applications.

Definition 1.11: Big and small oh

Consider (X_n) and (f_n) two sequences of real numbers.

1. We say that $X_n = O(f_n)$ if $\left|\frac{X_n}{f_n}\right| \to c < \infty$ (i.e. $\left|\frac{X_n}{f_n}\right|$ is bounded for all sufficiently large n)
2. We say that $X_n = o(f_n)$ if $\left|\frac{X_n}{f_n}\right| \to 0$ as $n \to \infty$

Notice that $X_n = O(1) \Leftrightarrow X_n$ is bounded uniformly in n, that is there exists $M < \infty$ such that $|X_n| \leq M$ for all n. Naturally, $X_n = o(1) \Leftrightarrow X_n \to 0$ as $n \to \infty$

Definition 1.12: Big and small oh-P

Consider (X_n) be a sequence of random variable and (f_n) a sequence of real numbers.

1. We say that $X_n = O_p(f_n)$ if $\forall \varepsilon > 0, \exists c \geq 0$ and $n_0 \in \mathbb{N}$ such that $\forall n \geq n_0$,
$$\mathbb{P}(|X_n| > c f_n) < \varepsilon$$
2. We say that $X_n = o_p(f_n)$ if $\frac{X_n}{f_n} \xrightarrow{p} 0$

Notice that the definition of $O_p(\cdot)$ is equivalent to the random variable being bounded in probability. If we have $X_n = O_p(1)$, we can equivalently write that for any $\varepsilon > 0$ there exists a constant $c < \infty$ such that

$$\limsup_{n \to \infty} \mathbb{P}(|X_n| > c) < \varepsilon$$

Finally, $X_n = o_p(1)$ is equivalent to $X_n \xrightarrow{p} 0$ as $n \to \infty$.

1.6.2 Consistency of the least squares estimator

Define the OLS estimator $\hat{\beta}$ and the estimation error:

$$\hat{\beta} = \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right)$$

To show that $\hat{\beta} \xrightarrow{p} \beta$ (Consistency), we need to show that $\left(\frac{X'X}{n}\right) \xrightarrow{p} Q$, where Q is finite and invertible, and $\left(\frac{X'\varepsilon}{n}\right) \xrightarrow{p} 0$.

First, consider the case where the data $\{(x_i, \varepsilon_i)\}_{i=1}^n$ is iid. Additionally, we assume contemporaneous exogeneity: $\mathbb{E}(x_i\varepsilon_i) = 0$, and we impose some moment conditions: $\mathbb{E}(x_i x_i') < \infty$ and $\mathbb{E}|x_i\varepsilon_i| < \infty$. Notice that $\mathbb{E}(x_i x_i')$ is invertible by assumption ii) (full rank). Therefore, by Khinchine's WLLN, we obtain that

$$\frac{1}{n}\sum_{i=1}^n x_i x_i' \xrightarrow{p} \mathbb{E}(x_i x_i') \equiv Q \text{ and } \frac{1}{n}\sum_{i=1}^n x_i\varepsilon_i \xrightarrow{p} \mathbb{E}(x_i\varepsilon_i) = 0$$

Note that Slutsky's Theorem guarantees

$$\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \xrightarrow{p} Q^{-1}$$

Therefore, by Continuous Mapping Theorem, we obtain

$$\hat{\beta} \xrightarrow{p} \beta + \mathbb{E}(x_i x_i')^{-1}\mathbb{E}(x_i\varepsilon_i) = \beta + O_p(1)\cdot o_p(1) = \beta$$

Notice that the assumptions imposed here are stronger than what we need to establish consistency. Indeed, only assuming that $\frac{1}{n}\sum_{i=1}^n x_i x_i' \xrightarrow{p} Q$ and imposing weak exogeneity is enough to guarantee consistency.

Weak exogeneity (or contemporaneous uncorrelation), written as $E(x_i\epsilon_i) = 0$, is the necessary population moment condition required to ensure OLS consistency. Strict exogeneity, written as $E(\varepsilon_i|X) = 0$ is a much stronger assumption, requiring the error term to be uncorrelated with all regressors in the sample, and is the condition needed to guarantee OLS unbiasedness. Since Strict Exogeneity implies Mean Independence ($\mathbb{E}(\varepsilon_i|x_i) = 0$), and Mean Independence in turn implies Weak Exogeneity via the Law of Iterated Expectations[2], the required condition for consistency is nested within the condition for unbiasedness, first stated in the Gauss-Markov Assumptions.

[2] $\mathbb{E}(x_i\varepsilon_i) = \mathbb{E}(\mathbb{E}(x_i\varepsilon_i|x_i)) = \mathbb{E}(x_i\mathbb{E}(\varepsilon_i|x_i)) = 0$ if $\mathbb{E}(\varepsilon_i|x_i) = 0$.

1.6.3 Asymptotic Normality of the least squares estimator

We are interested in understanding the behavior of $\sqrt{n}(\hat{\beta} - \beta)$. We can rewrite the sample counterpart of this object as follows

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \varepsilon_i\right)$$

Consider the set of assumptions we imposed for consistency, where naturally, using WLLN and Slutsky's Theorem, we have

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \xrightarrow{p} Q^{-1}$$

Now, we need to add two additional moment assumptions: $\mathbb{E}(\|x_i\|^4) < \infty$ and $\mathbb{E}(\varepsilon_i^4) < \infty$.[3] We do not specify any structure on the errors yet. Under these assumptions, by Lindeberg-Lévy's CLT[4], we have that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{d} \mathcal{N}(0, \Omega^*), \text{ where } \Omega^* = \mathbb{E}(\varepsilon_i^2 x_i x_i')$$

By CMT, we get that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} Q^{-1}\mathcal{N}(0, \Omega^*) = \mathcal{N}(0, Q^{-1}\Omega^* Q^{-1})$$

Now, suppose that we impose spherical errors. We can write the following, using LIE in the first equality

$$\mathbb{V}(x_i \varepsilon_i) = \mathbb{E}\left(\mathbb{V}(x_i \varepsilon_i | X)\right) = \mathbb{E}\left(x_i x_i' \mathbb{V}(\varepsilon_i | X)\right) = \sigma^2 \mathbb{E}(x_i x_i') = \sigma^2 Q$$

Therefore, we have that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q^{-1} Q Q^{-1}) = \mathcal{N}(0, \sigma^2 Q^{-1})$$

The $\sqrt{n}$ normalization is used because it is the unique scaling factor that keeps the distribution of the estimator error "alive" for asymptotic analysis. The variance of the OLS estimator, $\mathbb{V}(\hat{\beta})$, is known to decrease at the rate of $\frac{1}{n}$ as the sample size n increases (e.g., in the simple case, $\mathbb{V}(\hat{\beta}) \approx \frac{\sigma^2}{n \cdot \mathbb{V}(X_i)}$). Because the variance collapses to zero at rate n, the standard deviation collapses at rate $\sqrt{n}$; consequently, multiplying the estimation error $(\hat{\beta} - \beta)$ by $\sqrt{n}$ exactly counteracts this collapse, yielding a distribution that converges to a non-degenerate Normal distribution.

1.7 Quick detour on intuition & collinearity

[3] These two assumptions are imposed to have a finite variance. Indeed, using Cauchy-Schwarz Inequality, it is easy to show that $\mathbb{E}(|\varepsilon_i x_i|)^2 \leq \sqrt{\mathbb{E}(\|x_i\|^4)}\sqrt{\mathbb{E}(\varepsilon_i^4)} < \infty$

[4] Note that Lindeberg-Lévy's CLT applies to the term $\frac{1}{\sqrt{n}}\sum x_i \varepsilon_i$ when the data is iid. We use the general notation Ω here, which is applicable even under heteroskedasticity (provided iid still holds). We will discuss later the Lindeberg-Feller condition for non-iid cases.

Remark 1.1: Some additional intuition on the variance

Consider the following expression for the variance

$$\mathbb{V}(\hat{\beta}) \approx \frac{\sigma^2}{n \cdot \mathbb{V}(X_i)}$$

This expression makes it clear *why* we want a lot of variation in our independent variables. Since $\mathbb{V}(\hat{\beta}) \propto 1/\mathbb{V}(X_i)$, the higher $\mathbb{V}(X_i)$, the lower the variance of the estimator.

This remark introduces the concept of multicollinearity, which is generally written as $rank(X) < k$, where k is the number of regressors. Note that this is saying that at least some columns of X are linearly dependent. Formally, this implies that $X'X$ is not invertible, and therefore the OLS estimates cannot be computed. A typical example is to consider a simple regression of wage on age, schooling and experience. Since experience is a combination of age and years of schooling, we have collinearity, and can't run OLS.

More formally, the uniqueness of the OLS solution and the invertibility of $X'X$ are directly tied to the rank of X. The rank of an $n \times K$ matrix X, denoted $\text{rank}(X)$, is the number of linearly independent columns (or rows) it contains. The condition for a unique OLS solution is that X must be full column rank, meaning $\text{rank}(X) = K$. If $\text{rank}(X) < K$, the columns of X are linearly dependent (perfect multicollinearity), which means the column space $\mathcal{R}(X)$ has a dimension less than K, and consequently, the $K \times K$ matrix $X'X$ is singular (non-invertible) according to Proposition 1.3. Without an invertible $X'X$, the OLS formula $\hat{\beta} = (X'X)^{-1}X'y$ cannot be computed, and the fitted vector $\hat{y}$ would be defined by a non-unique set of coefficients $\hat{\beta}$.

1.8 Function of Parameters

In most applications, a researcher is interested in a specific transformation of the coefficient vector β[5]. For example, one may be interested in a single coefficient β_j, or a ratio β_j/β_l. More generally, we can write the parameter of interest θ as a function of the coefficients: $\theta = r(\beta)$ for some function $r : \mathbb{R}^k \to \mathbb{R}^q$ with $k \geq q$. The estimate of θ is therefore

$$\hat{\theta} = r(\hat{\beta})$$

Now, suppose that the conditions for consistency hold, $r(\cdot)$ is continuously differentiable in a neighborhood of (the true) β and $R := \frac{\partial r(\beta)'}{\partial \beta}$ has rank q. We now introduce the Delta method. In this context, it basically says that if the transformation is smooth enough, we can show $\hat{\theta}$ is asymptotically normal.

[5]Note that the next two parts are not directly part of the first year course at Bocconi, but are in my opinion very important to build towards hypothesis testing.

Theorem 1.10: Delta Method

Let $\mu \in \mathbb{R}^k$ and $g : \mathbb{R}^k \to \mathbb{R}^J$. If $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \xi$, where $g(u)$ is continuously differentiable in a neighborhood of μ, then as $n \to \infty$

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \xrightarrow{d} \mathbf{G}'\xi$$

where $\mathbf{G}(u) = \frac{\partial g(u)}{\partial u'}$ and $\mathbf{G} = \mathbf{G}(\mu)$.

In particular, if $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ then as $n \to \infty$

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{GVG}')$$

In this context, we can write the following Taylor expansion

$$\hat{\theta} = r(\hat{\beta}) \approx r(\beta) + \frac{\partial r(\beta)}{\partial \beta'}\Big|_{\beta=\tilde{\beta}}(\hat{\beta} - \beta) \text{ where } \tilde{\beta} \text{ is on the line joining } \beta \text{ and } \hat{\beta}$$

Rewriting this, and normalizing by $\sqrt{n}$, we get

$$\sqrt{n}(r(\hat{\beta}) - r(\beta)) \approx \underbrace{\frac{\partial r(\beta)}{\partial \beta'}\Big|_{\beta=\tilde{\beta}}}_{\xrightarrow{p}\mathbf{R}} \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, R'VR) \text{ where } V = Q^{-1}\Omega Q^{-1}$$

Notice that the key step relies on the fact that $\hat{\beta} \xrightarrow{p} \beta$, which implies $\tilde{\beta} \xrightarrow{p} \beta$ and since the partial derivatives are continuous, the Jacobian converges in probability:

$$\frac{\partial r(\beta)}{\partial \beta'}\Big|_{\beta=\tilde{\beta}} \xrightarrow{p} \frac{\partial r(\beta)}{\partial \beta'}\Big|_{\beta=\beta} \equiv R$$

Once that is derived, we can estimate the asymptotic variance of $\hat{\theta}$ by $\hat{R}'\hat{V}\hat{R}$ where we can show that a simple estimate of V is consistent, and $\hat{R} = \frac{\partial r(\hat{\beta})}{\partial \beta'}$.

1.9 Asymptotic standard errors

If $\hat{\theta}$ is a scalar, the standard error of $\hat{\theta}$ is obtained as

$$s(\hat{\theta}) = \sqrt{\frac{1}{n}\hat{R}'\hat{V}\hat{R}}$$

A standard error is an estimator of the standard deviation of the sampling distribution of an estimator. Intuitively, it measures the precision of the estimate, quantifying the typical distance between your sample-based estimate $\hat{\theta}$ and the true population parameter θ. If $\hat{V}$ is an estimator of the variance-covariance matrix of $\hat{\beta}$, then standard errors are the square rooots of the diagonal elements of the matrix:

$$s(\beta_j) = \sqrt{\frac{1}{n}\hat{V}_{jj}}$$

1.10 On the way to hypothesis testing: object construction

In this last subsection, I will construct the objects which will then be useful to proceed with hypothesis testing.

1.10.1 T-statistic

Again, suppose that $\hat{\theta}$ is a scalar. Based on the standard error $s(\hat{\theta})$, we can obtain the standardized object called t-statistic or t-ratio

$$T_n(\theta) = \frac{\hat{\theta} - \theta}{s(\hat{\theta})}$$

Now, by asymptotic distribution of $\hat{\theta}$, we can show that this statistic is *asymptotically pivotal*

$$T_n(\theta) = \frac{\hat{\theta} - \theta}{\sqrt{\frac{1}{n}\hat{R}'\hat{V}\hat{R}}} = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\hat{R}'\hat{V}\hat{R}}} \xrightarrow{d} \frac{\mathcal{N}(0, R'VR)}{\sqrt{R'VR}} = \frac{\sqrt{R'VR}}{\sqrt{R'VR}}\mathcal{N}(0,1) = \mathcal{N}(0,1)$$

Asymptotically pivotal means that as $n \to \infty$, the statistic behaves predictably and does not depend on the unknown θ. This must not be true in finite samples, but this property simply says that the dependence on unkowns diminishes as n increases.

1.10.2 Confidence Intervals

One way to estimate the unknown parameter $\theta \in \mathbb{R}$ is to find a point estimator $\hat{\theta}$. If one is not interested in the precise value of θ, it may be reasonable to estimate θ by an interval $[L_n, U_n] \subset \mathbb{R}$ (called confidence interval) such that

$$\mathbb{P}(\theta \in [L_n, U_n]) \approx 1 - \alpha \text{ for some } \alpha \in (0,1)$$

The goal is to set the coverage probability equal to a pre-specified target such 95%, with $\alpha = 0.05$. Since we rely upon asymptotic approximation, here we consider the interval $[L_n, U_n]$ such that

$$\mathbb{P}(\theta \in [L_n, U_n]) \to 1 - \alpha, \text{ called the } 100(1-\alpha)\% \text{ asymptotic confidence interval}$$

One common way to construct the asymptotic confidence interval of θ is to base it upon the asymptotically pivotal object

$$T_n(\theta) = \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \xrightarrow{d} \mathcal{N}(0,1)$$

By the asymptotic distribution, for a desired confidence level $1 - \alpha$, we have

$$\mathbb{P}(-z_{\alpha/2} \leq T_n(\theta) \leq z_{\alpha/2}) \to 1 - \alpha$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of $\mathcal{N}(0,1)$. This is intuitive: the object we consider converges in distribution to a standard normal, so the probability of the statistic

falling between $-z_{\alpha/2}$ and $z_{\alpha/2}$ approaches $1-\alpha$ as the sample size increases.

By isolating the true parameter θ, we get:

$$\mathbb{P}\left(\hat{\theta} - z_{\alpha/2}s(\hat{\theta}) \leq \theta \leq \hat{\theta} + z_{\alpha/2}s(\hat{\theta})\right) \to 1-\alpha$$

Thus, the $100(1-\alpha)\%$ asymptotic confidence interval of θ is:

$$\left[\hat{\theta} - z_{\alpha/2}s(\hat{\theta}), \quad \hat{\theta} + z_{\alpha/2}s(\hat{\theta})\right]$$

Notice that in this context (frequentist), we treat the confidence interval as a function of the data and hence random, while θ is fixed.

1.10.3 Wald Statistic

Now, consider the case where $\theta \in \mathbb{R}^q$. We want to find the asymptotically pivotal object. Instead of the t-ratio, consider the quadratic form

$$\sqrt{n}(\hat{\theta} - \theta)' A \sqrt{n}(\hat{\theta} - \theta)$$

If we set $A = \hat{V}_\theta^{-1} \equiv \hat{R}'\hat{V}\hat{R}$ we can derive the Wald statistic (asymptotically pivotal) using the CMT

$$W_n(\theta) = \sqrt{n}(\hat{\theta} - \theta)'\hat{V}^{-1}\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}(\hat{\theta} - \theta)'(\hat{R}'\hat{V}\hat{R})^{-1}\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_q$$

The confidence region for vector θ is given by $\{\theta : W_n(\theta) \leq \chi^2_{q,\alpha}\}$ where $\chi^2_{q,\alpha}$ is the $100(1-\alpha)\%$ percentile of χ^2_q.

We have now all the tools to move on to hypothesis testing. Yay!

2 Hypothesis Testing

2.1 Concepts

For parameter of interest $\theta = r(\beta)$, consider two-sided testing problem

$$\mathcal{H}_0 : \theta = \theta_0 \quad \text{against} \quad \mathcal{H}_1 : \theta \neq \theta_0$$

where θ_0 is a hypothetical value. $\mathcal{H}_0$ is called null hypothesis and $\mathcal{H}_1$ is called alternative hypothesis. In hypothesis testing, we assume that there is a true but unknown value of the parameter of interest θ, and this value either satisfies $\mathcal{H}_0$ or does not. The goal is therefore to assess whether or not $\mathcal{H}_0$ is true by asking if $\mathcal{H}_0$ is consistent with the observed data. The question is: Are the true coefficients zero? To answer this question the testing method asks the question: Are the observed estimates compatible with the hypothesis, in the sense that the deviation from the hypothesis can be reasonably explained by stochastic variation? Or are the observed estimates incompatible with the hypothesis, in the sense that that the

observed estimates would be highly unlikely if the hypothesis were true (Hansen, 2022)?

Let us consider a test in the form of

$$\begin{cases} \text{Accept } \mathcal{H}_0 & \text{if } T_n \leq c \\ \text{Reject } \mathcal{H}_0 & \text{if } T_n > c \end{cases}$$

where T_n is test the statistic and c the critical value.

In testing, we have two kinds of errors:

	Accept	**Reject**
$\mathcal{H}_0$ true	Correct	Type I
$\mathcal{H}_1$ true	Type II	Correct

Type I error probability can we written as

$$P(\text{Reject}|\mathcal{H}_0 \text{ true}) = P(T_n > c|\theta = \theta_0)$$

and represents false positives, e.g. an innocent person is convicted. Type II error probability can be written as

$$P(\text{Accept}|\mathcal{H}_1 \text{ true}) = P(T_n \leq c|\theta \neq \theta_0)$$

and represents false negative, e.g. a criminal is not convicted.

2.1.1 Optimality and The Neyman-Pearson Lemma

The primary goal in constructing a test is to manage the trade-off between the two error types. In classical testing, we choose to control the Type I error probability (the size of the test) at a fixed maximum level, α. We can choose c to satisfy

$$\mathbb{P}(\text{Type I Error}) \to \alpha$$

Given this fixed size, we then seek to find a test that minimizes the Type II error probability, which is equivalent to maximizing the power function, defined as $\pi(\theta) = \mathbb{P}(\text{Reject } \mathcal{H}_0|\mathcal{H}_1 \text{ true})$.

The **Neyman-Pearson (NP) Lemma** provides the theoretical foundation for this approach, stating that the most powerful test for a simple null hypothesis against a simple alternative is based on the **Likelihood Ratio (LR)**. This establishes the LR test as the theoretical benchmark for test optimality, and all other tests, such as the Wald and LM tests, are evaluated by their asymptotic equivalence to the LR test. More on the trinity of hypothesis testing later.

The UMP test can be written as follows $\varphi(x) = \mathbb{1}\{LR < k\}$ for some $k \in (0, 1)$ such that the probability of falsely rejecting the null is α. The likelihood ration can be written as

$$\Lambda = \frac{f_0(x)}{f_1(x)} = \frac{L(x_1, ..., x_n, \theta_0)}{L(x_1, ..., x_n, \theta_1)}$$

2.2 Linear hypotheses and joint significance

2.2.1 T-test

Based on the t-ratio constructed above, we can test $\mathcal{H}_0 : \theta = \theta_0$, for which it holds that $T_n(\theta) \xrightarrow{d} \mathcal{N}(0,1)$. This implies that $\mathbb{P}(|T_n(\theta_0)| > z_{\alpha/2}|\mathcal{H}_0) \to \alpha$. Therefore, the asymptotic size α test is

$$\text{Accept } \mathcal{H}_0 \text{ if } |T_n(\theta_0)| \leq z_{\alpha/2} \quad \text{and} \quad \text{Reject } \mathcal{H}_0 \text{ if } |T_n(\theta_0)| > z_{\alpha/2}$$

Example 2.1: Single Linear Hypothesis

For a single linear hypothesis $\mathcal{H}_0 : \mathbf{c}'\beta = \gamma$, the test statistic is written as the T-test:

$$T = \frac{c'\hat{\beta} - \gamma}{\text{SE}(c'\hat{\beta})}$$

For the exact-sample test under the classical assumptions (spherical errors and normal errors $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$), the statistic follows a t-distribution:[a]

$$T \sim t_{n-k} \quad \text{under } \mathcal{H}_0$$

The corresponding rejection rule for the exact-sample two-sided test is to Reject $\mathcal{H}_0$ if $|T| > t_{n-k,\alpha/2}$.

[a] Recall that as the degrees of freedom increases, the t-distributions converge to the standard normal

In the exact-sample T-test, the test statistic follows the t-distribution with $n - k$ degrees of freedom, where n is the sample size and k is the total number of parameters estimated in the unrestricted model (including the intercept). The quantity $n - k$ represents the number of observations available to estimate the error variance after accounting for the k parameters that have been fit. It reflects the number of independent data points available for estimating the residual variation.

2.2.2 F-test

A joint linear hypothesis involves testing J restrictions on the parameter vector β simultaneously. We want to test $\mathcal{H}_0 : R\beta = c$ against $\mathcal{H}_1 : R\beta \neq c$, where R is a known $J \times K$ matrix of constants with full rank, where J is the number of restrictions, and c is a $J \times 1$ vector of known constants.

Example 2.2: Joint significance testing

Testing $\mathcal{H}_0 : \beta_2 = 1, \beta_3 = -1$ simultaneously for a model with $K = 4$ parameters:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

This is subject to $\mathcal{H}_0 : \beta_2 = 1, \beta_3 = -1$, which implies $R\beta = c$:

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

The test for joint hypotheses is based on verifying that the estimated restrictions $R\hat{\beta} - c$ are close to zero. This test utlimately uses the quadratic form we derived previously to obtain the following test statistic:

$$W = n(R\hat{\beta} - c)'[RVR']^{-1}(R\hat{\beta} - c) \xrightarrow{d} \chi^2_J$$

Notice that this can be done when σ^2 is known. If σ^2 is unknown, an alternative, exact-sample statistic is the F-test, which follows the $F_{J,n-k}$ distribution under $\mathcal{H}_0$ (under classical assumptions). As before, the F-distribution converges to a χ^2 distribution.

The most intuitive way to understand the F-test is by viewing it as a comparison between two models: the unrestricted model (OLS) and the restricted model (RLS) where the constraints $R\beta = c$ are imposed.

$$F = \frac{(RSS_R - RSS_{UR})/J}{RSS_{UR}/(n-k)} \sim F_{J,n-k} \quad \text{under } \mathcal{H}_0$$

The acceptance rule is to **Reject** $\mathcal{H}_0$ if $F \geq F_{(J,n-k),\alpha}$.

2.2.3 Restricted Least Squares Derivation

The RLS estimator $\tilde{\beta}$ is obtained by minimizing the sum of squared residuals subject to the linear constraints $R\beta = c$, using the Lagrangian method:

$$\mathcal{Q}(\beta, \lambda) = (y - X\beta)'(y - X\beta) - 2\lambda'(R\beta - c)$$

The first-order conditions (FOCs) are:

$$\frac{\partial \mathcal{Q}}{\partial \beta}\Big|_{\tilde{\beta},\tilde{\lambda}} = 0 \implies -2X'y + 2(X'X)\tilde{\beta} - 2R'\tilde{\lambda} = 0$$

$$\frac{\partial \mathcal{Q}}{\partial \lambda}\Big|_{\tilde{\beta},\tilde{\lambda}} = 0 \implies R\tilde{\beta} = c$$

Solving the system (see appendix) yields the RLS estimator $\tilde{\beta}$:

$$\tilde{\beta} = \hat{\beta} + (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(c - R\hat{\beta})$$

If the null hypothesis $\mathcal{H}_0 : R\beta = c$ is true, the RLS estimator $\tilde{\beta}$ is BLUE and is more efficient than the unrestricted OLS estimator $\hat{\beta}$.

$$Var(\hat{\beta}) - Var(\tilde{\beta}) = \sigma^2(X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}R(X'X)^{-1} \geq 0$$

The implication is the an OLS with a restriction that reflects the true values is always more efficient than an unrestricted OLS.

Example 2.3: Significance of a Regression

A question of interest can be to know whether a regression as a whole is significant

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$

The test is a joint test of the hypothesis that all coefficients except the constant are zero

$$\mathcal{H}_0 : \beta_2 = ... = \beta_k = 0$$

Notice that the number of restrictions here is $k-1$, and the degrees of freedom of the unrestricted model $n-k$. We can define RSS_R as the restricted RSS, regress $y = \beta_1 + \varepsilon$ (under $\mathcal{H}_0$), and use the residuals to compute $\tilde{\varepsilon}_R' \tilde{\varepsilon}_R = TSS$. The F-test is therefore computed as

$$F = \frac{(RSS_R - RSS_{UR})/k-1}{RSS_{UR}/(n-k)} = \frac{R^2/k-1}{(1-R^2)/n-k} \sim F_{k-1,n-k}$$

2.3 The Trinity of Testing

Before introducing the full Trinity of tests, we first need to establish the properties of the Maximum Likelihood Estimator (MLE), which forms the basis for the LR and LM statistics.

2.3.1 The ML Estimator

Let $f(y_i \mid \theta)$ be the probability density function (or probability mass function) of the data, where θ is the vector of parameters. As introduced above, the **Likelihood Function** for n independent observations is:

$$L(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$$

The **Log-Likelihood Function** is:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \ln f(y_i \mid \theta)$$

The **Maximum Likelihood Estimator (MLE)**, denoted $\hat{\theta}_{ML}$, is the value of θ that maximizes $\mathcal{L}(\theta)$. It is found by solving the first-order condition:

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\Big|_{\hat{\theta}_{ML}} = \mathbf{0}$$

Definition 2.1: Score

The vector of first derivatives of the log-likelihood function is called the **Score Function**:

$$\mathbf{S}(\theta) = \frac{\partial \mathcal{L}(\theta)}{\partial \theta}$$

Under regularity conditions, the expected value of the score function evaluated at the true parameter is zero: $\mathbb{E}[\mathbf{S}(\theta)] = \mathbf{0}$.

Definition 2.2: Hessian

The matrix of second derivatives of the log-likelihood function is the **Hessian Matrix**:

$$\mathbf{H}(\theta) = \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta \partial \theta'}$$

Definition 2.3: Information Matrix

The **Information Matrix** is defined as the negative expected value of the Hessian:

$$\mathbf{I}(\theta) = -\mathbb{E}[\mathbf{H}(\theta)] = -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta \partial \theta'}\right]$$

Under regularity conditions[6], the MLE is consistent, asymptotically efficient, and asymptotically normally distributed:

$$\sqrt{n}(\hat{\theta}_{ML} - \theta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V})$$

The asymptotic covariance matrix $\mathbf{V}$ is the inverse of the Information Matrix:

$$\mathbf{V} = \mathbf{I}(\theta)^{-1}$$

Remark 2.1: The Fisher Information Identity

The Information Matrix can also be written as the expected outer product of the score function:

$$\mathbf{I}(\theta) = \mathbb{E}[\mathbf{S}(\theta)\mathbf{S}(\theta)']$$

In practice, we use the estimated (asymptotic) variance $\widehat{AVar}(\hat{\theta}_{ML})$. This can be obtained in two common ways by using different estimators for the Information Matrix $\mathbf{I}(\theta)$:

- Using the outer product of the scores (this estimates the Information Matrix using the empirical version of the Fisher Identity)

$$\widehat{AVar}(\hat{\theta}_{ML}) = \left[\sum_{i=1}^{n} \mathbf{S}_i(\hat{\theta}_{ML})\mathbf{S}_i(\hat{\theta}_{ML})'\right]^{-1}$$

[6] More on this when we introduce extremum estimators.

- Using the negative of the Hessian (this estimates the Information Matrix using the observed (actual) Hessian from the maximum point):

$$\widehat{AVar}(\hat{\theta}_{ML}) = [-\mathbf{H}(\hat{\theta}_{ML})]^{-1}$$

With that, we can turn to the tests.

2.3.2 Wald Test

The Wald test measures the distance between the **unrestricted estimate** $\hat{\theta}$ and the value imposed by the null hypothesis. Consider $\mathcal{H}_0 : r(\theta) = r(\theta_0)$.

$$W = n \cdot \left(r(\hat{\theta}) - r(\theta_0)\right)' \left[R(\hat{\theta})'\hat{V}(\theta)R(\hat{\theta})\right]^{-1} \left(r(\hat{\theta}) - r(\theta_0)\right) \xrightarrow{d} \chi^2_J$$

A problem with the Wald test is that it is not invariant to the formulation of the restrictions. For instance, if we test $\mathcal{H}_0 : \beta_1\beta_2 = 1$ and $\mathcal{H}'_0 : \beta_1 = 1/\beta_2$, the Wald test provides different answers in finite sample, even if the two are asymptotically equivalent. Additionally, it has a tendency to over-reject.

To build additional intuition, consider the univariate case

$$W_{univ} = \frac{(\hat{\theta} - \theta_0)^2}{Var(\hat{\theta})} \sim^{\mathcal{H}_0} \chi^2_1$$

The numerator $(\hat{\theta} - \theta_0)^2$ measures the squared difference between our best estimate, $\hat{\theta}$, and the value hypothesized by the null, θ_0. Higher deviation means stronger evidence against $\mathcal{H}_0$. The denominator $\widehat{Var}(\hat{\theta})$ measure precision. If we have a low variance it indicates a highly precise estimate (low $\widehat{Var}(\hat{\theta})$). A small deviation in the numerator is then magnified by the small denominator, resulting in a large W_{univ}. We are confident in $\hat{\theta}$, so we strongly reject $\mathcal{H}_0$. On the other hand, a flat likelihood function indicates an imprecise estimate (high $\widehat{Var}(\hat{\theta})$). The denominator is large, making it harder to reject $\mathcal{H}_0$. The data does not provide enough information to reliably distinguish $\hat{\theta}$ from θ_0.

2.3.3 The Lagrange Multiplier or Score Test

The LM test measures how close the score function $S(\theta)$ is to zero when evaluated at the **restricted estimate** $\tilde{\theta}_R$. If $\mathcal{H}_0$ is true, $\tilde{\theta}_R$ should be close to the true parameter, and the score should be near zero. It relies only on the restricted estimation.

$$LM = \mathbf{S}(\tilde{\theta}_R)'\mathbf{I}(\tilde{\theta}_R)^{-1}\mathbf{S}(\tilde{\theta}_R) \xrightarrow{d} \chi^2_J$$

Example 2.4: LM

Consider the linear regression model under normality:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

We want to test the restriction $\mathcal{H}_0 : \beta_1 = 1$. The restricted model is:

$$y_i - x_{i1} = \beta_0 + \beta_2 x_{i2} + e_i$$

The MLE estimates for the restricted model are $\tilde{\beta}_0, \tilde{\beta}_2$ and $\tilde{\beta}_1 = 1$. Since the unrestricted likelihood whose gradient needs to be evaluated at the constraint estimates is given by:

$$\mathcal{L} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

Under the classical linear regression with normality, we want to verify that the Score (gradient) is close to zero at the restricted estimates $\tilde{\beta}$. The derivatives of the log-likelihood are:

$$\frac{\partial \mathcal{L}}{\partial \beta_0}\Big|_{\tilde{\beta},\tilde{\sigma}^2} = \frac{1}{\tilde{\sigma}^2}\sum \tilde{e}_i \approx 0$$

$$\frac{\partial \mathcal{L}}{\partial \beta_1}\Big|_{\tilde{\beta},\tilde{\sigma}^2} = \frac{1}{\tilde{\sigma}^2}\sum \tilde{e}_i x_{i1} \approx 0$$

$$\frac{\partial \mathcal{L}}{\partial \beta_2}\Big|_{\tilde{\beta},\tilde{\sigma}^2} = \frac{1}{\tilde{\sigma}^2}\sum \tilde{e}_i x_{i2} \approx 0$$

The FOCs for $\beta_0, \beta_2, \sigma^2$ are satisfied (they form the constraint optimisation). The LM test focuses on the condition for the restricted parameter β_1.

LM Test (Auxiliary Regression): The LM test is implemented by running an auxiliary regression: first, regress the restricted residuals $\tilde{e}_i$ on all regressors x_{i1} and x_{i2}. The LM statistic is then related to the R^2 of this auxiliary regression.

Intuition: If $\beta_1 \neq 1$ (i.e., $\mathcal{H}_0$ is false), the residuals from the restricted model $\tilde{e}_i$ will still display correlation with the excluded regressor x_{i1}. Hence, the covariance between $\tilde{e}_i$ and x_{i1} (which is the score component for β_1) will be high, resulting in a high R^2 and leading to rejection of $\mathcal{H}_0$.

2.3.4 The Likelihood Ratio Test

We introduce the following notation. Let $Q_n(\theta) := \mathcal{L}(\theta)$, and $Q(\theta) = \mathbb{E}[\ln f(Y_i, X_i \mid \theta)]$. The LR test compares the maximum log-likelihood under the **unrestricted model** ($\hat{\theta}$) to the maximum log-likelihood under the **restricted model** ($\tilde{\theta}$).

$$LR = 2\left[Q_n(\hat{\theta}) - Q_n(\tilde{\theta})\right] \xrightarrow{d} \chi^2_J$$

Example 2.5: LR example

Suppose we want to test $\mathcal{H}_0 : \beta_1 = 1$. We need to run our MLE imposing $\beta_1 = 1$, and then run it without any restriction, obtaining $Q_n(\tilde{\beta}$ and $Q_n(\hat{\beta})$ respectively. We therefore obtain

$$LR = 2\left(Q_n(\hat{\beta}) - Q_n(\tilde{\beta}\right) \xrightarrow{d} \chi^2_1$$

Figure 2.3.4 shows a graphical represnetation of the 3 tests.

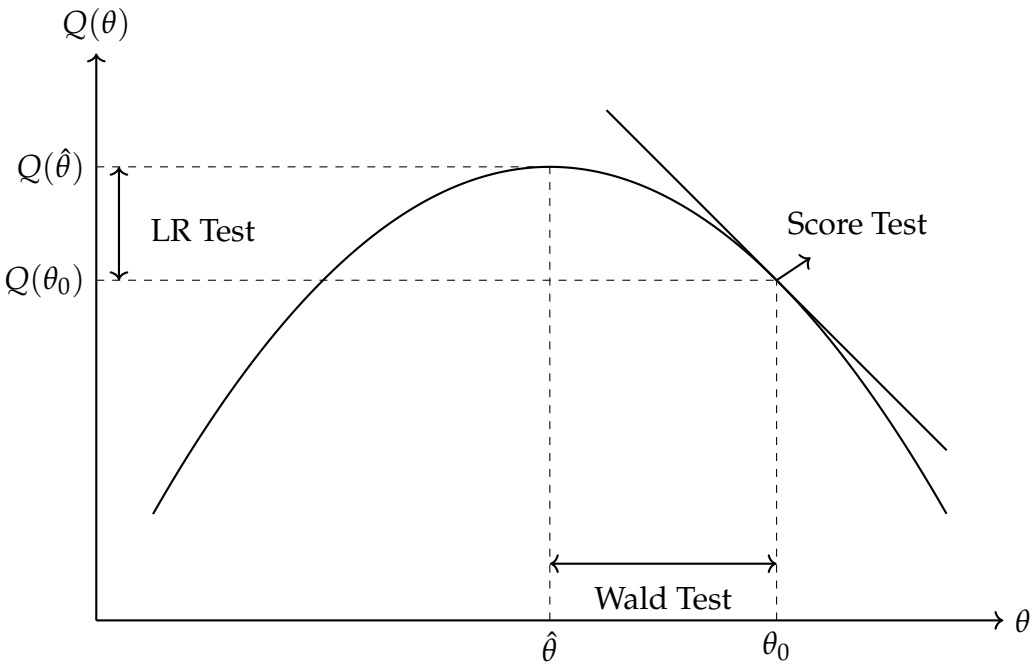

Figure 2: The trinity of hypothesis testing

3 Generalized Linear Regression

The whole point of this section is to relax assumption iv). That is, we do not assume spherical errors anymore. Indeed, we now consider a case where $\mathbb{V}(\varepsilon \mid X) = \Omega \neq \sigma^2 I_n$[7], which we denote assumption iv)'. This can different shape. The most common one would be to think about **heteroskedasticity**, that is a case where

$$\mathbb{V}(\varepsilon_i) = \sigma_i^2 \implies \Omega = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n^2 \end{pmatrix}$$

Another interesting example is one with **serial correlation**, for instance where ε_t follows an $AR(1)$ process:

[7]For consistency in the Generalized Least Squares (GLS) derivation, we define the full $n \times n$ covariance matrix as $\Omega \equiv \mathbb{V}(\varepsilon \mid X)$, which therefore absorbs the common scalar σ^2 from the initial assumption (iv).

$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t, \text{ with } |\rho| < 1, v_t \sim iid(0, \sigma^2)$$

$$\implies \Omega = \frac{\sigma^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{pmatrix}$$

3.1 Finite-Sample Properties

Consider the Gauss-Markov assumptions i) to iii), with assumption iv)'. It is easy to show that the OLS estimtor is still unbiased, since we rely on exogeneity assumption iii) to show unbiasedness. The conditional variance can be computed as follows

$$\mathbb{V}(\hat{\beta}|X) = \mathbb{E}\left[(\hat{\beta} - \mathbb{E}[\hat{\beta}|X](\hat{\beta} - \mathbb{E}[\hat{\beta}|X])' \mid X\right]$$

$$= \mathbb{E}\left[(\beta + (X'X)^{-1}X'\varepsilon - \beta)((\beta + (X'X)^{-1}X'\varepsilon - \beta)])' \mid X\right]$$

$$= \mathbb{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right] = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

OLS is therefore not BLUE anymore, since it is not efficient. We discuss later how to fix this.

3.2 Asymptotic Properties

First, for consistency, recall that we rely on a combination of a WLLN, Slutsky's Theorem, and assumption iii), we obtain that $\hat{\beta} \xrightarrow{p} \beta$. Since $\{x_i\varepsilon_i\}$ is no longer iid, we need to use a stronger version of the WLLN. Notably, we are lacking the **identically distributed** part. Based on Theorem 1.5, we have a suitable LLN that we can use for such cases, since we do not rely on any sort of identically distributed assumption. Notice that conditions of Theorem 1.5 are actually much stronger than what we need (see Appendix A.1).

Now, for asymptotic normality in presence of non-spherical errors, it is not possible to use Lindeberg-Levy's CLT (Theorem 1.6). Instead, we must use Lindeberg-Feller's CLT (Theorem 1.7). Using this CLT, provided that the Lindeberg condition holds, and following derivations done in the general case above, we get

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} Q^{-1}\mathcal{N}(0, \Omega^*) = \mathcal{N}(0, Q^{-1}\Omega^*Q^{-1})$$

Below is an elegant example of when Lindeberg-Feller's CLT can be applied.

Consider

$$y_i = \beta x_i + u_i, \quad i = 1, \dots, n,$$

where $E(u_i^2) = \sigma^2$. Suppose the data are grouped into J unequal categories:

$$Y_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i, \quad X_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \quad n = \sum_{j=1}^{J} n_j.$$

Estimate β by the no-intercept OLS regression of Y_j on X_j:

$$\hat{\beta} = \left(\sum_{j=1}^{J} X_j^2\right)^{-1} \sum_{j=1}^{J} X_j Y_j.$$

We want to derive the limiting distribution of $\hat{\beta}$ and construct a test of the null hypothesis that $\beta = 0$ based on the grouped data.

First, let us write the model for a specific group $\mathcal{G}_j$:

$$Y_j = \beta X_j + U_j \text{ where } U_j = \frac{1}{n_j}\sum_{i\in\mathcal{G}_j} u_i$$

We need to put assumption on our model. A standard assumption is that $u_i \sim \mathcal{N}(0, \sigma^2)$ and is independent and identically distributed (iid). Based on this assumption, notice that for arbitrary group $\mathcal{G}_j$, we can compute the expectation

$$\mathbb{E}\left(U_j\right) = \mathbb{E}\left(\frac{1}{n_j}\sum_{i\in\mathcal{G}_j} u_i\right) = \frac{1}{n_j}\sum_{i\in\mathcal{G}_j}\mathbb{E}(u_i) = 0$$

and the variance

$$\mathbb{V}(U_j) = \mathbb{V}\left(\frac{1}{n_j}\sum_{i\in\mathcal{G}_j} u_i\right) = \frac{1}{n_j^2}\sum_{i\in\mathcal{G}_j}\sum_{k\in\mathcal{G}_j}\mathbb{C}\text{ov}(u_i u_k)$$

$$= \frac{1}{n_j^2}\left(\sum_{i\in\mathcal{G}_j}\mathbb{V}(u_i) + \sum_{i\in\mathcal{G}_j}\sum_{k\neq i, k\in\mathcal{G}_j}\mathbb{C}\text{ov}(u_i, u_k)\right) = \frac{\sigma^2}{n_j} \equiv \sigma_j^2$$

The last equality comes from the fact that $\forall i \neq k, \mathbb{C}\text{ov}(u_i, u_k) = 0$.

We can re-write $\hat{\beta}$ as follows:

$$\hat{\beta} = \beta + \left(\sum_{j=1}^{J} X_j^2\right)^{-1} \sum_{j=1}^{J} X_j U_j$$

We are interested in the behavior of

$$\sqrt{J}(\hat{\beta} - \beta) = \sqrt{J}\left(\sum_{j=1}^{J} X_j^2\right)^{-1} \sum_{j=1}^{J} X_j U_j = \frac{\frac{1}{\sqrt{J}}\sum_{j=1}^{J} X_j U_j}{\frac{1}{J}\sum_{j=1}^{J} X_j^2}$$

The key fact to notice here is that the variance varies for each group $\mathcal{G}_j$ (i.e. heteroskedasticity), which implies we can't apply Lindeberg-Levy's Central Limit Theorem, since this theorem only works for iid data [8] Therefore, we need to use the Lindeberg-Feller CLT. I take the definition from a previous course, since we do not have the formal definition in this context.

We need to assume some **Lindeberg Condition**. This condition requires that the sum of the extreme-value variances becomes negligible compared to the total variance as $n \to \infty$.

In this context, we are interested in writing the Lindeberg condition to the sequence $Z_j := X_j U_j$, which is an independent but non-iid sequence of random variables. Notice that by LIE,

$$\mathbb{E}(Z_j) = \mathbb{E}[\mathbb{E}(X_j U_j | X_j)] = \mathbb{E}[X_j \mathbb{E}(U_j | X_j)] = 0$$

The variance can be computed as

$$\mathbb{V}(Z_j) = \mathbb{E}(X_j^2 U_j^2) = \mathbb{E}[X_j^2 \mathbb{E}(U_j^2 | X_j)] = \sigma_j^2 X_j^2 \equiv \tilde{\sigma}_j^2$$

We can now define

$$C_J := \sqrt{\sum_j \tilde{\sigma}_j^2}$$

We impose the following condition

$$\lim_{J \to \infty} \frac{\max_{j \in \{1,\dots,J\}} (X_j^2 / n_j)}{\sum_j (X_j^2 / n_j)} = 0$$

which is a sufficient condition for

$$\lim_{J \to \infty} \max_{j \in \{1,\dots,J\}} \frac{\tilde{\sigma}_j^2}{C_J^2} = 0$$

Now, assume the Lindeberg Condition holds, that is, assume, for any $\varepsilon > 0$:

$$\lim_{J \to \infty} \frac{1}{C_J^2} \sum_{j=1}^{J} \mathbb{E}\left[Z_j^2 \mathbb{1}\{|Z_j| \geq \varepsilon C_J\} \right] = 0$$

We can finally apply Lindeberg-Feller's Theorem:

$$\frac{\sum_{j=1}^J Z_j}{\sqrt{\sum_j \tilde{\sigma}_j^2}} = \frac{\sum_{j=1}^J X_j U_j}{C_J} \xrightarrow{d} \mathcal{N}(0,1)$$

[8] We cannot use the classical i.i.d. CLT because $\mathbb{V}(X_j U_j)$ varies with j through n_j. As J grows, the collection $\{X_j U_j\}_{j=1}^J$ forms a *triangular array* of independent but non-identically distributed random variables, so a Lindeberg–Feller CLT is required.

We can rewrite our object of interest as

$$\frac{1}{\sqrt{J}}\sum_{j=1}^{J} X_j U_j \frac{C_J}{\sqrt{J}} \cdot \frac{\sum_{j=1}^{J} X_j U_j}{C_J}$$

We analyze the limit of $\left(\frac{C_J}{\sqrt{J}}\right)^2$ since it represents the asymptotic variance of the numerator in the estimator:

$$\lim_{J\to\infty}\left(\frac{C_J}{\sqrt{J}}\right)^2 = \lim_{J\to\infty}\frac{C_J^2}{J} = \lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}\frac{\sigma^2 X_j^2}{n_j} := \sigma^2\Omega$$

By Slutsky's Theorem, we have that

$$\frac{1}{\sqrt{J}}\sum_{j=1}^{J} X_j U_j \xrightarrow{d} \mathcal{N}(0, \sigma^2\Omega)$$

Now, assume that

$$\frac{1}{J}\sum_{j=1}^{J} X_j^2 \xrightarrow{p} Q \implies \left(\frac{1}{J}\sum_{j=1}^{J} X_j^2\right)^{-1} \xrightarrow{p} Q^{-1} \text{ by Continuous Mapping Theorem}$$

We can therefore conclude that by Linderberg-Feller CLT and Slutsky's Theorem,

$$\sqrt{J}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \frac{\sigma^2\Omega}{Q^2})$$

Finally, we can perform the following T-test:

$$T = \frac{\hat{\beta}}{\sqrt{\sigma^2 \frac{\left(\sum_{j=1}^{J} \frac{X_j^2}{n_j}\right)}{\left(\sum_{j=1}^{J} X_j^2\right)^2}}}$$

Notice that under $H_0 : \beta = 0, T \xrightarrow{d} \mathcal{N}(0, 1)$

If σ^2 is unknown, we can use a standard method to estimate it.

3.3 Generalized Least Squares (GLS)

3.3.1 Theoretical GLS

The Generalized Least Squares (GLS) allows to adjust the linear regression model for it to satisfy the Gauss-Markov Conditions. The resulting OLS estimator equals the estimator that would minimize the generalized sum of squares

$$S(\beta) = (y - X\beta)'\Omega^{-1}(y - X\beta) \implies \hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

The original model, $y = X\beta + \varepsilon$, has non-spherical errors, meaning $\mathbb{V}(\varepsilon \mid X) = \Omega$ (where Ω is an $n \times n$ matrix, and we are using your new, consistent notation). This violates the standard Gauss-Markov assumption (spherical errors), making OLS inefficient.

The solution is to premultiply the entire model by an invertible $n \times n$ matrix R such that:

$$R'R = \Omega^{-1}$$

The transformed model is:

$$Ry = RX\beta + R\varepsilon$$
$$y^* = X^*\beta + \varepsilon^*$$

The errors in the new model, $\varepsilon^* = R\varepsilon$, now satisfy the spherical errors assumption, which is the whole point of the transformation:

$$\mathbb{V}(\varepsilon^* \mid X) = \mathbb{E}(\varepsilon^*\varepsilon^{*\prime} \mid X) = \mathbb{E}(R\varepsilon\varepsilon' R' \mid X)$$
$$= R\mathbb{E}(\varepsilon\varepsilon' \mid X)R' = R\Omega R'$$

Since $R'R = \Omega^{-1}$, it follows that $\Omega = (R'R)^{-1} = R^{-1}(R')^{-1}$. Substituting this back:

$$R\Omega R' = R[R^{-1}(R')^{-1}]R' = (RR^{-1})[(R')^{-1}R'] = I_n I_n = I_n$$

The transformed error term, ε^*, now has a variance proportional to the identity matrix (I_n), meaning it is spherical. OLS applied to the transformed model (y^* regressed on X^*) is the Best Linear Unbiased Estimator (BLUE).In fact, if you apply OLS to the transformed model $(X^{*\prime}X^*)^{-1}X^{*\prime}y^*$, you will find it simplifies directly to the GLS formula:

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

Theorem 3.1: Aitken Theorem

Under assumptions i) to iv), the GLS estimator $\hat{\beta}_{GLS}$ is efficient relative to all other unbiased estimators of β. This is a consequence of the transformation.

Assuming suitable regularity conditions[9], the GLS estimator is consistent:

$$\hat{\beta}_{\text{GLS}} - \beta = \left(\frac{X'\Omega^{-1}X}{n}\right)^{-1} \frac{X'\Omega^{-1}\varepsilon}{n} \xrightarrow{p} 0$$

The asymptotic normality is given by:

$$\sqrt{n}(\hat{\beta}_{\text{GLS}} - \beta) \xrightarrow{d} \mathcal{N}\left(0, D_{\Omega^{-1}}^{-1}\right)$$

[9]On top of the usual exogeneity and moment conditions, we need to assume the $n \times n$ covariance matrix Ω must be positive definite, ensuring Ω^{-1} exists and is positive definite. This is necessary for the transformation $R'R = \Omega^{-1}$. Additionally, the limit matrix $D_{\Omega^{-1}} = \text{plim } \frac{X'\Omega^{-1}X}{n}$ must be finite and invertible (positive definite). This ensures the asymptotic variance is well-defined and finite.

where the $K \times K$ matrix $D_{\Omega^{-1}}$ is the probability limit of the normalized $X'\Omega^{-1}X$ term:

$$D_{\Omega^{-1}} = \text{plim } \frac{X'\Omega^{-1}X}{n} > 0$$

Note: The matrix $D_{\Omega^{-1}}$ here is analogous to the Q matrix in the spherical OLS case, representing the asymptotic precision of the transformed data. Unlike Ω^* in the OLS sandwich, $D_{\Omega^{-1}}$ requires no "sandwiching" because the GLS transformation has already ensured efficiency.

3.3.2 Feasible GLS

For all the results above, we need to assume Ω is known. The feasible version of GLS (FGLS) is a version where Ω is replaced by its estimator, $\hat{\Omega}$

$$\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y$$

Typically, $\hat{\beta}_{FGLS}$ is biased, but consistent and asymptotically normal and efficient. We can write the following two sufficient conditions for GLS and FGLS to be *asymptotically equivalent*

$$\text{plim } \frac{X'(\hat{\Omega}^{-1} - \Omega^{-1})X}{n} = 0 \text{ and plim } \frac{X'(\hat{\Omega}^{-1} - \Omega^{-1})\varepsilon}{\sqrt{n}} = 0$$

Condition 1 ensures the denominator (or the non-stochastic part, $\frac{X'\Omega^{-1}X}{n}$) of the FGLS estimator converges to the same limit as the GLS estimator. Condition 2 ensures the numerator (or the stochastic part, $\frac{X'\Omega^{-1}\varepsilon}{\sqrt{n}}$) of the FGLS estimator has the same asymptotic distribution as the GLS estimator. I prove the proposition below.

Proof. We want to show that

$$\text{plim}\frac{1}{N}X'(\hat{\Omega}^{-1} - \Omega^{-1})X = 0 \quad \text{and} \quad \text{plim}\frac{1}{\sqrt{N}}X'(\hat{\Omega}^{-1} - \Omega^{-1})u = 0.$$

$$\implies \sqrt{N}(\hat{\beta}_{GLS} - \beta) - \sqrt{N}(\hat{\beta}_{FGLS} - \beta) \xrightarrow{p} 0$$

Consider the following notation, which simplifies the algebra a lot:

$$Q_N := \frac{X'\Omega^{-1}X}{N} \quad \hat{Q}_N := \frac{X'\hat{\Omega}^{-1}X}{N} \quad W_N := \frac{X'\Omega^{-1}u}{\sqrt{N}} \quad \hat{W}_N := \frac{X'\hat{\Omega}^{-1}u}{\sqrt{N}}$$

Therefore, we want to show that

$$Q_N^{-1}W_N - \hat{Q}_N^{-1}\hat{W}_N \xrightarrow{p} 0$$

We can now add and subtract $\hat{Q}_N^{-1}W_N$, which yields:

$$(Q_N^{-1}W_N - \hat{Q}_N^{-1}W_N) + (\hat{Q}_N^{-1}W_N - \hat{Q}_N^{-1}\hat{W}_N)$$

We focus first on the first difference. By factoring and using the property of difference of inverse matrices [10], we can rewrite it as

$$(Q_N^{-1} - \hat{Q}_N^{-1})W_N = \left[Q_N^{-1}(\hat{Q}_N - Q_N)\hat{Q}_N^{-1}\right] W_n$$

Now, by our hypothesis, we know that $\hat{Q}_N - Q_N = o_p(1)$. Additionally, Assuming $\text{plim}\left(\frac{1}{N}X'\Omega^{-1}X\right) = Q$, a finite non-singular matrix, it follows by the Continuous Mapping Theorem that $\text{plim}(Q_N^{-1}) = Q^{-1}$, which implies $Q_N^{-1} = O_p(1)$. We also need to show that $\hat{Q}_N^{-1} = O_p(1)$. Notice that we know that $\text{plim}(\hat{Q}_N - Q_N) = 0$ and that $\text{plim}Q_N = Q$. Therefore, we can rewrite

$$\text{plim}\hat{Q}_N = \text{plim}(\hat{Q}_N - Q_N + Q_N) = \text{plim}(\hat{Q}_N - Q_N) + \text{plim}(Q_N) = Q$$

By the same argument as before (CMT), this implies $\hat{Q}_N^{-1} = O_p(1)$. Finally, note that $W_n = O_p(1)$ by central limit theorem. Notice that for any $Z_n = O_p(1)$, $A_n = o_p(1)$, $Z_nA_n = o_p(1)$, we have that the first difference is therefore

$$\left[Q_N^{-1}(\hat{Q}_N - Q_N)\hat{Q}_N^{-1}\right] W_n = O_p(1)o_p(1)O_p(1)O_p(1) = o_p(1)$$

We can move on to the second difference, which we rewrite as follows:

$$(\hat{Q}_N^{-1}W_N - \hat{Q}_N^{-1}\hat{W}_N) = \hat{Q}_N^{-1}(W_N - \hat{W}_N)$$

We have shown $\hat{Q}_N^{-1} = O_p(1)$, so we focus on the term $(W_N - \hat{W}_N)$, which we rewrite as

$$(W_N - \hat{W}_N) = \frac{1}{\sqrt{N}}X'(\Omega^{-1} - \hat{\Omega}^{-1})u$$

We are given that $\text{plim}\frac{1}{\sqrt{N}}X'(\hat{\Omega}^{-1} - \Omega^{-1})u = 0$. Notice that by continuity[11] of the plim operator, it follows that

$$\text{plim}(W_N - \hat{W}_N) = \text{plim}\left[-\frac{1}{\sqrt{N}}X'(\hat{\Omega}^{-1} - \Omega^{-1})u\right] = 0$$

Therefore, $(W_N - \hat{W}_N) = o_p(1)$, which implies $\hat{Q}_N^{-1}(W_N - \hat{W}_N) = O_p(1)o_p(1) = o_p(1)$

We have shown that

$$Q_N^{-1}W_N - \hat{Q}_N^{-1}\hat{W}_n = (Q_N^{-1}W_N - \hat{Q}_N^{-1}W_N) + (\hat{Q}_N^{-1}W_N - \hat{Q}_N^{-1}\hat{W}_N)\hat{W}_N = o_p(1) + o_p(1) = o_p(1)$$

$$\implies Q_N^{-1}W_N - \hat{Q}_N^{-1}\hat{W}_n \xrightarrow{p} 0$$

□

[10] For any two invertible matrices A and B of the same dimension,

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}.$$

This identity follows from adding and subtracting $A^{-1}BB^{-1}$ and is valid whenever both inverses exist.

[11] If $X_n \xrightarrow{p} X$, then for any constant a, $aX_n \xrightarrow{p} aX$. This follows from the continuity of convergence in probability (Continuous Mapping Theorem), and is often informally referred to as linearity of the probability limit.

Note that asymptotic efficiency of FGLS does not require the efficiency of the estimator of θ. Only consistency of $\hat{\theta}$ is required to achieve full efficiency for the FGLS estimator.

3.3.3 Application: Heteroskedasticity

This is a problem often encountered in cross-sectional and panel models. In these studies, we are interested in studying members of a *population*, say at a given point in time. We can think about firms, industries, geographical units (e.g. counties, states) or individual consumers. Members of such population may vary in many different characteristis (size, productivity, demographic composition,...), which can imply that individuals/firms are drawn from distributions with different variances: heteroskedasticity.

Let us define $\mathbb{V}(\varepsilon_i|X_i) = \sigma_i^2 := \sigma^2\omega_i^2$. The transformed model satisfying the GM conditions[12] is

$$\frac{y_i}{\omega_i} = \left(\frac{x_i}{\omega_i}\right)'\beta + \frac{\varepsilon_i}{\omega_i}$$

In this case, R is a diagonal matrix with $\frac{1}{\omega_i}$ on the diagonal, and Ω is a diagonal matrix of $\sigma^2\omega_i^2$. The weighted least squares estimator is therefore given by

$$\hat{\beta}_{WLS} = \left(\sum_i \frac{x_i x_i'}{\omega_i^2}\right)^{-1} \sum_i \frac{y_i x_i'}{\omega_i^2}$$

The idea is basically to reweight based on how noisy this observation is. If the observation has a high variance (and therefore a high ω_i^2), the denominator increases so we put less weight on this observation. We now introduce an example.

Example 3.1: Grouped-data Regression model

Rather than observing all individuals, we may observe only group averages (case in many very large datasets, or anonymized datasets etc).
The true data generating process satisfies all Gauss-Markov assumptions (including full independence):

$$y_{ij} = x_{ij}'\beta + \varepsilon_{ij} \quad j = 1, \ldots, M_i, i = 1, \ldots, n$$

where $\varepsilon_{ij} \sim iid(0, \sigma^2)$.
However, the estimable model is defined by the group means:

$$\bar{y}_i = \bar{x}_i'\beta + \bar{\varepsilon}_i$$

with $\bar{y}_i = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i}$, $\bar{x}_i = \sum_{j=1}^{M_i} \frac{x_{ij}}{M_i}$, and $\bar{\varepsilon}_i = \sum_{j=1}^{M_i} \frac{\varepsilon_{ij}}{M_i}$.

[12] It is easy to show that spherical errors hold in the transformed model: $Var(\varepsilon^* \mid X) = \mathbb{E}(\varepsilon^*\varepsilon^{*\prime} \mid X) = \frac{1}{\omega_i^2}\mathbb{E}(\varepsilon_i\varepsilon_i'|x_i) = \sigma^2$

The estimable model exhibits heteroskedasticity, as the variance of the average error is:

$$\mathbb{V}(\bar{\varepsilon}_i) = \mathbb{V}\left(\sum_{j=1}^{M_i} \frac{\varepsilon_{ij}}{M_i}\right) = \frac{\sigma^2}{M_i}$$

Therefore, the variance of the error term $\bar{\varepsilon}_i$ is inversely proportional to the group size M_i. Observations from larger groups (M_i large) are more precise (smaller variance) than observations from smaller groups.
GLS is an OLS on the transformed model, where the transformation uses the inverse square root of the variance, $\frac{1}{\sqrt{\mathbb{V}(\bar{\varepsilon}_i)}} = \frac{\sqrt{M_i}}{\sigma}$. Since σ is constant across all observations, the necessary weight ω_i is $\sqrt{M_i}$.
The transformed model is:

$$\sqrt{M_i}\bar{y}_i = \sqrt{M_i}\bar{x}_i'\beta + \sqrt{M_i}\bar{\varepsilon}_i$$

$$\text{or } y_i^* = x_i^{*\prime}\beta + \varepsilon_i^*$$

The transformed error $\varepsilon_i^* = \sqrt{M_i}\bar{\varepsilon}_i$ is now homoskedastic:

$$\mathbb{V}(\varepsilon_i^*) = \mathbb{V}(\sqrt{M_i}\bar{\varepsilon}_i) = M_i\mathbb{V}(\bar{\varepsilon}_i) = M_i\frac{\sigma^2}{M_i} = \sigma^2$$

Instead of using Weighted Least Squares (WLS), we can choose to use OLS, while only adjusting the standard errors. This approach does not require us to specify the functional form of the heteroskedasticity. The appropriate covariance matrix for the OLS estimator when heteroskedasticity is present is given by the sandwich formula, derived from the asymptotic variance:

$$\mathbb{V}(\hat{\beta} \mid X) = \left(\sum x_i x_i'\right)^{-1}\left(\sum \sigma_i^2 x_i x_i'\right)\left(\sum x_i x_i'\right)^{-1}$$

The estimator of this covariance matrix, known as the White (1980) estimator (or the Heteroskedasticity-Consistent Covariance Matrix Estimator, HCCME), replaces the unknown σ_i^2 with the squared OLS residuals $\hat{\varepsilon}_i^2$:

$$\widehat{\mathbb{V}(\hat{\beta} \mid X)} = \left(\sum x_i x_i'\right)^{-1}\left(\sum \hat{\varepsilon}_i^2 x_i x_i'\right)\left(\sum x_i x_i'\right)^{-1}$$

This estimator is consistent because it relies on the condition that the difference between the true normalized σ_i^2 and the estimated normalized $\hat{\varepsilon}_i^2$ vanishes asymptotically:

$$\operatorname{plim} \frac{1}{n}\sum(\sigma_i^2 - \hat{\varepsilon}_i^2)x_i x_i' = 0$$

Standard errors calculated as the square root of the diagonal elements of this matrix are usually referred to as heteroskedasticity-consistent standard errors or simply White standard errors. In Stata, this is typically implemented using the 'vce(robust)' option.

Now, we introduce a theorem that provides necessary and sufficient conditions for equivalence between OLS and GLS.

Theorem 3.2: Kruskal's Theorem

Under assumptions i) to iii) and iv)', OLS is efficient if and only if

- The column space of X, denoted $\mathcal{R}(X)$ (i.e. the space spanned by the K columns of X), is spanned by K egeinvectors of the covariance matrix Ω
- The column space of the matrix product $X\Omega$ is that same as the column space of X, that is $\mathcal{R}(X\Omega) = \mathcal{R}(X)$

I find the second condition a bit more intuitive. $\mathcal{R}(X\Omega) = \mathcal{R}(X)$ means that applying the covariance structure Ω to the explanatory variables X does not "move" them out of their column space. The condition $\mathcal{R}(\Omega X) = \mathcal{R}X$ is the formal linear algebra requirement that ensures this proportionality holds, meaning the weighting matrix Ω^{-1} does not change the resulting projection (the coefficients $\hat{\beta}$).

3.3.4 Heteroskedasticity Tests

We introduce two tests for heteroskedasticity. First, the Goldfeld-Quandt Test (1965). We assume the heteroskedastic variance σ^2 is monotonically related to one observable variable, z_i. It is then a simple F test where we separate observations into two groups after reordering by z_i, omitting r central observations to improve power. Under $\mathcal{H}_0$ (homoskedasticity):

$$F = \frac{s_1^2}{s_2^2} \sim F_{m-k,m-k}$$

where $m = \frac{n-r}{2}$ and $s_j^2 = \hat{\varepsilon}'\hat{\varepsilon}/(m-k)$.

The second test is the White Test. It Does not specify anything about the form of heteroskedasticity. We regress the squared residuals $\hat{\epsilon}^2$ on a constant and all p unique first moments, second moments, and cross-products of the original regressors. The test statistic is based on the R^2 from the auxiliary regression:

$$\text{White Test} = nR^2 \xrightarrow{p} \chi_p^2$$

It is also possible to apply the trinity of hypothesis testing to this context (see Limodio's slides).

4 Identification and Instrumental Variables

In this section, we relax the assumption of conditional mean independence, that is assumption iii). The asymptotic results derived in earlier section are all based on $\mathbb{E}(\varepsilon|X) = 0$.[13]

[13] A weaker condition, contemporaneous uncorrelatedness, which is implied by conditional mean independence, and written as $\mathbb{E}(x_i\varepsilon_i) = 0$ is sufficient to ensure the consistency of the OLS estimator.

However, in most cases, our regression has an endogeneity problem, that is $\mathbb{E}(x_i\varepsilon_i) \neq 0$. Using our asymptotic results from above, by Khinchine's WLLN, we obtain that

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} \mathbb{E}(x_i x_i') \equiv Q \text{ and } \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{p} \mathbb{E}(x_i \varepsilon_i) \neq 0$$

Therefore, by Slutsky's Theorem and CMT, we understand the OLS estimator is inconsistent

$$\beta \xrightarrow{p} \beta + Q^{-1}\mathbb{E}(x_i\varepsilon_i) \neq \beta$$

There are different causes of endogeneity, which I detail below.

4.1 Cause of Endogeneity

4.1.1 Measurement error

The true (unobservable) model starts with conditional mean independence $\mathbb{E}(\varepsilon_i|x_i^*) = 0$:

$$y_i = x_i^{*\prime}\beta + \varepsilon_i$$

We observe a measurement error u_i such that $x_i = x_i^* + u_i$, where $\mathbb{E}(u_i) = 0$. The key assumptions are that the measurement error u_i is independent of both the true regressor x_i^* and the equation error ε_i. These are strong assumptions, implying that the true value x_i^* reveals no information about the sign, size, or value of the measurement error u_i.

The regression model we estimate uses the observed regressor x_i and has a composite error term v_i:

$$y_i = x_i'\beta + v_i$$

where $v_i = \varepsilon_i - u_i'\beta$.

By construction, the observed regressor x_i and the new error v_i are correlated:

$$\mathbb{E}(x_i v_i) = \mathbb{E}[(x_i^* + u_i)(\varepsilon_i - u_i'\beta)] \neq 0$$

Specifically, $\mathbb{E}(x_i v_i) = \mathbb{E}(u_i v_i) = -\mathbb{E}(u_i u_i')\beta \neq 0$ because u_i and v_i are necessarily correlated as both are functions of the measurement error u_i.

Therefore, the OLS estimator $\hat{\beta}$ of equation is in general inconsistent. The regressor x_i becomes endogenous when measured with error, and the resulting parameter estimates will be biased, with the extent of this bias depending on the magnitude and pattern of the error.

Example 4.1: Special case: bivariate regression

The linear model is $y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i$ with conditional mean independence $\mathbb{E}(\epsilon_i|\tilde{x}_i^*) = 0$. The true regressor $\tilde{x}_i^*$ is unobserved, and we instead observe the

error-ridden version $\tilde{x}_i = \tilde{x}_i^* + \tilde{u}_i$.

Suppose the measurement error $\tilde{u}_i$ is independent of the true regressor $\tilde{x}_i^*$ and the equation error ϵ_i, with $\mathbb{E}(\tilde{u}_i) = 0$ and variance $\mathbb{V}(\tilde{u}_i) = \sigma_{\tilde{u}}^2$.

Asymptotic Properties of $\hat{\beta}_1$

The probability limit of the slope estimator $\hat{\beta}_1$ is:

$$\operatorname{plim} \hat{\beta}_1 = \frac{\operatorname{plim} \frac{1}{n} \sum(\tilde{x}_i - \bar{\tilde{x}})(y_i - \bar{y})}{\operatorname{plim} \frac{1}{n} \sum(\tilde{x}_i - \bar{\tilde{x}})^2} = \frac{\mathbb{C}\text{ov}(\tilde{x}_i, y_i)}{\mathbb{V}(\tilde{x}_i)} = \beta_1 \frac{\mathbb{V}(\tilde{x}_i^*)}{\mathbb{V}(\tilde{x}_i^*) + \sigma_{\tilde{u}}^2}$$

Since the term $\frac{\mathbb{V}(\tilde{x}_i^*)}{\mathbb{V}(\tilde{x}_i^*)+\sigma_{\tilde{u}}^2}$ is less than 1, this is the typical case of "attenuation bias". Therefore, $\hat{\beta}_1$ is inconsistent, and the amount of inconsistency is small when the "noise to signal ratio" ($\sigma_{\tilde{u}}^2 / \mathbb{V}(\tilde{x}_i^*)$) is small.

Asymptotic Properties of $\hat{\beta}_0$

The intercept estimator $\hat{\beta}_0$ is also inconsistent:

$$\operatorname{plim} \hat{\beta}_0 = \operatorname{plim} \bar{y} - \operatorname{plim} \hat{\beta}_1 \operatorname{plim} \bar{\tilde{x}} = \beta_0 + \beta_1 \operatorname{plim} \bar{\tilde{x}}^* + \operatorname{plim} \bar{\epsilon} - \operatorname{plim}(\operatorname{plim} \hat{\beta}_1 \operatorname{plim} \bar{\tilde{x}})$$

$$\operatorname{plim} \hat{\beta}_0 = \beta_0 + \beta_1 \operatorname{E}(\tilde{x}^*) \frac{\mathbb{V}(\tilde{x}^*)}{\mathbb{V}(\tilde{x}^*) + \sigma_{\tilde{u}}^2}$$

(The formula in the source is missing a β_1 term, assuming $\operatorname{plim} \bar{\epsilon} = 0$ and $\operatorname{plim} \bar{u} = 0$). Therefore, $\hat{\beta}_0$ is also inconsistent, unless $\operatorname{E}(\tilde{x}^*) = 0$ or $\beta_1 = 0$. The direction of inconsistency depends on the sign of β_1 and $\operatorname{E}(\tilde{x}^*)$.

In general, if only one regressor is measured with error, the estimator of its coefficient is asymptotically shrunk to zero. If more variables are measured with error, then very little can be said about the direction. Therefore, measurement error in more than one explanatory variable gives inconsistent OLS, not only for the parameters associated with the variables measured with error but for all parameters in general.

4.1.2 Simultaneity

Simultaneity arises when some of the independent variables are jointly determined with the dependent variable.
Consider the following linear market model which uses a structural form representation consisting of two equations:

- Demand equation: $q_t^d = \alpha p_t + X_t^d \beta + u_{1t}$
- Supply equation: $q_t^s = \gamma p_t + X_t^s \delta + u_{2t}$

Here, X_t^d and X_t^s are predetermined variables that capture exogenous shifts in supply and demand, satisfying conditional mean independence. These are structural equations derived from theory, each describing a particular aspect of the economy.
The market equilibrium condition, $q_t^d = q_t^s = q_t$, ensures that price (p_t) and quantity (q_t) are jointly determined, making them both endogenous variables.

Our concern is the estimation of these structural form equations. To see the endogeneity problem, we rewrite the model in its reduced form by solving for p_t and q_t in terms of X_t^d, X_t^s, u_{1t}, and u_{2t}. The reduced form for price is:

$$p_t = \frac{1}{\alpha - \gamma}\left(X_t^s\delta - X_t^d\beta + u_{2t} - u_{1t}\right)$$

Clearly, p_t is correlated with both error terms u_{1t} and u_{2t}. Applying OLS to each structural form equation separately will therefore give inconsistent parameter estimates. OLS applied to the reduced form equations will provide consistent estimates of the reduced form parameters, but this is a nonlinear relation to the structural form parameters of interest.

4.1.3 Omitted Variable Bias

Consider the true model, which is called the long regression

$$y_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + u_i, \text{ with } \mathbb{E}(u_i|x_{1i}, x_{2i}) = 0$$

Suppose that we observe only x_{1i} and y_i, and therefore run the following short model, omitting x_{2i}:

$$y_i = x_{1i}'\beta_1 + \varepsilon_i, \text{ with } \varepsilon_i = x_{2i}'\beta_2 + u_i$$

Now, if we consider

$$\mathbb{E}(x_{1i}\varepsilon_i) = \mathbb{E}(x_{1i}(x_{2i}'\beta_2 + u_i)) = \beta_2\mathbb{E}(x_{1i}x_{2i}') \neq 0$$

unless $\beta_2 = 0$ or $\mathbb{E}(x_{1i}x_{2i}') = 0$. This implies that the estimator is not consistent in most cases.

Now, consider the estimator of β_1 written in matrix form:

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y = \beta_1 + (X_1'X_1)^{-1}X_1'\varepsilon = \beta_1 + (X_1'X_1)^{-1}X_1'X_2' + (X_1'X_1)^{-1}X_1'u$$

The last term vanishes by orthogonality condition, but the bias term $(X_1'X_1)^{-1}X_1'X_2'$ remains.

A well known example comes from Card, 2001, where he shows that unobserved ability is an omitted variable when trying to evaluate the causal effect of education on earnings. Individuals with higher ability are likely to be more sucessful on the labore market by earning higher wages, and are likely to acquire more education. As such, unobserved ability affects both education $\left(\mathbb{E}(x_{1i}x_{2i}') \neq 0\right)$ and earnings $(\beta_2 \neq 0)$, and the regressor, education, is correlated with the error term.

4.2 Instrumental Variables and Identification

4.2.1 Set-up

In this section, I introduce one of the solutions to this endogeneity problem: Instrumental Variables (IV).[14] Consider the model

$$y_i = x_i'\beta + \varepsilon_i, \text{ where } \mathbb{E}(x_i\varepsilon_i) \neq 0$$

We first intoduce z_i, which is an instrumental variable, for which we assume $\mathbb{E}(z_i\varepsilon_i) = 0$. Notice that z_i can overlap with x_i, that is any x_i that is exogenous is its own instrument. Note that we must have most instruments than endogenous regressors, that is $dim(z_i) \geq dim(x_i)$.

Now, consider what is called the *reduced form*[15] relationship between x_i and z_i, found by linear projection of x_i on z_i:

$$x_i = z_i'\Pi + u_i$$

If we impose the moment condition $\mathbb{E}(z_i u_i') = 0$, we can recover Π:

$$\mathbb{E}(z_i u_i') = \mathbb{E}(z_i(x_i - z_i'\Pi)) = 0 \implies \left(\mathbb{E}(z_i z_i')\right)^{-1}\mathbb{E}(z_i x_i')$$

In matrix notation, the reduced form can be written as

$$X_{(n\times k)} = Z_{(n\times \ell)}\Pi_{(\ell\times k)} + U_{(n\times k)}$$

and we can consistently Π by OLS:

$$\hat{\Pi} = (Z'Z)^{-1}Z'X$$

If we plug this matrix reduced form equation into the original model, we get

$$y = X\beta + \varepsilon = (Z\Pi + U)\beta + \varepsilon = Z(\Pi\beta) + (U\beta + \varepsilon) \equiv Z\lambda + v$$

Note that λ is a valid projection coefficient since

$$\mathbb{E}(z_i v_i) = \mathbb{E}(z_i u_i')\beta + \mathbb{E}(z_i\varepsilon_i) = 0$$

Therefore, the OLS from y_i on z_i yields a consistent estimator of λ:

$$\hat{\lambda} = (Z'Z)^{-1}Z'y$$

We therefore have the following set of reduced form equations.

[14] I present things in a different order than Limodio's slides, but all the content should be there.

[15] Note that this is not directly what comes to mind when mentioning reduced form: this is more a first stage.

Definition 4.1: Reduced form equations

The reduced form equations in matrix notation are

1. $y = Z\lambda + v$
2. $X = Z\Pi + U$

Understand that so far we have just established:

1. The direct relationship between Z and y (what we think about when talking about reduced form usually): does the instrument have a direct impact on the outcome. For instance, if we instrument education by distance to school, the reduced form is the regression of income on distance to school directly.
2. The relation between X and Z, which is usually what we think about as first stage, and is therefore related to the relevance condition, which will be introduced shortly.

The question is now to understand if we can recover the original parameter of interest β from (λ, Π): this is an **identification question**, that is a population question.

4.2.2 Identification

We say that the parameters β are **identified** if β can be recovered from (λ, Π) through the equation

$$\lambda_{(\ell \times 1)} = \Pi_{(\ell \times k)} \beta_{(\ell \times 1)}$$

It is easy to see that we basically want Π to be invertible. From Proposition 1.3, this is equivalent to Π being full column rank, that is having k linearly independent columns. Therefore, a sufficient condition for identification of β is

$$\text{rank}\Pi = k$$

To build intuition a bit more, recall that Π represents the "coefficients" of the impact of Z on X. Full rank means that there is non-zero covariance between at least k elements of Z and X. That is, at least k instruments should be correlated with the endogenous regressors X: this is a relevance condition. We must now distinguish between different cases.

First, consider the **just-identified** case, where $\ell = k$. The full rank condition guarantees inveribility, so β is identified as

$$\hat{\beta} = \hat{\Pi}^{-1}\hat{\lambda}$$

Indeed, plugging $\hat{\Pi}$ and $\hat{\lambda}$, we obtain

$$\hat{\Pi}^{-1}\hat{\lambda} = \left[(Z'Z)^{-1}Z'X\right]^{-1}(Z'Z)^{-1}Z'y = \left[(Z'Z)^{-1}Z'X\right]^{-1}(Z'Z)^{-1}Z'(X\beta + \varepsilon)$$

$$= \hat{\beta} + \left[(Z'Z)^{-1}Z'X\right]^{-1}(Z'Z)^{-1}Z'\varepsilon = \hat{\beta}_{IV}$$

where the last equality comes by orthogonality of Z' and ε.

In the just-identified case ($\ell = k$), the IV estimator $\hat{\beta}_{\text{IV}} = \hat{\Pi}^{-1}\hat{\lambda}$ formalizes the intuition of "reduced form divided by first stage" that we usually think about when talking about IV. The numerator, $\hat{\lambda}$, is the estimated RF effect—the total impact of the instruments (Z) on the outcome (y). The denominator, $\hat{\Pi}$, is the estimated first-stage effect—the total impact of the instruments (Z) on the endogenous regressor (X). Since the instruments affect the outcome only through the endogenous regressor (the exclusion restriction), the structural coefficient β represents the causal effect of X on y. Therefore, dividing the total effect ($\hat{\lambda}$) by the mechanism's strength ($\hat{\Pi}$) effectively isolates the specific causal effect of interest, $\hat{\beta}_{\text{IV}}$, which is the change in y for a unit change in X induced only by the instrument. This calculation essentially normalizes the total impact by the degree of "relevance" provided by the instruments.

Now, consider the **overidentified** case, where $\ell > k$. We basically treat the reduced form relationship as having an error term **e**:

$$\lambda = \Pi\beta + e$$

and then minimize the sum of squared residuals $e'e$ with respect to β which yields a solution of the form

$$\Pi'\lambda = (\Pi'\Pi)\beta$$

Again, the full rank condition ensures that $(\Pi'\Pi)$ is invertible, which yields

$$\hat{\beta} = (\Pi'\Pi)^{-1}\Pi'\lambda$$

Finally, in the **underidentified** case ($\ell < k$), we do not have enough instrument to explain the endogenous regressors. Note that what really matters again is the rank: if we do not have full column rank in Π, some columns are linear combinations of each other (linear dependence) implying that we "lose" relevant information to instrument the endogenous variables.

4.3 Instrumental Variables and moments conditions

As we hinted at before, the instrumental variable estimator, $\hat{\beta}_{IV}$ is the method of moments estimator that solves, the following moment condition (validity)

$$\mathbb{E}(z_i\varepsilon_i) = \mathbb{E}(z_i(y_i - x_i'\beta)) = 0$$

The sample analog of this moment condition is

$$\frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i'\beta) = 0$$

Assuming the rank condition (relevance), which implies that $\sum_{i=1}^{n} z_i x_i'$ is invertible, the IV estimator can be written as

$$\hat{\beta}_{IV} = \left(\sum_{i=1}^{n} z_i x_i'\right)^{-1} \sum_{i=1}^{n} z_i y_i \text{ or } \hat{\beta}_{IV} = (Z'X)^{-1}Z'y$$

Using the matrix form, we can take the expectation and observe that

$$\mathbb{E}(\hat{\beta}_{IV}|Z,X) = \beta + (Z'X)^{-1}Z'\mathbb{E}(\varepsilon|Z,X)$$

But notice that the main reason we use an IV is because $\mathbb{E}(\varepsilon|X) \neq 0$, which implies $\mathbb{E}(\varepsilon|Z,X) \neq 0$. Therefore, it is important to remember that **the IV estimator is generally biased in finite samples**.

Moving to the asymptotics, as per usual, we can re-write

$$\hat{\beta}_{IV} = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} z_i \varepsilon$$

We impose the following regularity conditions.

- $\{(x_i, z_i, \varepsilon_i)\}_{i=1}^n$ is iid
- Validity: $\mathbb{E}(z_i\varepsilon_i) = 0$
- Relevance: $\mathbb{E}(z_i x_i')$ is invertible
- Moment existence condition: $\mathbb{E}(z_i x_i') < \infty$ and $\mathbb{E}|z_i\varepsilon_i| < \infty$

Therefore, by Khinchine's WLLN, we obtain that

$$\frac{1}{n}\sum_{i=1}^{n} z_i x_i' \xrightarrow{p} \mathbb{E}(z_i x_i') \equiv Q_{IV} \text{ and } \frac{1}{n}\sum_{i=1}^{n} z_i \varepsilon_i \xrightarrow{p} \mathbb{E}(z_i\varepsilon_i) = 0$$

By Slutsky's Theorem and Continuous Mapping Theorem, we obtain

$$\hat{\beta}_{IV} \xrightarrow{p} \beta + \mathbb{E}(z_i x_i')^{-1}\mathbb{E}(z_i\varepsilon_i) = \beta + O_p(1)\cdot o_p(1) = \beta$$

This shows that under **validity** and **relevance** of the instrument, $\hat{\beta}_{IV}$ is consistent.
Moving to asymptotic normality, we impose two additional moment assumptions: $\mathbb{E}(\|z_i\|^4) < \infty$ and $\mathbb{E}(\varepsilon_i^4) < \infty$. Under these assumptions, by (Multivariate) Lindeberg-Lévy's CLT, we have that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i \xrightarrow{d} \mathcal{N}(0, \Omega_{IV}^*), \text{ where } \Omega_{IV}^* = \mathbb{E}(\varepsilon_i^2 z_i z_i')$$

By CMT, we get that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} Q_{IV}^{-1}\mathcal{N}(0, \Omega_{IV}^*) = \mathcal{N}(0, Q_{IV}^{-1}\Omega_{IV}^* Q_{IV}^{-1})$$

We now consider the estimation of the asymptotic variance for the IV estimator, $\hat{\beta}_{IV}$. A consistent estimator of the asymptotic variance is given by the sandwich formula:

$$\widehat{\text{Var}}(\hat{\beta}_{IV}) = \left(\frac{1}{n}\sum z_i x_i'\right)^{-1} \cdot \frac{1}{n}\sum (y_i - x_i'\hat{\beta}_{IV})^2 z_i z_i' \cdot \left(\frac{1}{n}\sum x_i z_i'\right)^{-1}$$

where the middle term estimates the Ω^*_{IV} matrix. The residuals used are the IV residuals, $\hat{\varepsilon}_{IV} = y - X\hat{\beta}_{IV}$.

If the errors are conditionally homoskedastic, meaning $\mathbb{E}(\varepsilon_i^2|z_i) = \sigma^2$ does not depend on z_i, then $\mathbb{E}(\varepsilon_i^2 z_i z_i') = \sigma^2 \mathbb{E}(z_i z_i')$. In this case, the asymptotic variance simplifies to:

$$\widehat{\text{Var}}(\hat{\beta}_{IV}) = s_{IV}^2 \left(\sum \frac{1}{n} z_i x_i' \right)^{-1} \cdot \frac{1}{n} \sum z_i z_i' \cdot \left(\sum \frac{1}{n} x_i z_i' \right)^{-1}$$

where s_{IV}^2 is a consistent estimator of the error variance σ^2:

$$s_{IV}^2 = \frac{\hat{\varepsilon}_{IV}' \hat{\varepsilon}_{IV}}{n - k}$$

To show consistency of s_{IV}^2, we substitute $y = X\beta + \varepsilon$ into the residuals $\hat{\varepsilon}_{IV}$:

$$\hat{\varepsilon}_{IV} = y - X\hat{\beta}_{IV} = y - X(Z'X)^{-1}Z'y = \left(I_n - X(Z'X)^{-1}Z' \right) \varepsilon$$

The consistency of s_{IV}^2 for σ^2 follows from the fact that $\text{plim} \frac{\varepsilon'\varepsilon}{n} = \sigma^2$ and the validity condition.

Example 4.2: Special case: $L = k = 2$

Consider the simple linear model $y_i = \alpha + \beta \tilde{x}_i + \varepsilon_i$, where the regressor $\tilde{x}_i$ is endogenous, $\mathbb{E}(\tilde{x}_i \varepsilon_i) \neq 0$. The vector of regressors is $x_i = (1, \tilde{x}_i)'$ (including the constant). The vector of instruments is $z_i = (1, \tilde{z}_i)'$, assuming the validity condition $\mathbb{E}(\tilde{z}_i \varepsilon_i) = 0$.

The relevance condition holds if the matrix $\mathbb{E}(z_i x_i')$ has full rank, which is equivalent to $\mathbb{C}\text{ov}(\tilde{z}_i, \tilde{x}_i) \neq 0$:[a]

$$\mathbb{E}(z_i x_i') = \begin{pmatrix} 1 & \mathbb{E}(\tilde{x}_i) \\ \mathbb{E}(\tilde{z}_i) & \mathbb{E}(\tilde{z}_i \tilde{x}_i) \end{pmatrix}$$

The IV estimator for the slope coefficient β and the intercept α are given by:

$$\hat{\beta}_{IV} = \frac{\widehat{\mathbb{C}\text{ov}}(\tilde{z}_i, y_i)}{\widehat{\mathbb{C}\text{ov}}(\tilde{z}_i, \tilde{x}_i)} = \frac{\sum(\tilde{z}_i - \bar{\tilde{z}})(y_i - \bar{y})}{\sum(\tilde{z}_i - \bar{\tilde{x}})(\tilde{x}_i - \bar{\tilde{x}})} \quad \text{and} \quad \hat{\alpha}_{IV} = \bar{y} - \hat{\beta}_{IV} \bar{\tilde{x}}$$

Under the assumption of homoskedasticity ($\mathbb{E}(\varepsilon_i^2|z_i) = \sigma^2$), the asymptotic variance of the slope estimator $\hat{\beta}_{IV}$ can be simplified and estimated as:

$$\mathbb{V}(\hat{\beta}_{IV}) = \sigma^2 \frac{\mathbb{V}(\tilde{x}_i)}{\mathbb{C}\text{ov}(\tilde{z}_i, \tilde{x}_i)^2}$$

This can also be expressed using the squared correlation coefficient, $\text{Corr}(\tilde{z}_i, \tilde{x}_i)^2$:

$$\mathbb{V}(\hat{\beta}_{IV}) = \frac{\sigma^2}{\text{Corr}(\tilde{z}_i, \tilde{x}_i)^2 \mathbb{V}(\tilde{x}_i)}$$

For a given distribution of the regressor $\tilde{x}_i$, the variance $\mathbb{V}(\hat{\beta}_{IV})$ has the minimal possible value ($\sigma^2/\mathbb{V}(\tilde{x}_i)$) when the correlation between $\tilde{x}_i$ and $\tilde{z}_i$ is unit. In general, the lower the correlation, the higher the variance of $\hat{\beta}_{IV}$.

1. Good instruments are not only uncorrelated with the regression error (which guarantees the consistency of the IV estimator) but also highly correlated with the explanatory variables (which gives the precision of the IV estimator).

2. The standard errors of IV regression are always higher than the OLS standard errors. We see that by the fact that as mentioned just above, $\mathbb{V}(\hat{\beta}_{IV})$ is minimized whenever $\text{Corr}(\tilde{z}_i, \tilde{x}_i) = 1$, that is whenever $\tilde{x}_i$ is an insturment for itself, the OLS case.

[a] Note that here, $\text{Cov}(\tilde{z}_i, \tilde{x}_i) = \mathbb{E}(\tilde{z}_i\tilde{x}_i) - \mathbb{E}(\tilde{z}_i)\mathbb{E}(\tilde{x}_i) = \det(\mathbb{E}(z_i x_i'))$. Obviously, if $\det(\mathbb{E}(z_i x_i')) = 0$, the matrix is not invertible.

4.4 Two-Stage Least Squares (2SLS)

4.4.1 2SLS derivation

Let us now introduce the Two-Stage Least Squares (2SLS) estimator. We will get back to this when we cover GMM. For now, consider to overidentified case where, for any positive definite matrix $W_{\ell\times\ell}$, β can be expressed as

$$\beta = (\Pi' W \Pi)^{-1}\Pi' W \lambda$$

In the 2SLS case, we use

$$\hat{W} = \frac{1}{n}\sum_{i=1}^{n} z_i z_i' = \frac{1}{n} Z'Z$$

The 2SLS estimator is therefore

$$\tilde{\beta}_{2SLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y = (X'P_Z X)^{-1}X'P_Z y$$

The idea is that we are regressing X on Z first (first stage), to get the fitted values $\hat{X}$ and then regressing y on $\hat{X}$. This is what is called **Theil's interpretation**.

4.4.2 Asymptotic properties of 2SLS

The Two-Stage Least Squares (2SLS) estimator $\tilde{\beta}_{2SLS}$ is a generalization of the IV estimator used in the over-identified case ($\ell > k$). We analyze its consistency using asymptotic theory.
The 2SLS estimator can be written as:

$$\tilde{\beta}_{2SLS} = (\mathrm{X}'\mathrm{P}_z\mathrm{X})^{-1}\mathrm{X}'\mathrm{P}_z y$$

Substituting $y = X\beta + \varepsilon$ and manipulating the terms, we get:

$$\tilde{\beta}_{2SLS} = \beta + (\mathrm{X}'\mathrm{P}_z\mathrm{X})^{-1}\mathrm{X}'\mathrm{P}_z\varepsilon$$

Now, we replace the projection matrix $P_z = Z(Z'Z)^{-1}Z'$ and rewrite the expression using sample averages (multiplying and dividing by n):

$$\tilde{\beta}_{2SLS} = \beta + \left[\frac{1}{n}X'Z\left(\frac{1}{n}Z'Z\right)^{-1}\frac{1}{n}Z'X\right]^{-1}\frac{1}{n}X'Z\left(\frac{1}{n}Z'Z\right)^{-1}\frac{1}{n}Z'\varepsilon$$

We rely on the standard regularity conditions (i.i.d. observations, moment existence) and the two core IV assumptions: validity and relevance. By the WLLN (Khinchine's or similar), the sample averages converge in probability to their population expectations:

$$\begin{aligned}
\operatorname{plim}\frac{1}{n}X'Z &= \operatorname{plim}\frac{1}{n}\sum x_i z_i' = \mathbb{E}(x_i z_i') \equiv Q_{xz} \\
\operatorname{plim}\frac{1}{n}Z'X &= \operatorname{plim}\frac{1}{n}\sum z_i x_i' = \mathbb{E}(z_i x_i') \equiv Q_{zx} \quad \text{(Note: } Q_{zx} = Q_{xz}'\text{)} \\
\operatorname{plim}\frac{1}{n}Z'Z &= \operatorname{plim}\frac{1}{n}\sum z_i z_i' = \mathbb{E}(z_i z_i') \equiv Q_{zz} \\
\operatorname{plim}\frac{1}{n}Z'\varepsilon &= \operatorname{plim}\frac{1}{n}\sum z_i \varepsilon_i = \mathbb{E}(z_i \varepsilon_i) = 0 \quad \text{(Validity assumption)}
\end{aligned}$$

Applying the WLLN and the Continuous Mapping Theorem (CMT) to the asymptotic expression for $\tilde{\beta}_{2SLS}$:

$$\operatorname{plim}\tilde{\beta}_{2SLS} = \beta + \left[Q_{xz}Q_{zz}^{-1}Q_{zx}\right]^{-1}Q_{xz}Q_{zz}^{-1}\operatorname{plim}\left(\frac{1}{n}Z'\varepsilon\right)$$

Substituting the validity condition $\operatorname{plim}(\frac{1}{n}Z'\varepsilon) = 0$:

$$\operatorname{plim}\tilde{\beta}_{2SLS} = \beta + \left[Q_{xz}Q_{zz}^{-1}Q_{zx}\right]^{-1}Q_{xz}Q_{zz}^{-1}\cdot 0$$

$$\operatorname{plim}\tilde{\beta}_{2SLS} = \beta$$

This shows that the 2SLS estimator $\tilde{\beta}_{2SLS}$ is consistent under the standard validity and relevance conditions. The relevance condition ensures that the matrix $[Q_{xz}Q_{zz}^{-1}Q_{zx}]$ is invertible.

Having established the consistency of the 2SLS estimator $\tilde{\beta}_{2SLS}$, we now move to its asymptotic normality. We start with the expression for the difference between the 2SLS estimator and the true parameter vector, scaled by $\sqrt{n}$:

$$\sqrt{n}(\tilde{\beta}_{2SLS} - \beta) = \left[\frac{1}{n}X'Z\left(\frac{1}{n}Z'Z\right)^{-1}\frac{1}{n}Z'X\right]^{-1}\frac{1}{n}X'Z\left(\frac{1}{n}Z'Z\right)^{-1}\frac{1}{\sqrt{n}}Z'\varepsilon$$

Suppose the sample is i.i.d. is we have the valid moment conditions. By the Multivariate Lindeberg-Lévy Central Limit Theorem, the scaled sum of the moment condition terms converges in distribution to a normal distribution:

$$\frac{1}{\sqrt{n}}Z'\varepsilon = \frac{1}{\sqrt{n}}\sum z_i\varepsilon_i \xrightarrow{d} \mathcal{N}(0, \Omega_{zz})$$

where the asymptotic variance of the moment vector is $\Omega_{zz} = \mathrm{V}(z_i\varepsilon_i) = \mathbb{E}(\varepsilon_i^2 z_i z_i')$.

We apply the Continuous Mapping Theorem (CMT) by taking the probability limit of the deterministic (non-stochastic) parts of the $\sqrt{n}(\tilde{\beta}_{2SLS} - \beta)$ expression, and substituting the CLT result for the stochastic part. Recall the probability limits:

$$\begin{aligned}
\operatorname{plim} \frac{1}{n} \mathrm{X}'\mathrm{Z} &= Q_{xz} \\
\operatorname{plim} \frac{1}{n} \mathrm{Z}'\mathrm{Z} &= Q_{zz} \\
\operatorname{plim} \frac{1}{n} \mathrm{Z}'\mathrm{X} &= Q_{zx}
\end{aligned}$$

Applying these limits to the scaled expression:

$$\sqrt{n}(\tilde{\beta}_{2SLS} - \beta) \xrightarrow{d} \left[Q_{xz} Q_{zz}^{-1} Q_{zx} \right]^{-1} Q_{xz} Q_{zz}^{-1} \cdot \mathcal{N}(0, \Omega_{zz})$$

Therefore, the asymptotic distribution is:

$$\sqrt{n}(\tilde{\beta}_{2SLS} - \beta) \xrightarrow{d} \mathcal{N}(0, V_{2SLS})$$

with the asymptotic variance V_{2SLS} given by the generalized "sandwich" formula:

$$V_{2SLS} \equiv \left[Q_{xz} Q_{zz}^{-1} Q_{zx} \right]^{-1} Q_{xz} Q_{zz}^{-1} \Omega_{zz} Q_{zz}^{-1} Q_{zx} \left[Q_{xz} Q_{zz}^{-1} Q_{zx} \right]^{-1}$$

This is the standard, heteroskedasticity-robust formula for the asymptotic variance of the 2SLS estimator. Notice that the 2SLS estimator is inefficient comapred to OLS. The variance is higher by design.

To use the derived asymptotic normality for inference (t-tests, confidence intervals), we need a consistent estimator for the asymptotic variance V_{2SLS}. We obtain such consistent estimator of the asymptotic variance V_{2SLS} by replacing all population moment matrices (Q_{xz}, Q_{zz}, and Ω_{zz}) with their consistent sample analogs.

The sample analog for the matrix $\Omega_{zz} = \mathbb{E}(\varepsilon_i^2 z_i z_i')$ is obtained by using the 2SLS residuals $\tilde{\varepsilon}_{i,2SLS} = y_i - x_i'\tilde{\beta}_{2SLS}$:

$$\frac{1}{n} \sum (y_i - x_i'\tilde{\beta}_{2SLS})^2 z_i z_i'$$

The consistent estimator $\hat{V}_{2SLS}$ is then the robust (Heteroskedasticity-Consistent) sandwich estimator, obtained by substituting the sample analogs into the general formula for V_{2SLS}.

If we assume conditional homoskedasticity of the errors with respect to the instruments, i.e., $\mathbb{E}(\varepsilon_i^2 | z_i) = \sigma^2$, the asymptotic variance simplifies. Under homoskedasticity, the middle matrix Ω_{zz} simplifies to:

$$\Omega_{zz} = \mathbb{E}(\varepsilon_i^2 z_i z_i') = \sigma^2 \mathbb{E}(z_i z_i') = \sigma^2 Q_{zz}$$

Substituting this into the V_{2SLS} formula and simplifying yields:

$$V_{2SLS} = \sigma^2 \left[Q_{xz} Q_{zz}^{-1} Q_{zx} \right]^{-1}$$

We replace the unknown population variance σ^2 with its consistent sample estimator, s^2:

$$s^2 = \frac{(y - \mathrm{X}\tilde{\beta}_{2SLS})'(y - \mathrm{X}\tilde{\beta}_{2SLS})}{n - k} = \frac{\tilde{\varepsilon}'_{2SLS}\tilde{\varepsilon}_{2SLS}}{n - k}$$

The final homoskedastic estimator for the asymptotic variance is obtained by substituting s^2 and the sample moments for Q_{xz}, Q_{zz}, and Q_{zx}:

$$\hat{V}_{2SLS} = s^2 \left[\frac{1}{n}\mathrm{X}'\mathrm{Z} \left(\frac{1}{n}\mathrm{Z}'\mathrm{Z} \right)^{-1} \frac{1}{n}\mathrm{Z}'\mathrm{X} \right]^{-1}$$

4.5 Validity Tests and Weak Instruments

4.5.1 Testing for the Validity of Instruments: The J-Test

In the over-identified case ($\ell > k$), we have more instruments ($\mathbf{z}_i$, dimension ℓ) than endogenous regressors (x_i, dimension k). This provides $\ell - k$ extra moment conditions that can be used to test the validity of the instruments.

The core idea is that if the population moment conditions $\mathbb{E}(z_i\varepsilon_i) = 0$ are true (the null hypothesis $\mathcal{H}_0$), the sample analog $\frac{1}{n}\sum z_i(y_i - x_i'\tilde{\beta}_{2SLS})$ should be close to zero.
This provides a basis for a model specification test of the null hypothesis ($\mathcal{H}_0$) against the alternative hypothesis ($\mathcal{H}_1$):

$$\mathcal{H}_0 : \mathbb{E}(z_i\varepsilon_i) = 0 \quad \text{versus} \quad \mathcal{H}_1 : \mathbb{E}(z_i\varepsilon_i) \neq 0$$

The test statistic is known as the J_n statistic (or Sargan's test in the homoskedastic case, or Hansen's J test in the heteroskedastic/GMM case):

$$J_n = n \left(\frac{1}{n}\sum z_i(y_i - x_i'\tilde{\beta}_{2SLS}) \right)' \left(s^2 \frac{1}{n}\sum z_i z_i' \right)^{-1} \left(\frac{1}{n}\sum z_i(y_i - x_i'\tilde{\beta}_{2SLS}) \right)$$

Under the null hypothesis $\mathcal{H}_0$, the J_n test statistic is asymptotically distributed as a chi-squared distribution with degrees of freedom equal to the number of overidentifying restrictions:

$$J_n \xrightarrow{d} \chi^2(\ell - k) \quad \text{under } \mathcal{H}_0$$

The J_n test is a test of the overidentifying restrictions. If we reject $\mathcal{H}_0$ (the p-value is small), it suggests that at least some of the moment conditions are invalid, implying that at least some of the instruments $\mathbf{z}_i$ are not truly exogenous. If the errors are conditionally heteroskedastic, the test must be constructed differently and should be based on an efficient GMM estimator, which will be discussed in the GMM topic.

4.5.2 Weak Instruments

Instruments are considered weak when they are only weakly correlated with the endogenous explanatory variables they instrument. They fail to provide sufficient variation to effectively isolate the exogenous component of the endogenous regressors. This can lead to biased and inconsistent parameters estimates, as well as inflated standard errors, making inference unreliable.

A way to test for weak instruments is to use a First-stage F-statistic. It aims at assessing the joint significance of the instruments in the first-stage regression. The initial rule of thumb, determined by Staiger and Stock, 1994 was an F-stat above 10 was enough to claim the instrument is not weak. New papers such as Lee et al., 2022 claims that is should actually be 100.

5 GMM and Extremum Estimation

In this section, we talk about the Generalized method of moments (GMM), and introduce extremum estimators.

5.1 Intro to GMM

Many estimation methods in econometrics are method-of-moments estimators, in which the k-dimensional parameter of interest θ_0 is assumed to satisfy an unconditional moment condition

$$\mathbb{E}(g(z_i, \theta_0)) = 0$$

for some J-dimensional vector of functions $g(z_i, \theta_0)$ of the observed data vector z_i and the parameter value $\theta \in \Theta$. If we assume that θ_0 is the unique solution of the population moment, and we are in a just-identified case ($J = k$). The method of moment estimator $\hat{\theta}$ is defined as a solution to the sample analogue of the population moment condition:

$$\frac{1}{n}\sum_{i=1}^{n} g(z_i, \hat{\theta}) = 0$$

The OLS estimation derived in section 1.2.2 is a perfect example of a method of moment estimator in this context, with the population moment condition being

$$\mathbb{E}(x_i(y_i - x_i'\beta)) = 0^{16}$$

Similarly, the IV estimator (denoting ζ_i the vector of instruments) is based on the moment condition

$$\mathbb{E}(\zeta_i(y_i - x_i'\beta)) = 0$$

The MLE estimator can also be written as a method-of-moments estimators

$$\mathbb{E}(s(z_i, \theta_0))$$

[16] Just to understand notation better, in this case we have $z_i \equiv (x_i', y_i')$ and $g(z_i, \beta) \equiv x_i(y_i - x_i'\beta)$

where $s(z_i, \theta_0)$ is the score function.

So far, we have dealt with cases where the number of equations is the same as the number of unknown parameters. Now, consider the overidentified where we have more moment conditions that parameters to estimate ($J > k$). For instance, we might have knowledge that the distribution of a scalar random variable z_i is not skewed, which allows us to have the moment $\mathbb{E}\left((z - \mathbb{E}(z))^3\right) = 0$. If we want t0 estimate $\theta_0 = \mathbb{E}(z)$, we have the following moment condition

$$\mathbb{E}\left[\begin{pmatrix} z_i - \theta_0 \\ (z_i - \theta_0)^3 \end{pmatrix}\right] = 0$$

In general, whenever $J > k$, the system of equations

$$g_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta) = 0$$

is overdetermined, meaning that there is no solution to this system. We provide the solution below.

Definition 5.1: Generalized Method of Moments Estimator

The Generalized Method of Moments (GMM) Estimator is defined as the estimator minimizing the weighted Euclidean norm, that is

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} g_n(\theta)' W_n g_n(\theta) \equiv \arg\min_{\theta \in \Theta} J_n(\theta)$$

where W_n is a symmetric positive matrix which converges in probability to a positive definite matrix W_0.

The idea is that since we are in an over-identified case, we can't make each moment equal to ero in the sample. Instead, we try to minimize the sum of squares, that is we try to make all theoretical moments close to zero. The weight matrix just gives more importance to the moments that we theoretically believe are true.

Consider the basic linear model. We can write

$$g_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i'\beta) = \frac{1}{n}Z'(y - X\beta) \implies J_n(\beta) = \left[\frac{1}{n}Z'(y - X\beta)\right]' W_n \left[\frac{1}{n}Z'(y - X\beta)\right]$$

Taking the first-order condition with respect to β, we get[17]

$$\frac{\partial J_n(\beta)}{\partial \beta} = 0 \Leftrightarrow 2\left[\frac{\partial g_n(\hat{\beta})}{\partial \beta}\right]' W_n g_n(\hat{\beta}) = 0 \Leftrightarrow \left[\frac{1}{n}Z'X\right]' W_n \left[\frac{1}{n}Z'(Y - X\beta)\right] = 0$$

[17] We use the following matrix multiplication rule:

$$\frac{\partial}{\partial x}\left[f(x)'Bf(x) = 2\frac{\partial f(x)}{\partial x}' Bf(x)\right]$$

$$\Leftrightarrow \frac{1}{n}(X'Z)W_n\frac{1}{n}Z'Y = \frac{1}{n}(X'Z)W_n\left[\frac{1}{n}Z'X\hat{\beta}\right] \Leftrightarrow \hat{\beta} = \left[X'ZW_nZ'X\right]^{-1}X'ZW_nZ'y$$

Now, notice that if we impose $W_n = I_n$ and $Z = X$, we obtain that

$$\hat{\beta} = (X'XI_nX'X)^{-1}X'XI_nX'y = (X'X)^{-1}X'y = \hat{\beta}_{OLS}$$

If we impose $W_n = (Z'Z)^{-1}$, we obtain

$$\hat{\beta} = \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1}X'Z(Z'Z)^{-1}Z'y = \hat{\beta}_{2SLS}$$

With these easy examples, we see that the different estimators are just version of GMM with a different weight matrix W_n. This leads us to our next concept, extremum estimators.

5.2 Extremum Estimators

The concept of extremum estimators is a class of estimator that is derived by finding extreme values - either a maximum of a minimum- of some objective function $Q_n(\theta)$:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} Q_n(\theta)$$

We discuss general asymptotic theory based on this class of estimators.

5.2.1 General Consistency

We start by introducing the general consistency theorem, which I prove (but this can be skipped).

Theorem 5.1: General Consistency Theorem (Newey and McFadden, 1994)

Suppose that

1. Θ is compact
2. The limit function Q_* is uniquely maximized at θ_0 (Identification)
3. Q_* is continuous in $\theta \in \Theta$
4. $\sup_{\theta\in\Theta} \mid Q_n(\theta) - Q_*(\theta) \mid \xrightarrow{p} 0$ for some $Q_* : \Theta \to \mathbb{R}$ (Uniform Convergence)

Then, we acheive consistency:

$$\hat{\theta} \xrightarrow{p} \theta_0$$

Proof. Pick any $\varepsilon > 0$. Since $\hat{\theta}$ maximizes $Q_n(\theta)$ by definition, we have that

$$Q_n(\hat{\theta}) > Q_n(\theta_0) - \frac{\varepsilon}{3}$$

By condition 4, for any $\theta \in \Theta$, we have that

$$| Q_n(\theta) - Q_*(\theta) | < \frac{\varepsilon}{3} \text{ with probability approaching 1 (w.p.a.1)}^{18}$$

Therefore, w.p.a.1, we have

$$| Q_n(\hat{\theta}) - Q_*(\hat{\theta}) | < \frac{\varepsilon}{3} \implies Q_n(\hat{\theta}) - Q_*(\hat{\theta}) < \frac{\varepsilon}{3} \text{ since } (Q_n(\hat{\theta}) - Q_*(\hat{\theta})) > 0 \text{ and}$$

$$| Q_n(\theta_0) - Q_*(\theta_0) | < \frac{\varepsilon}{3} \implies Q_*(\theta_0) - Q_n(\theta_0) < \frac{\varepsilon}{3} \text{ since } (Q_n(\theta_0) - Q_*(\theta_0)) < 0$$

Combining these inequalities, w.p.a.1, we get

$$Q_*(\hat{\theta}) + \frac{\varepsilon}{3} > Q_n(\hat{\theta}) > Q_n(\theta_0) - \frac{\varepsilon}{3} > Q_*(\theta_0) - \frac{2\varepsilon}{3} \implies Q_*(\hat{\theta}) > Q_*(\theta_0) - \varepsilon \quad (\star)$$

By definition of convergence in probability, we want $\mathbb{P}(\hat{\theta} \in \mathcal{N}) \to 1$ for any open neighborhood $\mathcal{N} \subset \Theta$ containing θ_0. Pick such set. Since $\mathcal{N}$ is open, $\mathcal{N}^c$ is closed. Additionally, by condition 1, Θ is compact, which implies that $\Theta \cap \mathcal{N}^c$ is also compact. Since Q_* is continuous (by condition 3), Weierstrass theorem guarantees that there exists $\theta_* \in \Theta \cap \mathcal{N}^c$ such that

$$\sup_{\Theta \cap \mathcal{N}^c} Q_*(\theta) = Q_*(\theta_*)$$

Since Q_* is uniquely maximized at θ_0 by condition 4, we know that $Q_*(\theta_0) > Q_*(\theta_*)$. Set $\varepsilon' := Q_*(\theta_0) - Q_*(\theta_*) > 0$. Now, if we use equation $(\star)$ and set $\varepsilon = \varepsilon'$, we get

$$Q_*(\hat{\theta}) > Q_*(\theta_0) - \varepsilon' = Q_*(\theta_*) = \sup_{\Theta \cap \mathcal{N}^c} Q_*(\theta) \text{ w.p.a.1}$$

This implies that $\hat{\theta} \in \mathcal{N}$ with probability approaching 1, so we are done. □

The proof is not easy, so I try to give some intuition below. The estimator $\hat{\theta}$ is chosen because it makes the sample objective function ($Q_n(\theta)$) as high as possible. Therefore, $Q_n(\hat{\theta})$ must be slightly greater than $Q_n(\theta_0)$.

The key step is establishing that the sample function $Q_n(\theta)$ is an excellent approximation of the true function $Q_*(\theta)$ everywhere (Uniform Convergence). This means that if we evaluate the true function Q_* at the sample peak $\hat{\theta}$, the result $Q_*(\hat{\theta})$ must be very close to the true peak value $Q_*(\theta_0)$. That's the meaning of the inequality $Q_*(\hat{\theta}) > Q_*(\theta_0) - \varepsilon$: the $\hat{\theta}$ chosen by the sample must give an outcome on the true function that is nearly optimal.

Finally, we use the Identification condition, which states that θ_0 is the unique point that maximizes Q_*. Because no other point $\theta \neq \theta_0$ can give a value of Q_* as high as $Q_*(\theta_0)$, the $\hat{\theta}$ that gives a nearly optimal value $Q_*(\hat{\theta})$ must itself be forced into a small neighborhood

[18] This holds because uniform convergence implies pointwise convergence. Additionally, w.p.a.1 is equivalent to convergence in probability. Recall that an event event A_n holds with probability approaching 1 means $\lim_{n\to\infty} P(A_n) = 1$.

around θ_0. If $\hat{\theta}$ were far away, the uniqueness condition would guarantee that $Q_*(\hat{\theta})$ would be significantly lower, which contradicts our finding from step 2.

In practice, most efforts are devoted to check condition 2 and 4. For condition 4, we typically need some kind of uniform law of large numbers.

Lemma 5.1: Uniform Law of Large Numbers (ULLN)

Suppose that

1. $\{z_i\}_{i=1}^n$ is i.i.d.
2. $g(z, \theta)$ is almost surely continuous at each $\theta \in \Theta$, and Θ is compact.
3. There is $d(z)$ such that $|g(z, \theta)| \leq d(z)$ for all $\theta \in \Theta$ and almost every z, and $E[d(z)] < \infty$.

Then, we have the uniform convergence condition:

$$\sup_{\theta \in \Theta} \mid \bar{g}(\theta) - E[g(z, \theta)] \mid \xrightarrow{p} 0$$

where $\bar{g}(\theta) = \frac{1}{n} \sum_{i=1}^n g(z_i, \theta)$ is the sample counterpart of the population moment $E[g(z, \theta)]$. Additionally, $E[g(z, \theta)]$ is continuous at each $\theta \in \Theta$.

Note that if we take $\frac{1}{n} \sum_{i=1}^n g(z_i)$, we are back in the usual WLLn case, where we need to assume $\mathbb{E}(g(z)) < \infty$. Here, we have that $g(\cdot)$ is a function of θ, so we need to assume something stronger. Indeed, to guarantee uniform convergence, the moment condition needs to hold for the worst possible value of the function across the entire parameter space. The condition $\mathbb{E}(\sup_{\theta \in \Theta} |g(z, \theta)|) < \infty$ ensures that the function $g(z, \theta)$ is well-behaved over the entire compact set Θ. This condition is known as the dominance condition.

5.2.2 General asymptotic normality

Suppose that we have achieved consistency. We now want to derive the asymptotic normal distribution in the form

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$$

We derive some the asymptotic distribution and then provide the formal theorem. Suppose that $Q_n(\theta)$ is continuously twice differentiable. We can look at the FOC for $\hat{\theta}$:

$$\frac{\partial Q_n(\hat{\theta})}{\partial \theta} = 0$$

We can write the Mean Value Theorem (MVT) around θ_0 as follows:

$$0 = \frac{\partial Q_n(\theta_0)}{\partial \theta} + \frac{\partial^2 Q_n(\tilde{\theta})}{\partial \theta \partial \theta'} (\hat{\theta} - \theta_0)$$

where $\tilde{\theta}$ is a point on the line joining θ_0 and $\hat{\theta}$. Assuming the inverse exists, we can therefore rewrite

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left(\frac{\partial^2 Q_n(\tilde{\theta})}{\partial\theta\partial\theta'}\right)^{-1} \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}$$

We now move on to the theorem to understand what conditions are necessary and sufficient for asymptotic normality (I do not prove this theorem).

Theorem 5.2: General asymptotic normlality theorem (Newey and McFadden, 1994)

Suppose that

a. $\hat{\theta} \xrightarrow{p} \theta_0$ and $\theta_0 \in \text{int}\Theta$

b. $Q_n(\theta)$ is continuously twice differentiable in a neighborhood $\mathcal{N}$ of θ_0

c. $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} \xrightarrow{d} \mathcal{N}(0, \Sigma)$

d. There exists $H(\theta)$ that is continuous at θ_0 such that

$$\sup_{\theta\in\mathcal{N}} \mid \frac{\partial^2 Q_n(\tilde{\theta})}{\partial\theta\partial\theta'} - H(\theta) \mid \xrightarrow{p} 0$$

and $H = H(\theta_0)$ non-singular.

Then, we obtain asymptotic normality

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Sigma H^{-1})$$

5.3 Linking the two concepts

As we have shown above, the GMM estimator is nothing but an extremum estimator where

$$Q_n(\theta) = -J_n(\theta) = -g_n(\theta)'\hat{W}g_n(\theta) \text{ and } \hat{\theta}_W = \arg\max_{\theta\in\Theta} Q_n(\theta)$$

Theorem 5.1 and 5.2 can be applied directly, but it would be nice to have some more primitive conditions. The following theorem provides the primitive conditions for consistency.

Theorem 5.3: Consistency of GMME

Suppose $\{z_i\}_{i=1}^n$ is iid and

(i) Θ is compact

(ii) $\hat{W} \xrightarrow{p} W_0$ and W_0 is symmetric and positive definite

(iii) $g(z, \theta)$ is almost surely continuous at each $\theta \in \Theta$

(iv) $\mathbb{E}(\sup_{\theta\in\Theta}|g(z,\theta)|) < \infty$

(v) $W_0\mathbb{E}(g(z,\theta)) = 0$ only if $\theta = \theta_0$

Proof. It is sufficient to check conditions 1-4 of Theorem 5.1. Condition 1 is guaranteed by i). We now check condition 4. In this context, we have that

$$Q_*(\theta) = -\mathbb{E}(g(z,\theta))'W\mathbb{E}(g(z,\theta))$$

We want to show that

$$\sup_{\theta\in\Theta}|Q_n(\theta) - Q_*(\theta)| = \sup_{\theta\in\Theta}|J_n(\theta) - J_*(\theta)| \xrightarrow{p} 0$$

We can rewrite the expression in the absolute value as

$$|Q_n(\theta) - Q_*(\theta)| = |g_n(\theta)'\hat{W}_n g_n(\theta) - \mathbb{E}[g(z,\theta)]'W_0\mathbb{E}[g(z,\theta)]|$$

Let $\bar{g}(\theta) = g_n(\theta)$ and $\mu(\theta) = \mathbb{E}[g(z,\theta)]$. Also, let $\hat{W} = \hat{W}_n$ and $W = W_0$. Therefore, we have something of the form

$$|Q_n(\theta) - Q_*(\theta)| = |\bar{g}(\theta)'\hat{W}\bar{g}(\theta) - \mu(\theta)'W\mu(\theta)|$$

We use the identity $\bar{g}(\theta) = \mu(\theta) + [\bar{g}(\theta) - \mu(\theta)]$ and substitute it into the sample objective term:

$$|(\mu + (\bar{g} - \mu))'\hat{W}(\mu + (\bar{g} - \mu)) - \mu'W\mu|$$

Expanding the quadratic form:[19]

$$|\mu'\hat{W}\mu + \mu'\hat{W}(\bar{g} - \mu) + (\bar{g} - \mu)'\hat{W}\mu + (\bar{g} - \mu)'\hat{W}(\bar{g} - \mu) - \mu'W\mu|$$

We rearrange these terms to isolate the errors in the moments and the errors in the weighting matrix:

$$|\{(\bar{g} - \mu)'\hat{W}(\bar{g} - \mu)\} + \{\mu'(\hat{W} + \hat{W}')(\bar{g} - \mu)\} + \{\mu'(\hat{W} - W)\mu\}|$$

Using the Triangle Inequality[20], we separate the terms into the T_1, T_2, T_3:

$$|Q_n(\theta) - Q_*(\theta)| \leq T_1(\theta) + T_2(\theta) + T_3(\theta)$$

where:

- $T_1(\theta) = |\{\bar{g}(\theta) - \mu(\theta)\}'\hat{W}\{\bar{g}(\theta) - \mu(\theta)\}|$
- $T_2(\theta) = |\mu(\theta)'(\hat{W} + \hat{W}')\{\bar{g}(\theta) - \mu(\theta)\}|$
- $T_3(\theta) = |\mu(\theta)'(\hat{W} - W)\mu(\theta)|$

[19] $(A+B)'W(A+B) = A'WA + A'WB + B'WA + B'WB$

[20] $|a+b+c| \leq |a| + |b| + |c|$

Now, $\sup_{\theta\in\Theta} T_1(\theta) \xrightarrow{p} 0$ and $\sup_{\theta\in\Theta} T_2(\theta) \xrightarrow{p} 0$ by ULLN, which is guaranteed by iid assumption as well as i), iii) and iv). From ii), we obtain that $\sup_{\theta\in\Theta} T_3(\theta) \xrightarrow{p} 0$, and condition 4 is therefore verified. ULLN also implies that $\mathbb{E}[g(z,\theta)]$ is continuous in $\theta \in \Theta$, which in turns implies that $Q_*(\theta)$ is also continuous in $\theta \in \Theta$, so condition 3 is verified.

Next, we verify condition 2: the limit function Q_* is uniquely maximized at θ_0. By the population moment condition, $\mathbb{E}[g(z,\theta_0)] = 0$, which implies $Q_*(\theta_0) = 0$. Since W_0 is positive definite, we have $Q_*(\theta) \leq 0$ for all $\theta \in \Theta$. Thus, it is sufficient to show that $Q_*(\theta) < 0$ for all $\theta \neq \theta_0$.

Since W_0 is symmetric and positive definite, there exists a matrix R such that $W_0 = R'R$. For any $\theta \neq \theta_0$, assumption (v) implies $W_0\mathbb{E}[g(z,\theta)] \neq 0$. This further implies that:

$$R\mathbb{E}[g(z,\theta)] \neq 0$$

Therefore, for any $\theta \neq \theta_0$:

$$\begin{aligned} Q_*(\theta) &= -\mathbb{E}[g(z,\theta)]'W_0\mathbb{E}[g(z,\theta)] \\ &= -\{R\mathbb{E}[g(z,\theta)]\}'\{R\mathbb{E}[g(z,\theta)]\} \\ &< 0 \end{aligned}$$

where the strict inequality follows from the fact that the inner product of a non-zero vector with itself is strictly positive. Thus, θ_0 is the unique maximizer of Q_*, and Condition 2 is verified. □

This proof is hard, but allows us to see how setting primitive conditions yields the nice and general properties mentioned initially. We are now moving to the asymptotic normality counterpart. This is more difficult, but I really do find helpful to see how the asymptotic variance is derived.

Theorem 5.4: Asymptotic Normality of GMM

Suppose the assumptions for GMM consistency are satisfied. Additionally, assume:

(i) $\theta_0 \in \operatorname{int}\Theta$

(ii) $g(z,\theta)$ is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 with probability one

(iii) $\mathbb{E}[|g(z,\theta_0)|^2] < \infty$, $\mathbb{E}[|\frac{\partial g(z,\theta_0)}{\partial\theta'}|] < \infty$, and $\mathbb{E}\left[\sup_{\theta\in\mathcal{N}}\left|\frac{\partial^2 g^{(j)}(z,\theta)}{\partial\theta\partial\theta'}\right|\right] < \infty$ for $j = 1,\ldots,\dim g$

(iv) $G'WG$ is nonsingular, where $G = \mathbb{E}\left[\frac{\partial g(z,\theta_0)}{\partial\theta'}\right]$

Then, the GMM estimator is asymptotically normal:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}\right)$$

where $\Omega = E[g(z,\theta_0)g(z,\theta_0)']$ is the asymptotic variance of the moment conditions.

Proof. We provide a proof by construction, which is a bit different from what we have done in the last proof. We are not going to show why each primitive condition implies that 5.2 holds. Instead, we will construct the asymptotic distribution, highlighting where the assumptions come to play.

Consider $Q_n(\theta) = -\frac{1}{2}\left[g_n(\theta)'\hat{W}g_n(\theta)\right] \equiv f(g(\theta))$. Assuming i) and ii) allows us to write the MVT:

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left(\frac{\partial^2 Q_n(\tilde{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}$$

Let us take the first derivative:

$$\frac{\partial Q_n(\theta)}{\partial\theta} = -\left(\frac{\partial g_n(\theta)'}{\partial\theta}\right)\left(\frac{\hat{W}+\hat{W}'}{2}\right)g_n(\theta) = -\left(\frac{\partial g_n(\theta)}{\partial\theta'}\right)'\left(\frac{\hat{W}+\hat{W}'}{2}\right)g_n(\theta)$$

where the first equality comes from the Chain rule[21] and the second inequality from the Transpose rule.[22] Now, if we the second assumption of iii), that is $\mathbb{E}[|\frac{\partial g(z,\theta_0)}{\partial\theta'}|] < \infty$, we have that

$$\frac{\partial g_n(z,\theta_0)}{\partial\theta'} \xrightarrow{p} \mathbb{E}\left[\frac{\partial g(z,\theta_0)}{\partial\theta'}\right] \equiv G$$

Additionally, by consistency assumptions,

$$\frac{\hat{W}+\hat{W}'}{2} \xrightarrow{p} W \text{ by WLLN}$$

Now, we move to the variance, which can be written as follows by iid

$$\mathbb{V}(g_n(z_i,\theta_0)) - \mathbb{E}(g_n(z_i,\theta_0)g_n(z_i,\theta_0)') \equiv \Omega < \infty$$

It is finite by the first assumption of iii), $\mathbb{E}[|g(z,\theta_0)|^2] < \infty$ and Cauchy-Schwarz inequality. By Multivariate CLT, we have that

$$\sqrt{n}g_n(z_i,\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(z_i,\theta_0) \xrightarrow{d} \mathcal{N}(0,\Omega)$$

[21] To take the derivative with respect to θ, we use the rule for a quadratic form $f(u) = u'Mu$ where u is a function of θ:

$$\frac{\partial f}{\partial\theta} = \frac{\partial u'}{\partial\theta}Mu + \frac{\partial u'}{\partial\theta}M'u = \frac{\partial u'}{\partial\theta}(M+M')u$$

[22] Transpose rule:

$$\frac{\partial\bar{g}(\theta)'}{\partial\theta} = \left(\frac{\partial\bar{g}(\theta)}{\partial\theta'}\right)'$$

Therefore, by CMT, we have that

$$\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \xrightarrow{d} -G'W \cdot \mathcal{N}(0,\Omega) = \mathcal{N}(0, G'W\Omega WG) \equiv \mathcal{N}(0,\Sigma)$$

We can move to the denominator. Differentiating again with respect to θ', we treat the score as a product of two functions of θ: the Jacobian term and the moment term. Using the product rule $(f \cdot g)' = f'g + fg'$, the j-th column of the Hessian (derivative with respect to θ_j) is expressed as:

$$\frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta_j} = \underbrace{-\left(\frac{\partial^2 g_n(\theta)'}{\partial\theta\partial\theta_j}\right)\left(\frac{\hat{W}+\hat{W}'}{2}\right)g_n(\theta)}_{T_1(\theta)} \underbrace{-\left(\frac{\partial g_n(\theta)'}{\partial\theta}\right)\left(\frac{\hat{W}+\hat{W}'}{2}\right)\left(\frac{\partial g_n(\theta)}{\partial\theta_j}\right)}_{T_2(\theta)}$$

Now, assuming the last part of iii), that is $\mathbb{E}\left[\sup_{\theta\in\mathcal{N}}\left|\frac{\partial^2 g^{(j)}(z,\theta)}{\partial\theta\partial\theta'}\right|\right] < \infty$ for $j = 1,\dots,\dim g$, we have by ULLN that

$$T_1(\theta) = O_p(1)\cdot O_p(1)\cdot o_p(1) = o_p(1)$$

noting that $\mathbb{E}(g(z,\theta_0)) = 0$ by GMM assumption. We also have that, by ULLN,

$$T_2(\theta) \xrightarrow{p} G'WG \equiv H$$

which we assume to be non-singular in assumption iv).

Therefore, by CMT, we obtain the desired result

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Sigma H^{-1}) \text{ with } H^{-1}\Sigma H^{-1} = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$$

□

5.4 Optimal GMM

We can now introduce the optimal weighting matrix, that is the weighting matrix that minimizes the asymptotic variance.

Theorem 5.5: Optimal Weighting Matrix for GMM

The GMM asymptotic variance

$$V(W) = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$$

is minimized in the matrix sense (i.e., $V(W) - V(\Omega^{-1})$ is positive semi-definite) by choosing $W = \Omega^{-1}$. Under this choice, the variance collapses to:

$$\text{Avar}(\hat{\theta}) = (G'\Omega^{-1}G)^{-1}$$

Proof. We want to show that for any symmetric positive definite matrix W, $V(W) - V(\Omega^{-1})$ is a positive semi-definite matrix.

Let $A = (G'WG)^{-1}G'W$ and $B = G$. Note that by construction, $AB = I$. We are comparing:

$$V(W) = A\Omega A' \quad \text{and} \quad V(\Omega^{-1}) = (G'\Omega^{-1}G)^{-1}$$

Consider the following matrix, which is positive semi-definite by construction (as it is of the form XX'):

$$(A - (G'\Omega^{-1}G)^{-1}G'\Omega^{-1})\Omega(A - (G'\Omega^{-1}G)^{-1}G'\Omega^{-1})' \geq 0$$

Expanding this product:

$$\begin{aligned} &A\Omega A' - A\Omega\Omega^{-1}G(G'\Omega^{-1}G)^{-1} - (G'\Omega^{-1}G)^{-1}G'\Omega^{-1}\Omega A' \\ &+ (G'\Omega^{-1}G)^{-1}G'\Omega^{-1}\Omega\Omega^{-1}G(G'\Omega^{-1}G)^{-1} \end{aligned}$$

Using the fact that $\Omega\Omega^{-1} = I$ and $AG = I$:

$$\begin{aligned} &= A\Omega A' - (AG)(G'\Omega^{-1}G)^{-1} - (G'\Omega^{-1}G)^{-1}(G'A') + (G'\Omega^{-1}G)^{-1}(G'\Omega^{-1}G)(G'\Omega^{-1}G)^{-1} \\ &= A\Omega A' - (G'\Omega^{-1}G)^{-1} - (G'\Omega^{-1}G)^{-1} + (G'\Omega^{-1}G)^{-1} \\ &= A\Omega A' - (G'\Omega^{-1}G)^{-1} \end{aligned}$$

Since we started with a quadratic form that is ≥ 0, we have established that:

$$(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1} \geq (G'\Omega^{-1}G)^{-1}$$

The equality holds when $A = (G'\Omega^{-1}G)^{-1}G'\Omega^{-1}$, which is satisfied when $W = \Omega^{-1}$. □

The problem is of course that the optimal weight Ω is a function of θ_0, and is therefore infeasible. This is why we introduce what is called the Two-step GMM.

Definition 5.2: Two Step GMM

The two step GMM is computed as follows.

1. Compute a 1st step GMME $\hat{\theta}_W$ using some W_n (e.g. $W_n = I_n$):
$$\hat{\theta}_W = \arg\min_{\theta\in\Theta} g_n(\theta)'W_n g_n(\theta)$$
2. Estimate $\hat{\Omega}$ as follows
$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(z_i, \hat{\theta}_W)g(z_i, \hat{\theta}_W)'$$
3. Compute the 2-step GMME using the estimated optimal weight matrix $\hat{\Omega}^{-1}$
$$\hat{\theta}_{2step} = \arg\min_{\theta\in\Theta} g_n(\theta)'\hat{\Omega}^{-1}g_n(\theta)$$

Note that $\hat{\theta}_{2step}$ is more efficient than $\hat{\theta}_W$. Additionally, steps 2-3 can be repeated (which is called repeated or iterated GMM). While the 2-step GMM estimator is asymptotically efficient, the estimation of Ω depends on the initial (sub-optimal) $\hat{\theta}_W$. This suggests that repeating steps 2 and 3 iteratively—using the latest $\hat{\theta}$ to update $\hat{\Omega}$—can improve finite sample performance.

The procedure is as follows. We continue updating $\hat{\Omega}(\hat{\theta}_{i-1})$ and re-estimating $\hat{\theta}_i$ until the difference $\|\hat{\theta}_i - \hat{\theta}_{i-1}\|$ falls below a small tolerance ϵ_θ (e.g., 10^{-6}). Note that all iterations beyond the second step share the same asymptotic distribution as the 2-step estimator. The primary gain is higher efficiency in finite samples.

5.5 Testing in GMM

5.5.1 Sargan-Hansen Test

One of the main tests in this context has to do with overidentified moment validity. The null/alternative hypothesess can be written as

$$\mathcal{H}_0 : \mathbb{E}(g(z_i, \theta)) = 0 \text{ for some } \theta \in \Theta \text{ v.s. } \mathcal{H}_1 : \mathbb{E}(g(z_i, \theta)) \neq 0 \text{ for all } \theta \in \Theta$$

Rejecting the null hypothesis here means that at least one moment condition does not hold in the data.

Definition 5.3: J-statistic

The J-statistic is written as follows

$$J_n = n \min_{\theta \in \Theta} g_n(\theta)' \hat{\Omega}^{-1} g_n(\theta) = n g_n(\hat{\theta})' \hat{\Omega}^{-1} g_n \hat{\theta})$$

The intuition is that if the model is correct, $g_n(\hat{\theta})$ and J_n have to be close to zero. We basically need $g_n(\hat{\theta})' g_n(\hat{\theta})$ to go to zero faster than n goes to infinity. That is, we shoud have $g_n(\hat{\theta})' g_n(\hat{\theta}) = o(\frac{1}{n^\delta})$ for $\delta > 1$. This is a sufficient condition if we assume that $\hat{\Omega}^{-1} \xrightarrow{p} \Omega^{-1}$, meaning that $\hat{\Omega}^{-1} = O(1)$.

In other words, we know that $\sqrt{n} g_n(\hat{\theta})$ converges to a normal distribution. Since $\sqrt{n} g_n(\hat{\theta}) = O_p(1)$, it follows that:

$$g_n(\hat{\theta}) = O_p\left(\frac{1}{\sqrt{n}}\right)$$

Therefore, the quadratic form $g_n(\hat{\theta})' \hat{\Omega}^{-1} g_n(\hat{\theta})$ is $O_p(\frac{1}{n})$. When you multiply this by the n out front in the J-statistic formula, you get an $O_p(1)$ object, which allows it to converge to a χ^2 distribution rather than collapsing to zero or diverging to infinity. Therefore, under $\mathcal{H}_0$, we have that

$$J_n \xrightarrow{d} \chi^2_{(J-k)}$$

Note that J-test (or Sargan-Hansen test) is only applicable in overidentified models ($J > k$). A rejection suggests either that the moment conditions are invalid (instruments are not

exogenous) or that the functional form of the model is misspecified. It requires the use of the efficient (optimal) weighting matrix $\hat{\Omega}^{-1}$; if a sub-optimal W is used, the statistic does not follow a χ^2 distribution.

5.5.2 Hausman Test

The test compares two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, that behave differently under $\mathcal{H}_0$ and $\mathcal{H}_1$. $\hat{\theta}_1$ is consistent and efficient under $\mathcal{H}_0$, but inconsistent under $\mathcal{H}_1$. On the other hand, $\hat{\theta}_2$ is consistent under both $\mathcal{H}_0$ and $\mathcal{H}_1$, but is inefficient under $\mathcal{H}_0$. The idea is that if $\mathcal{H}_0$ is true, both estimators should converge to the same value, so the difference $(\hat{\theta}_2 - \hat{\theta}_1)$ should be close to zero. If it is far from zero, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$.

Example 5.1: Hausman Exogeneity Test: OLS vs. 2SLS

Consider a linear regression model $y_i = x_i'\beta_0 + \varepsilon_i$, where we suspect some regressors in x_i may be endogenous. We test the following hypotheses:

- $\mathcal{H}_0 : E[x_i\varepsilon_i] = 0$ (Exogeneity)
- $\mathcal{H}_1 : E[x_i\varepsilon_i] \neq 0$ (Endogeneity)

We compare two estimators with different properties:

- $\hat{\beta}_{OLS}$: Consistent and efficient under $\mathcal{H}_0$, but inconsistent under $\mathcal{H}_1$.
- $\hat{\beta}_{2SLS}$: Consistent under both $\mathcal{H}_0$ and $\mathcal{H}_1$, but inefficient under $\mathcal{H}_0$.

Let $\hat{q} = \hat{\beta}_{2SLS} - \hat{\beta}_{OLS}$. Under $\mathcal{H}_0$, $\hat{q} \xrightarrow{p} 0$ because both estimators are consistent. Under $\mathcal{H}_1$, $\hat{q} \xrightarrow{p} \text{plim}\, \hat{\beta}_{2SLS} - \text{plim}\, \hat{\beta}_{OLS} \neq 0$.

The Hausman test statistic H is defined as:

$$H = n(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})'\hat{V}_q^{-1}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}) \xrightarrow{d} \chi^2(k)$$

where k is the dimension of β_0. Crucially, because $\hat{\beta}_{OLS}$ is efficient under $\mathcal{H}_0$, Hausman showed that the variance of the difference simplifies to:

$$\hat{V}_q = \hat{V}_{2SLS} - \hat{V}_{OLS}$$

We reject exogeneity if $H > \chi^2_{1-\alpha}(k)$.

In practice, it is often difficult to know with certainty which instruments are truly valid. If both estimators use invalid instruments, the test loses its ability to distinguish between the models.

Additionally, the test relies on the difference in asymptotic variances $[\widehat{\text{Var}}(\hat{\beta}_{2SLS}) - \widehat{\text{Var}}(\hat{\beta}_{OLS})]$. If the additional instruments in the second-stage estimator do not actually improve the

asymptotic variance (i.e., they add no new information), this matrix difference may not be invertible, causing the test to fail.

Finally, if the "untrusted" instrument is invalid but happens to be uncorrelated with both the "trusted" instruments and the endogenous variables, $\hat{\beta}_1$ (the efficient estimator) might remain consistent even under $\mathcal{H}_1$. In this specific case, the test would fail to reject the null, despite the instrument being technically invalid.

6 Treatment Effects

6.1 Setup

We start by defining the notation we will use throughout this section. First, D denotes participation or treatment, where $D = 1$ if the individual/unit is treated and $D = 0$ otherwise. Now, we can define $Y(0)$, the potential outcome in absence of treatment, and $Y(1)$ the potential outcomes if treated. Note that these potential outcomes are for the same unit/individual under counterfactual cases. We observe D and

$$Y = Y(1)\mathbb{1}\{D = 1\} + Y(0)\mathbb{1}\{D = 0\} = DY(1) + (1 - D)Y(0)$$

The key insight is that we observe either $Y(1)$ or $Y(0)$, but not both. As econometricians, we are interest in the effect of treatment, that is $Y(1) - Y(0)$.

6.2 Objects of interest

6.2.1 Building the objects

A popular object of interest is the Average Treatment Effect (ATE):

$$\theta_{ATE} = ATE = \mathbb{E}(Y(1) - Y(0))$$

or it's conditional version

$$\theta_{ATE_x} = ATE_x = \mathbb{E}(Y(1) - Y(0)|X = x)$$

Similarly, we can define the Average Treatment Effect on the Treated (ATT):

$$\theta_{ATT} = ATT = \mathbb{E}(Y(1) - Y(0)|D = 1)$$

which is conditional on treatment. It answers the question: how does a program change the outcome for treated units compared to what they would have experienced if they had not participated?

We can finally define the Local Average Treatment Effect (LATE):

$$\theta_{LATE} = LATE = \mathbb{E}(Y(1) - Y(0)|D(z) \neq D(z'))$$

which measures the effect of treatment on persons who change state in response to a change in Z (they are at the margin of being treated). $D(z)$ is the conditional random variable D given $Z = z$. When the instruments are indicator variables denoting different policy regimes, LATE is interpreted as the response to policy changes for those who change participation status in response to the change.

6.2.2 Understanding the objects

First, note that if there was an homogenous response, that is if all individuals were affected by a policy intervention in the same way, $\Delta \equiv Y(1) - Y(0)$, all measures would be identical

$$\theta_{ATE} = \theta_{ATT} = \theta_{LATE}$$

In general, this is not the case. We can first think about a case where the is a homogenous response to a policy conditional on X, $\Delta(X) \equiv Y(1) - Y(0)$. That is, within a group (created based on observables), people respond in the same way. Formally, this means

$$\theta_{ATE_x} = \theta_{ATT_x} = \theta_{LATE_x}$$

To reach identification, we need to add a support requirement (sometimes called overlap assumption)

$$0 < \mathbb{P}(D = 1|X = x) < 1, \forall x$$

This means that within a group, we always have non-treated and treated observations. Now, understand that having similar conditional response does not necessarily generalize to the unconditional case. Indeed, ATE averages ATE_x over the whole population while ATT averages ATT_x over the treated subpopulation only.

Now, in general, even conditional on X, there is unobservable heterogeneity. In this case, we would write

$$\Delta + U(1) - U(0) \equiv Y(1) - Y(0)$$

$U(1) - U(0)$ represents the idiosyncratic gain from treatment (heterogeneity that the econometrician cannot see). But notice that generally,

$$\mathbb{E}(U(1) - U(0)|D = 1, X = x) \neq \mathbb{E}(U(1) - U(0)|D = 0, X = x)$$

which implies that individuals choose to participate in a program based on their own unobserved potential gains. This is the first time we encounter selection bias: people who expect to benefit more from the treatment ($U(1) - U(0)$ is high) are more likely to select into treatment ($D = 1$). This is known as the "Roy Model" type selection. In this case, the treated group is not a random sample of the population, but a group that likely has higher-than-average returns to treatment. Consequently, $ATT_x \neq ATE_x$.

6.3 The counterfactual and Selection Bias

6.3.1 The Problem

A key concept in the study of treatment effects is the counterfactual. Consider the conditional ATT_x

$$\theta_{ATT_x} = \mathbb{E}(Y(1)|D=1, X=x) - \mathbb{E}(Y(0)|D=1, X=x)$$

It is obvious that we do not observe $\mathbb{E}(Y(0)|D=1, X=x)$: this represents what would have happened to someone that was treated if they were not treated. Now, an idea would be to use $\mathbb{E}(Y(0)|D=1, X=x)$ as a proxy for it. ATT_x would therefore be rewritten as

$$\theta_{ATT_x} = \mathbb{E}(Y(1)|D=1, X=x) - \mathbb{E}(Y(0)|D=0, X=x)$$

$$= \mathbb{E}(Y(1)|D=1, X=x) - \mathbb{E}(Y(0)|D=1, X=x) + \mathbb{E}(Y(0)|D=1, X=x) - \mathbb{E}(Y(0)|D=0, X=x)$$

Therefore, we have the following

$$\underbrace{\mathbb{E}(Y|D=1, X) - \mathbb{E}(Y|D=0, X)}_{\text{Observed Difference}} = \underbrace{\mathbb{E}(Y(1) - Y(0)|D=1, X)}_{\text{ATT}_x} + \underbrace{\mathbb{E}(Y(0)|D=1, X) - \mathbb{E}(Y(0)|D=0, X)}_{\text{Selection Bias}}$$

ATT_x is the "true effect" for the people who actually took the treatment. The selection bias term represents the difference in the starting point. It asks: "Even without the treatment, would the treated group have performed differently than the untreated group?". If the treated group is more motivated/healthier/wealthier by nature, then $\mathbb{E}(Y(0)|D=1) > \mathbb{E}(Y(0)|D=0)$, and we will overestimate the treatment effect.

6.3.2 Some solutions

First, if we are using experimental data, randomized experiments can solve the problem of selection bias. The idea is randomization solves this by making the treatment D independent of potential outcomes $(Y(0), Y(1))$. This ensures that $\mathbb{E}[Y(0)|D=1] = \mathbb{E}[Y(0)|D=0]$, effectively killing the selection bias and leaving you with just the treatment effect. Formally, we need the treatment D to be assigned independently of the subjects' characteristics. Mathematically, this is expressed as:

$$(R) \quad (Y(0), Y(1)) \perp D$$

We need to consider cases with non-experimental data too. In the absence of a truly randomized experiment (R), we often rely on the CI Assumption (also known as "Selection on Observables"). We assume that the treatment assignment is "as good as random" once we control for a set of observed covariates X. Formally,

$$(CI) \quad (Y(0), Y(1)) \perp D \mid X$$

Under CI, the selection bias term we derived earlier vanishes for individuals with the same X. This implies: $E[Y(0)|D=1, X] = E[Y(0)|D=0, X] = E[Y(0)|X]$. Consequently, we can identify ATE_x and ATT_x simply by comparing treated and untreated units within the same X groups.

6.4 Estimation

6.4.1 Outcome specification

In many economic contexts, (CI) is too strong. Even if we control for X (like education or age), there is often unobserved heterogeneity ($U_1 - U_0$) that affects both the decision to participate (D) and the potential outcomes. To understand this selection on unobservables problem, consider the following structural model

$$Y(0) = g_0(X) + U_0 \quad \text{and} \quad Y(1) = g_1(X) + U_1$$

where $g(X)$ represents the mean outcome based on observables and U represents the unobservable components. Notice that this implies that $\mathbb{E}[U_0|X] = \mathbb{E}[U_1|X] = 0$. Now, remember that

$$Y = (1 - D)Y(0) + DY(1)^{23}$$

By defining $\Delta(X) = g_1(X) - g_0(X)$ as the mean gain, we can write the observed outcome Y as:

$$Y = g_0(X) + D \cdot \Delta(X) + [U_0 + D(U_1 - U_0)]$$

Using this notation, we can clearly see why these effects differ when there is unobserved heterogeneity:

$$ATE(x) = \mathbb{E}[Y(1) - Y(0)|X = x] = g_1(x) - g_0(x) \text{ and } ATT(x) = ATE(x) + \underbrace{\mathbb{E}[U_1 - U_0|D = 1, X = x]}_{\text{Unobserved Gain}}$$

These two are equal only in cases where there are no unobservable components of the gain ($U_1 = U_0$), or if $U_1 - U_0$ does not determine who goes into the program (treatment uncorrelated with error).

Now, we can rewrite the observed outcome equation to highlight the specific estimation challenges:

$$Y = g_0(X) + D \cdot ATE(x) + \varepsilon$$

where $\varepsilon = U_0 + D(U_1 - U_0)$. To identify $ATE(x)$, we need the treatment D to be uncorrelated with the entire error term ε. However, endogeneity arises from two distinct sources:

- Correlation with U_0. This is the standard selection bias (levels). People with higher baseline outcomes might be more/less likely to participate.
- Correlation with $(U_1 - U_0)$. This is selection on gains. People who expect higher-than-average idiosyncratic benefits are more likely to participate.

Even if we only want to estimate the effect on those who actually participated (ATT), we still face an endogeneity problem. Recall

$$ATT(x) = \underbrace{g_1(X) - g_0(X)}_{ATE(x)} + \mathbb{E}(U_1 - U_0|D = 1, X = x)$$

[23] Recall it is an easy way to represent $Y = Y(1)\mathbb{1}\{D = 1\} + Y(0)\mathbb{1}\{D = 0\}$

$$\implies g_1(X) - g_0(X) = ATT(x) - \mathbb{E}(U_1 - U_0|D = 1, X = x)$$

Substituting this in our main expression for Y yields

$$Y = g_0(X) + D\left[ATT(x) - \mathbb{E}(U_1 - U_0|D = 1, X = x)\right] + U_0 + D(U_1 - U_0)$$

By rearranging the terms, we can write:

$$Y = g_0(X) + D \cdot ATT(x) + \{U_0 + D[U_1 - U_0 - \mathbb{E}(U_1 - U_0|D = 1)]\}$$

The problem remains: D is correlated with U_0. Even if there is no selection on gains, the simple difference in means is still biased if there is a selection on levels (U_0).

6.4.2 Participation Decision

In this framework, the choice to participate ($D = 0$ versus $D = 1$) is driven by a latent variable IN, which represents the net profit or net utility of treatment. An individual decides to participate if the expected utility is non-negative:

$$D = \mathbb{1}\{IN \geq 0\}$$

The index IN is typically modeled as a function of both observed and unobserved factors:

$$IN = g(Z, V)$$

Z is a vector of observable characteristics. Note that Z usually includes X (the covariates from the outcome equation) plus at least one instrument that affects the decision to participate but does not directly affect the potential outcomes. V is an unobserved random component representing idiosyncratic shocks to the participation decision. A common way to write this index is linearly:

$$IN = Z\gamma + V$$

Under this specification, the probability of participation (the propensity score) is:

$$\mathbb{P}(Z) = \mathbb{P}(D = 1|Z) = \mathbb{P}(V \geq -Z\gamma)$$

As before, the selection problem occurs if the unobservables in the participation decision (V) are correlated with the unobservables in the outcome equations (U_0, U_1).

6.5 Instrumental Variables

The core idea of IV is to find a variable Z that acts as a "shifter." It must be powerful enough to change the probability that someone takes the treatment (D), but "clean" enough that it has no independent effect on the outcome (Y). As always, the first condition is relevance: $\mathbb{P}(D = 1|X, Z)$ must actually depend on Z. This ensures there is "independent variation" in D generated by Z. The second condition is exogeneity or validity: Z must be uncorrelated with the unobserved potential outcomes. The dependence of Y on Z must operate only through D.

6.5.1 The structural setup for IV

Recall the outcome equation

$$Y = g_0(X) + D \cdot ATE(x) + [U_0 + D(U_1 - U_0)]$$

Additionally, we have the participation decision written as

$$D = \mathbb{1}\{g(Z, V) \geq 0\}$$

In this setup, we assume that Z is a variable that satisfies mean independence:

$$\mathbb{E}(U_0 + D(U_1 - U_0)|X, Z) = 0$$

This identifying condition is what allows us to isolate the effect of D on Y by using only the part of D that was triggered by the "random-like" movement of the instrument Z.

When we want to estimate an ATT, the outcome equation is

$$Y = g_0(X) + D \cdot ATT(x) + \{U_0 + D[U_1 - U_0 - \mathbb{E}(U_1 - U_0|D = 1)]\}$$

Therefore, the identifying assumption is

$$\mathbb{E}\{U_0 + D[U_1 - U_0 - \mathbb{E}(U_1 - U_0 \mid X, D = 1)] \mid X, Z\} = 0$$

This is a weaker identifying condition for the ATT. It acknowledges that while individuals might select into treatment based on gains (violating the ATE assumption), the instrument Z itself remains "clean" of the baseline levels and the specific variation in gains.

Regardless of whether you want ATE or ATT, the instrument must satisfy the relevance condition: $\mathbb{E}(D \mid X, Z) = \mathbb{P}(D = 1 \mid X, Z)$ must depend on Z. This means there must be independent variation in D generated by Z. If Z doesn't change the probability of participation, it cannot help us identify the effect.

Because we have assumed the instrument is independent of the error terms (U_0, U_1), when you take the expectation $\mathbb{E}(\varepsilon|X, Z)$, those terms drop out (become zero), leaving only the part of the model that depends on observables. For the ATE,

$$\mathbb{E}(Y|X, Z) = g_0(X) + \underbrace{\mathbb{E}(D|X, Z)}_{\mathbb{P}(D=1|X,Z)} \cdot ATE(x) + \underbrace{\mathbb{E}(U_0 + D(U_1 - U_0)|X, Z)}_{=0}$$

$$\implies \mathbb{E}(Y|X, Z) = g_0(X) + \mathbb{P}(D = 1 \mid X, Z) \cdot ATE(x)$$

For the ATT,

$$\mathbb{E}(Y|X, Z) = g_0(X) + \underbrace{\mathbb{E}(D|X, Z)}_{\mathbb{P}(D=1|X,Z)} \cdot ATT(x) + \underbrace{\mathbb{E}(\varepsilon_{ATT}|X, Z)}_{=0}$$

$$\implies \mathbb{E}(Y|X, Z) = g_0(X) + \mathbb{P}(D = 1 \mid X, Z) \cdot ATT(x)$$

6.5.2 IV estimation

Now, suppose we have a binary instrument $Z \in \{0,1\}$, you can now derive what is called the Wald Estimator by comparing the outcomes at $Z = 1$ and $Z = 0$. Subtracting the two reduced-form equations (for $Z = 1$ and $Z = 0$) cancels out the baseline $g_0(X)$:

$$\mathbb{E}(Y|X, Z = 1) - \mathbb{E}(Y|X, Z = 0) = [\mathbb{P}(D = 1|X, Z = 1) - \mathbb{P}(D = 1|X, Z = 0)] \cdot \theta$$

Solving for θ gives you the famous Wald formula:

$$\theta_{IV} = \frac{\mathbb{E}(Y|X, Z = 1) - \mathbb{E}(Y|X, Z = 0)}{\mathbb{P}(D = 1|X, Z = 1) - \mathbb{P}(D = 1|X, Z = 0)}$$

Notice that this relies on the assumption that $\mathbb{P}(D = 1|X, Z = 1) \neq \mathbb{P}(D = 1|X, Z = 0)$, which is once again the relevance condition: the instrument switching from 0 to 1 should have an impact on the probability of treatment.

This ratio reveals the core logic of instrumental variables. The numerator, $\mathbb{E}(Y|X, Z = 1) - \mathbb{E}(Y|X, Z = 0)$, is known as the Reduced Form; it captures the total effect of the instrument on the outcome of interest. However, since the instrument only affects the outcome indirectly through the treatment D, this raw difference understates the true treatment effect because not everyone in the $Z = 1$ group actually receives the treatment. To correct for this, we divide by the first stage, $\mathbb{P}(D = 1|X, Z = 1) - \mathbb{P}(D = 1|X, Z = 0)$, which measures how much the instrument actually impacts treatment participation. By dividing the reduced form by the first stage, we "blow up" the instrument's effect to recover the full impact of the treatment itself. Notice that this is exactly the same idea as what we found in Section 4: $\hat{\beta}_{\text{IV}} = \hat{\Pi}^{-1}\hat{\lambda}$.

6.5.3 LATE

We now introduce more formally the assumption of Local Average Treatment Effect (LATE). Imagine a binary instrument (e.g., being randomly assigned a voucher for a private school). For any individual, we can define their treatment status as $D(z)$. This gives us four types of people.

- Never-takers: People who never take the treatment regardless of the instrument ($D(1) = 0, D(0) = 0$)
- Always-takers: People who always take the treatment regardless of the instrument ($D(1) = 1, D(0) = 1$)
- Compliers: People who take the treatment if they get the instrument, and don't if they don't ($D(1) = 1, D(0) = 0$). These are the only people whose behavior is actually changed by the instrument
- Defiers: People who do the opposite of what the instrument suggests ($D(1) = 0, D(0) = 1$).

To identify LATE, we must assume Monotonicity (also known as the "No Defiers" assumption). It states that while the instrument might have no effect on some people (Always-takers/Never-takers), it must not push people in opposite directions. The idea is that we compare the same person in two counterfactual cases. Monotonicity implies that a person who does not go to private school receiving a voucher would not go to private school without the voucher either. Formally, we write $D_i(1) \geq D_i(0)$ for all i.

Under Monotonicity and the IV assumptions (Relevance and Validity), the Wald estimator identifies the Local Average Treatment Effect:

$$\theta_{LATE} = E[Y(1) - Y(0) \mid D(1) > D(0)]$$

This is the average effect of treatment only for the compliers. In our school example, this corresponds to the average effect on test score for people who choose private school because they have received a voucher.

While LATE provides high internal validity (it solves the selection problem for the marginal group), it has limitations regarding **external validity**. It does not provide information about:

- **Always-takers:** Those who participate regardless of the instrument's value.
- **Never-takers:** Those who refuse participation regardless of the instrument's value.

Therefore, a LATE estimate from a voucher program might not accurately predict the effect of a universal private schooling mandate (the ATE).

6.6 Some other topics

6.6.1 The Two-Step Heckman Estimator

I don't think it makes much sense introducing this without building everything from the usual limited dependent variables section. That is, it would be nice to start from probit/logit, moving on to some Tobit models and arrive at this point. I will maybe add this to the notes at some point. Still, let us discuss this quickly.

The Heckman estimator is designed to correct for **Sample Selection Bias**, which occurs when the observed sample is not a random draw from the population. This is viewed as a form of omitted variable bias where the omitted variable is the "selectivity" of the individuals.

The estimation proceeds in two stages:

1. **Selection Stage:** Estimate a Probit model of the participation decision $D_i = \mathbb{1}\{Z_i\gamma + v_i > 0\}$. From this, calculate the **Inverse Mills Ratio** (λ_i), which represents the probability density over the cumulative distribution of the selection shock.

2. **Outcome Stage:** Estimate the outcome equation (e.g., wages) via OLS, including the Inverse Mills Ratio as an additional covariate:

$$Y_i = X_i\beta + \sigma_{uv}\lambda_i + \epsilon_i$$

Including λ_i effectively controls for the unobserved factors that lead to selection, allowing for consistent estimation of β even when the sample is non-random.

6.6.2 Matching

Matching is a quasi-experimental design that re-establishes experimental conditions by pairing treated units with similar non-treated units. It does not require a particular specification of the participation decision or the outcome equations. The identification assumptions can be written as follows:

- **Conditional Independence (M1):** Selection is based entirely on observables X. Formally, $\mathbb{E}[Y_0 \mid D = 1, X] = \mathbb{E}[Y_0 \mid D = 0, X]$.
- **Common Support (M2):** For all X, $0 < \mathbb{P}(D = 1 \mid X) < 1$. This ensures every treated agent has a potential counterpart in the non-treated population.

The Estimator: The treatment effect for a treated individual is estimated by:

$$\hat{\Delta}_i = Y_{1i} - \sum_{j \in C} W_{ij} Y_{0j}$$

where W_{ij} are weights assigned to control units based on their similarity to treated unit i.

OLS can be interpreted as a form of matching. Under the classic conditional mean independence assumption ($E[u|T, X] = E[u|X]$), running OLS on:

$$Y = \alpha + \beta T + X'\gamma + v$$

provides consistent estimates for β, though these coefficients may not be strictly causal if unobservables are present.

Applying the **Frisch-Waugh-Lovell theorem**, we can think of OLS as orthogonalizing the treatment T with respect to X, effectively creating many "cells" of individuals with similar X and comparing treated vs. untreated units within those cells. Propensity score matching is often preferred over OLS because it does not require a linear specification and explicitly handles the challenge of common support. It reduces the dimensionality of X by matching units based on their predicted probability of treatment, $\mathbb{P}(X) = \mathbb{P}(D = 1 \mid X)$.

6.6.3 Difference-in-differences

I provide a super short intro to DiD here. The classical before-after estimator compares the outcomes of participants after they participate in the program with their outcomes before they participate. Consider the simple model where we have a panel and data on participatino before and after they participate in period k

$$y_{it} = x'_{it}\beta + d_i\Delta + u_{0it}$$

where $d_i = \mathbb{1}\{t > k\}$ and $u_{0it} = \alpha_i + \varepsilon_{it}$, where i is some individual heterogeneity (which can include heterogenous gain from treatment).

Suppose that selection is only allowed on the permanent error component α_i. This means that ε_{it} is a random error term, $\mathbb{E}(\varepsilon_{it}d_i) = 0$ and $\varepsilon_{it} \perp\!\!\!\perp \varepsilon_{is}$ for all $s \neq t$ but we allow $\mathbb{E}(\alpha_i d_i) \neq 0$. As we will see in the panel data section below, a consistent estimator of the average treatment effect is the fixed effect estimator

$$y_{it'} - y_{i\tau} = (x_{it'} - x_{i\tau})'\beta + d_i\Delta + (u_{it'} - u_{i\tau}) \text{ for } t' > k > \tau$$

In practice, we need a setting with absence of economy-wide factors affecting participants outcomes. The DiD estimator eliminates both common macro effects and individual specific fixed effects by substracting the before-after change for participant outcomes. The critical identifying assumption is that conditional on X, the biases are the same on average in different time periods before and after the period of participation in the program so that differencing the differences between participants and nonparticipants eliminates the bias.

Formally, the main assumption is the parallel trends assumption. This is the most critical identification condition. It requires that, in the absence of treatment, the average change in potential outcomes for the treated group would have been the same as the observed average change for the control group:

$$E(u_{0it'} - u_{0i\tau}|D_i = 1, X_i) = E(u_{0it'} - u_{0i\tau}|D_i = 0, X_i)$$

for time periods before ($\tau < k$) and after ($t' > k$) the treatment starts. Additionally, selection into treatment must be independent of the temporary shocks ε_{it}. Formally: $E(\varepsilon_{it}|D_i, X_i) = 0$. To ensure the biases cancel out, the time periods compared (t' and τ) should ideally be equally far from the treatment period k to satisfy the stationarity assumption (stationarity).

The DiD estimator effectively "washes out" α_i by differencing over time within units, and it "washes out" θ_t by differencing across the treated and control groups.

7 Panel Data

A panel data (pooled cross section and time series) regression differs from a regular time-series or cross-section regression in that it has a double subscript on all variables $\{y_{it}, x_{it}\}$

for $i = 1, ..., n$ and $t = 1, ..., T$. i denotes the cross-sectional dimension (individuals, households, firms, etc.) and t the time-series dimension (days, weeks, etc.). Different types of variables can be included in a panel. We can have variables that vary across individuals and time, such as wage, age, years of experience. Some variables are *time invariant* such as race. Finally, some variables vary only over time, but not across individuals, in general aggregate/macro factors (e.g. unemployment).

Another important bit of terminology is what we call complete or balanced panels, where individuals are observed over the entire sample periods. Unbalanced panels occur in case where some $\{it\}$ is missing. We focus on balanced panels. Typically, we will be concerned with situations where T is small and n is large: fixed T and $n \to \infty$.

7.1 Fixed Effects Models

Consider the following linear model

$$y_{it} = \alpha + x'_{it}\beta + u_{it}$$

with x_{it} a k dimensional vector of explanatory variables, not including a constant. For any t, we assume $x_{it} \sim i.i.d$ across individuals. An error component model (ECM) specifies the structure of the error. Let μ_i be an unobservable individual-specific effect, λ_t an unobservable time effect and v_{it} and idiosyncratic component, where we assume $v_{it} \sim iid(0, \sigma^2)$ on i and t. We call one-way ECM a model of the form

$$u_{it} = \mu_i + v_{it} \quad \text{or} \quad u_{it} = \lambda_t + v_{it}$$

We call a two-way ECM a model of the form

$$u_{it} = \mu_i + \lambda_t + v_{it}$$

If characteristics that have a direct effect on both dependent variable and explanatory variables are omitted, then explanatory variables are correlated with the regression error and, consequently, estimators of the coefficients will be inconsistent. A traditional response of econometrics to this problem has been instrumental variables models. However, it is often very difficult to find instruments. A major motivation for using panel data has been the ability to control for (possibly correlated) time-invariant heterogeneity without observing it.

7.1.1 General case

In fixed effects models, we treat μ_i as n fixed unknown parameters. Consider our one way ECM :

$$y_{it} = \alpha + \mu_i + x'_{it}\beta + v_{it}$$

The main assumption in this context is **strict exogeneity**

$$\mathbb{E}(v_{it}|x_{i1}, ..., x_{iT}) = 0, \forall t$$

We can rewrite this model as

$$y_{it} = \alpha + \sum_{j=1}^{n} \mu_j d_{ij,t} + x_{it}'\beta + v_{it}$$

where $d_{ij,t} = \mathbb{1}\{i = j\}$. Note that if we have perfect collinearity with α (if we sum through all $d_{ij,t}$, we get 1). Therefore, we need to drop the constant and write

$$y_{it} = \sum_{j=1}^{n} \mu_j d_{ij,t} + x_{it}'\beta + v_{it}$$

To estimate the model efficiently, we stack the observations, dropping the subscript t. For a single individual i, the model is:

$$y_i = \iota_T \mu_i + X_i\beta + v_i,$$

where the components are:

$$\underbrace{y_i}_{(T\times 1)} = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix}, \quad \underbrace{\iota_T}_{(T\times 1)} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \underbrace{X_i}_{(T\times k)} = \begin{pmatrix} x_{i1}' \\ x_{i2}' \\ \vdots \\ x_{iT}' \end{pmatrix}, \quad \underbrace{v_i}_{(T\times 1)} = \begin{pmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{iT} \end{pmatrix}$$

Stacking all n individuals, we define the matrix equation:

$$y = D\mu + X\beta + v$$

With the following block definitions:

$$\underbrace{y}_{(nT\times 1)} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \underbrace{\mu}_{(n\times 1)} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}, \quad \underbrace{v}_{(nT\times 1)} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \quad \underbrace{X}_{(nT\times k)} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

The matrix D is the matrix of individual dummies, defined using the Kronecker product:

$$\underbrace{D}_{(nT\times n)} = I_n \otimes \iota_T = \begin{pmatrix} \iota_T & 0 & 0 & \dots & 0 \\ 0 & \iota_T & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \iota_T \end{pmatrix}$$

Matrix D is a "selector." When you multiply $D\mu$, the first column of D picks out μ_1 and applies it to of person 1's observations (T of them), the second column picks μ_2 for person 2, and so on.

Example 7.1: Visualizing Panel Matrices ($n = 2, T = 2$)

Suppose we have two individuals, Alice ($i = 1$) and Bob ($i = 2$), observed for two years. Our model is $y_{it} = \mu_i + \beta x_{it} + v_{it}$.

For Alice ($i = 1$):

$$y_1 = \begin{pmatrix} y_{11} \\ y_{12} \end{pmatrix}, \quad X_1 = \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}, \quad \iota_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

For Bob ($i = 2$):

$$y_2 = \begin{pmatrix} y_{21} \\ y_{22} \end{pmatrix}, \quad X_2 = \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix}, \quad \iota_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The stacked outcome vector y and regressor matrix X (size $nT \times 1 = 4 \times 1$) are:

$$y = \left(\begin{array}{c} y_{11} \\ y_{12} \\ \hline y_{21} \\ y_{22} \end{array}\right), \quad X = \left(\begin{array}{c} x_{11} \\ x_{12} \\ \hline x_{21} \\ x_{22} \end{array}\right)$$

$D = I_2 \otimes \iota_2$ creates the columns that assign μ_1 to Alice and μ_2 to Bob:

$$D = \left(\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ \hline 0 & 1 \\ 0 & 1 \end{array}\right), \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

When we calculate $D\mu$, we get a vector where the first two rows are μ_1 and the last two are μ_2. This allows each individual to have their own "fixed" intercept over time.

We estimate the stacked model $y = D\mu + X\beta + v$ via OLS. Under the assumption that $E[vv'] = \sigma_v^2 I_{nT}$, OLS is efficient. However, the following conditions must hold for identification:

- **Time Dimension**: $T \geq 2$. If $T = 1$, μ_i cannot be calculated as it is perfectly confounded with the error term.
- **Variation**: X must not contain time-invariant regressors, as these would be perfectly collinear with the columns of D.

Direct OLS estimation is often unattractive because it requires handling the $(nT \times n)$ matrix $D$and inverting a $(n + k) \times (n + k)$ matrix, which is computationally taxing for large n. We are typically only interested in β, making the estimation of all n values of μ unnecessary.

7.1.2 The Within Estimator (Demeaning)

A numerically efficient way to estimate β is to use the **Frisch-Waugh-Lovell (FWL) Theorem**. In our stacked model $y = D\mu + X\beta + v$, the matrix of individual dummies D plays

the role of the variables we wish to partial out. Indeed, we write the partition as $[D : X]$.

We define the projection matrix M_D as:

$$M_D = I_{nT} - D(D'D)^{-1}D'$$

Recall $D = I_n \otimes \iota_T$. We derive $M_D = I_{nT} - D(D'D)^{-1}D'$ as follows:

1. $D'D = {}^{24}(I_n \otimes \iota_T)'(I_n \otimes \iota_T) = I_n \otimes \iota_T'\iota_T = TI_n$
2. $D(D'D)^{-1}D' = (I_n \otimes \iota_T)(\frac{1}{T}I_n)(I_n \otimes \iota_T') = I_n \otimes \frac{1}{T}J_T$ where $J_T = \iota_T\iota_T'$
3. $M_D = I_{nT} - I_n \otimes \frac{1}{T}J_T$

Now, look at the projection part $P_D = D(D'D)^{-1}D'$, which we have shown is equal to $I_n \otimes \frac{1}{T}J_T$, where J_T is a $T \times T$ matrix of all ones. When you multiply $\frac{1}{T}J_T$ by an individual's vector y_i, you get:

$$\frac{1}{T}\begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}\begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix} = \begin{pmatrix} \bar{y}_i \\ \vdots \\ \bar{y}_i \end{pmatrix}$$

This operation simply calculates the mean for that person and repeats it T times.

Now, apply the full $M_D = I_{nT} - P_D$ matrix to the vector y:

$$M_D y = (I_{nT}y) - (P_D y)$$

$$M_D y = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{nT} \end{pmatrix} - \begin{pmatrix} \bar{y}_1 \\ \bar{y}_1 \\ \vdots \\ \bar{y}_n \end{pmatrix} = \begin{pmatrix} y_{11} - \bar{y}_1 \\ y_{12} - \bar{y}_1 \\ \vdots \\ y_{nT} - \bar{y}_n \end{pmatrix}$$

Therefore, we get

$$M_D y = y - \bar{y}_i \quad \text{where} \quad \bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

Thus, the FE estimator $\hat{\beta}_{FE} = (X'M_D X)^{-1}X'M_D y$ obtained by FWL is the OLS estimator of the regression of $(y_{it} - \bar{y}_i)$ on $(x_{it} - \bar{x}_i)$. $\hat{\beta}_{FE}$ is the estimator that we obtain if we first project out the group effects μ_i by measuring the individual observations in deviations from group means. This is known as **within variation**: we study how individuals varies compared to its own mean.

Intuition: The fixed effects estimator exploits variation *within* each individual over time, effectively comparing each person to themselves. By subtracting individual-specific means (demeaning), we remove all time-invariant characteristics—both observed and unobserved—that might confound our estimates. This transformation asks: when individual i

[24] Using the properties of Kronecker products $(A \otimes B) \cdot (C \otimes D) = (AC \otimes BD)$

experiences a change in x_{it} relative to their own average, how does y_{it} change relative to their own average? This "within" comparison controls for any fixed individual heterogeneity μ_i, whether it's ability, preferences, or other unchanging traits.

The estimator is unbiased under the strict exogeneity assumption:

$$\mathbb{E}(v_{it}|x_{i1}, \ldots, x_{iT}) = 0 \quad \forall t$$

Substituting the model $y = D\mu + X\beta + v$ into the estimator formula:

$$\mathbb{E}(\hat{\beta}_W|X) = \mathbb{E}\left[(X'M_DX)^{-1}X'M_D(D\mu + X\beta + v)\right]$$

Since $M_DD = 0$, the μ term drops out:

$$\mathbb{E}(\hat{\beta}_W|X) = \beta + (X'M_DX)^{-1}X'M_D\mathbb{E}(v|X)$$

Under strict exogeneity, $\mathbb{E}(v|X) = 0$, so $\mathbb{E}(\hat{\beta}_W|X) = \beta$

The requirement for consistency is $E[(x_{it} - \bar{x}_i)(v_{it} - \bar{v}_i)] = 0$.

- **Small** T: Requires *Strict Exogeneity* (x_{it} uncorrelated with v_{is} for all s, t) because $\bar{v}_i$ contains errors from all periods.
- **Large** T: As $T \to \infty$, $\bar{v}_i$ and $\bar{x}_i$ converge to their expectations. In this case, *Contemporaneous Exogeneity* ($E[x_{it}v_{it}] = 0$) is sufficient for consistency.

This is why Fixed Effects is often considered "safer" in long panels (T is large) than in short panels where feedback loops can violate strict exogeneity.

Assuming the idiosyncratic errors are homoskedastic and serially uncorrelated ($V(v|X) = \sigma_v^2 I_{nT}$), the variance of the estimator is:

$$V(\hat{\beta}_W|X) = \sigma_v^2(X'M_DX)^{-1}$$

σ_v^2 is consistently estimated as

$$\hat{\sigma}_v^2 = \frac{RSS}{nT - n - k}$$

where RSS is the residiual sum of squares from the within-transformed model. Notice that we subtract $n + k$ because this is the number of unknown parameters (n fixed effects, k regressors).

While OLS is used for the within-transformed model, the disturbances $\tilde{v}_{it} = v_{it} - \bar{v}_i$ are correlated by construction. Usually, when disturbances are correlated, we use GLS to achieve efficiency. However, GLS is not an option here because the transformed idiosyncratic errors $\tilde{v}_{it} = v_{it} - \bar{v}_i$ necessarily sum to zero for each individual: $\sum_{t=1}^{T} \tilde{v}_{it} = 0$.

This constraint implies that the T disturbances for each individual are linearly dependent, making the resulting $nT \times nT$ covariance matrix $\mathbb{E}[M_D v(M_D v)'] = \sigma_v^2 M_D$ singular. Consequently, the standard GLS estimator is not defined for the within-transformed model.

To account for this correlation and potential heteroskedasticity, we typically **cluster standard errors** at the individual (or treatment) level. Indeed, OLS standard errors assume that the residuals are i.i.d. However, in the FE model, even if the original v_{it} were i.i.d., the transformed residuals $(v_{it} - \bar{v}_i)$ are correlated because they all share the same $\bar{v}_i$ term. This correlation is strictly internal to the individual (the "cluster"). If we ignore this, our standard errors will be biased (usually downward), leading to over-rejection of the null hypothesis (t-statistics that are artificially high).

Formally,

$$\frac{RSS}{nT-k} \xrightarrow{p} \frac{T-1}{T}\sigma_v^2$$

Since $T - 1 < T$, which is strictly smaller than the true variance. This is why we use

$$\hat{\sigma}_v^2 = \frac{RSS}{nT-n-k} = \frac{RSS}{nT-k} \cdot \frac{nT-k}{n(T-1)-k} \xrightarrow{p} \frac{T-1}{T}\sigma_v^2 \cdot \frac{T}{T-1} = \sigma_v^2$$

Once we have obtained a consistent estimate $\hat{\beta}$ via the within-estimator, we can recover the estimated fixed effects $\hat{\mu}_i$. Using the FWL theorem, the vector of fixed effects is:

$$\hat{\mu} = (D'D)^{-1}D'(y - X\hat{\beta})$$

For each individual i, this simplifies to a very intuitive result:

$$\hat{\mu}_i = \bar{y}_i - \bar{x}_i'\hat{\beta}$$

The fixed effect μ_i is defined as the constant level of y for individual i that cannot be explained by the variation in x_{it}. Therefore, it is simply the difference between their average outcome and the predicted average outcome based on their characteristics.

While $\hat{\beta}$ is consistent as $n \to \infty$, remember that $\hat{\mu}_i$ is **not** consistent if T is fixed. We only have T observations to estimate each specific μ_i, so the estimation error in $\hat{\mu}_i$ does not vanish unless the number of time periods grows ($T \to \infty$).

7.1.3 Two-way ECM with FE

Consider our two way ECM :

$$y_{it} = \alpha + \mu_i + \lambda_t + x_{it}'\beta + v_{it}$$

Here again, we have a multicollinearity problem. Indeed, the sum of μ_i and the sum of λ_t both give 1. In general, we keep α and drop one individual and one time FE, say $\mu_n = 0$ and $\lambda_T = 0$.

The OLS estimation of this model is very similar to the one-way ECM. Omitting the derivations, we get

$$y_{it} - \bar{y}_i - \bar{y}_t + \bar{y} = (x_{it} - \bar{x}_i - \bar{x}_t + \bar{x})'\beta + v_{it} - \bar{v}_i - \bar{v}_t + \bar{v}$$

In practice, this is a version of the within-estimator in which we are also removing period-specific shocks across all individuals.

7.1.4 Drawbacks of the fixed-effect model

One of the main drabacks of this model is that it fails to identify any components of β corresponding to regressors that are constant over time for a given individual (e.g. race, religion, etc.), since they are absorbed in the individual fixed effect. Indeed, they are perfectly collinear with the individual fixed effect D and are wiped out by the M_D transformation.

Coefficients of time-varying regressors are estimable, but these estimates may be very imprecise if most of the variation in a regressor is cross sectional rather than over time. We can diagnose this using the Stata command `xtsum`, which decomposes total variance into:

- **Between variation**: $\bar{x}_i - \bar{x}$, the variation across different individuals.
- **Within variation**: $x_{it} - \bar{x}_i$, the variation over time for the same individual.

If the within variation is small relative to the between variation, the FE estimator will have large standard errors because it relies solely on that thin "within" slice of data.

For example, in a short panel, education level is often time-invariant for most adults. Since M_D removes the cross-sectional differences in education levels, $\hat{\beta}_{educ}$ would be identified only by the few individuals who completed a degree during the sample period. This leads to a loss of power compared to a Random Effects (introduced below) or Pooled OLS model.

Additionally, prediction of the absolute conditional mean is not possible because the individual intercepts μ_i are treated as nuisance parameters and are not consistently estimated for fixed T. We can only predict *changes* in the dependent variable caused by **changes in time-varying regressors**.

7.2 Random Effects Models

Consider again the one way ECM:

$$y_{it} = \alpha + x'_{it}\beta + u_{it}$$

with x_{it} being i.i.d and $u_{it} = \mu_i + v_{it}$. In this case, individual effects μ_i are assumed to be random and are treated as drawings from the same distribution $\mu_i \sim i.i.d(0, \sigma^2_\mu)$. We further assume $v_{it} \sim i.i.d(0, \sigma^2_v)$. Additionally, we assume **strict exogeneity**

$$\mathbb{E}(v_{it}|x_{i1}, ..., x_{iT}) = 0, \forall t$$

and assume $\mu_i \perp\!\!\!\perp v_{it}$, $\mathbb{E}(\mu_i|x_{i1},...,x_{iT}) = 0$. Note that here, the is no dummy variable trap, so we do not need to drop any constant term.

Denote the variance-covariance matrix $\Omega \equiv \mathbb{V}(u)$. The diagonal elements of Ω can be written as

$$\mathbb{V}(u_{it}) = \mathbb{V}(\mu_i + v_{it}) = \sigma_\mu^2 + \sigma_v^2 + 2\mathbb{C}\text{ov}(\mu_i, v_{it}) = \sigma_\mu^2 + \sigma_v^2$$

For the covariance elements, we have

$$\mathbb{C}\text{ov}(u_{it}, u_{is}) = \mathbb{C}\text{ov}(\mu_i + v_{it}, \mu_i + v_{is}) = \mathbb{E}\left((\mu_i + v_{it})(\mu_i + v_{is})\right) = \sigma_\mu^2$$

This is true for any i across time periods within an individual. Now, $\mathbb{C}\text{ov}(u_{it}, u_{is}) = 0$ for any $i \neq j$ since we assumed independence. Therefore, we can write $\Omega = \sigma_v^2 I_{nT} + \sigma_\mu^2 I_n \otimes J_T$

Therefore, the variance-covariance matrix of the composite errors, $\Omega = E[uu']$, is an $nT \times nT$ block-diagonal matrix. Each $T \times T$ block corresponds to an individual i and has a specific structure where the diagonal is $\sigma_\mu^2 + \sigma_v^2$ and all off-diagonals are σ_μ^2.

$$\Omega = \begin{pmatrix} \begin{pmatrix} \sigma_\mu^2 + \sigma_v^2 & \cdots & \sigma_\mu^2 \\ \vdots & \ddots & \vdots \\ \sigma_\mu^2 & \cdots & \sigma_\mu^2 + \sigma_v^2 \end{pmatrix} & 0 & \cdots & 0 \\ 0 & \begin{pmatrix} \sigma_\mu^2 + \sigma_v^2 & \cdots & \sigma_\mu^2 \\ \vdots & \ddots & \vdots \\ \sigma_\mu^2 & \cdots & \sigma_\mu^2 + \sigma_v^2 \end{pmatrix} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \begin{pmatrix} \sigma_\mu^2 + \sigma_v^2 & \cdots & \sigma_\mu^2 \\ \vdots & \ddots & \vdots \\ \sigma_\mu^2 & \cdots & \sigma_\mu^2 + \sigma_v^2 \end{pmatrix} \end{pmatrix}$$

Rewriting the model as $y = X'\beta + u$, we can use a **pooled OLS** estimator:

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

OLS is unbiased and consistent, but not efficient because of the non-spherical errors. We can therefore use a **pooled GLS** estimator

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

which is consistent and BLUE.

Now, the **Between Estimator** discards all information regarding how individuals change over time and focuses exclusively on the variation *between* individuals. We apply the projection matrix $P_D = D(D'D)^{-1}D' = I_n \otimes \frac{1}{T}J_T$ to our model:

$$P_D y = P_D X\beta + P_D u$$

In scalar form, this is equivalent to running OLS on the individual time-averages:

$$\bar{y}_i = \alpha + \bar{x}_i'\beta + \bar{u}_i, \quad i = 1, \ldots, n$$

where $\bar{u}_i = \mu_i + \bar{v}_i$. The between estimator is written as

$$\hat{\beta}_B = (X'P_D X)^{-1} X'P_D y$$

- **Consistency**: $\hat{\beta}_B$ is consistent if $E[\bar{u}_i|\bar{x}_i] = 0$, which is satisfied under the RE assumption $E[\mu_i|X] = 0$.
- **Efficiency**: It is generally **inefficient** compared to GLS because it collapses nT observations into only n averages, throwing away the "within" variation.

We can also rewrite our within estimator as hat

$$\beta_W = (X'M_D X)^{-1} X'M_D y$$

which is also unbiased and consistent but inefficient.

Now, recall P (which we called P_D) and M (which we called M_D) are symmetric, idempotent, and orthogonal projectors ($P \times M = 0$). Now, since $P = I_n \otimes \frac{1}{T} J_T$, we have $TP = I_n \otimes J_T$. Additionally, recall $M = I_{nT} - I_n \otimes \frac{1}{T} J_T$. It is easy to see that $M + P = I_{nT}$. Therefore, writing Ω as a function of these projectors, we get

$$\Omega = \sigma_v^2 M + (\sigma_v^2 + T\sigma_\mu^2) P$$

This is what we call a **spectral decomposition**: it expresses Ω as a weighted sum of its eigen-projections. Now, because P and M are orthogonal, any power of Ω is found by simply raising the scalar coefficients to that power.To find the inverse, we take the reciprocal of the coefficients:

$$\Omega^{-1} = \frac{1}{\sigma_v^2} M + \frac{1}{\sigma_v^2 + T\sigma_\mu^2} P$$

For GLS, we specifically need the "weighting" matrix $\Omega^{-1/2}$ to transform the data so that the errors become spherical. Applying the same logic, we take the square root of the coefficients:

$$\Omega^{-1/2} = \frac{1}{\sigma_v} M + \frac{1}{\sqrt{\sigma_v^2 + T\sigma_\mu^2}} P$$

By substituting $M = I_{nT} - P$ back into this equation, we arrive at the quasi-demeaning factor θ:

$$\Omega^{-1/2} = \frac{1}{\sigma_v} [I_{nT} - (1 - \sqrt{\theta}) P] \text{ where } \theta = \frac{\sigma_v^2}{\sigma_v^2 + T\sigma_\mu^2}$$

θ is basically the ratio of the idiosyncratic variance to the total weighted variance. For each observation, the transformation is applied as follows:

$$y_{it} - (1 - \sqrt{\theta})\bar{y}_i = [x_{it} - (1 - \sqrt{\theta})\bar{x}_i]'\beta + r_{it}$$

where $\bar{y}_i$ and $\bar{x}_i$ are the individual-specific time averages. r_{it} is the transformed error term which is now spherical (homoskedastic and uncorrelated), allowing for efficient estimation. This is therefore consistent, BLUE and asymptotically normal.

Now, let's try to build some intuition. First, instead of subtracting the entire mean (as in Fixed Effects), we only subtract a portion of it, $(1 - \sqrt{\theta})$, which preserves the between-unit variation, and allows the model to estimate the effects of time-invariant variables and potentially increases efficiency.

Remark 7.1: Interpretation of θ

The parameter $\theta = \frac{\sigma_v^2}{\sigma_v^2 + T\sigma_\mu^2}$ determines the weight of the quasi-demeaning.

- If $\theta = 1$, we have **Pooled OLS**. This occurs when there is no individual-specific variance ($\sigma_\mu^2 = 0$).
- If $\theta = 0$, we have the **Within Estimator**. This occurs as the individual effects dominate ($\sigma_\mu^2 \to \infty$) or as the number of time periods $T \to \infty$.

Essentially, RE is a data-driven choice between OLS and FE based on the relative importance of within-unit and between-unit variance.

7.3 Fixed Effects vs. Random Effects

If both n and T are large, it makes no difference whether we treat individual specific effects μ_i as fixed or random because both the within estimator and the GLS estimator become the same. If T is small and n large, there are differences.

The advantage with FE is that there is no need to assume that individual effects are uncorrelated with covariates, but there is the issue of incidental parameters. For RE, a pro is that the number of parameters is fixed, allowing for efficient estimation, but it relies on a strong assumption on uncorrelation of individual effects with covariates.

The choice can be viewed through two lenses:

- **Inference Type:** Use FE for *conditional inference* on a specific set of unique units (e.g., OECD countries). Use RE for *unconditional inference* where units are treated as homogeneous drawings from a larger population (e.g., a random sample of individuals).
- **The Consistency-Efficiency Trade-off:** RE uses GLS to provide efficient estimates but requires $\text{Cov}(\mu_i, x_{it}) = 0$. FE remains consistent regardless of this correlation but is less efficient as it discards between-unit variation.

Decision Rule: The **Hausman Test** is used to statistically determine if the RE assumption of zero correlation holds. If the test rejects, the RE estimates are biased, and FE must be used.

8 Recent DiD paper

In this section, I discuss a recent paper that I presented during class (Borusyak et al., 2024).

8.1 Research Question

The central research question of this paper asks: How can we estimate dynamic treatment effects in event studies with staggered treatment adoption and heterogeneous causal effects in an efficient manner? This question is particularly relevant given the increasing prevalence of settings where treatment is rolled out to different units at different times, and where the effect of treatment may vary across units and over time.

8.2 Framework

The paper uses a household-level example to illustrate the framework. Each household i is treated (that is, has received a tax rebate) from period E_i onwards. We define $K_{it} = t - E_i$ as the time (in weeks) relative to treatment.

To study the average treatment effect, researchers typically use the canonical regression specification: $Y_{it} = \alpha_i + \beta_t + \tau D_{it} + \varepsilon_{it}$, where $D_{it} = \mathbb{1}\{K_{it} > 0\}$ is an indicator for whether unit i is treated at time t.

However, to study the dynamics of the treatment effect, we need a fully dynamic specification: $Y_{it} = \alpha_i + \beta_t + \sum_{k=-\infty}^{\infty} \tau_k \mathbb{1}\{K_{it} = k\} + \varepsilon_{it}$, where α_i represents individual fixed effects, β_t represents week fixed effects, and τ_k captures the dynamic causal effects at each relative time period k.

The estimation target in this framework is $\tau_w = \sum_{it \in \Omega_1} w_{it} \tau_{it} \equiv w'\tau$, where $\tau_{it} = Y_{it} - Y_{it}(0)$ represents the individual treatment effect for unit i at time t.

8.3 Assumptions

The paper relies on four key assumptions. Assumption 1 is the parallel trend assumption, which states that for all $(i, t) \in \Omega$, the expected value of the untreated potential outcome can be written as $\mathbb{E}[Y_{it}(0)] = \alpha_i + \beta_t$, where $Y_{it}(0)$ is the period t stochastic potential outcome of unit i if it is never treated.

Assumption 2 is the no anticipation assumption, which requires that for all non-treated units $(i, t) \in \Omega_0$, we have $Y_{it} = Y_{it}(0)$. This means that units do not change their behavior in anticipation of treatment.

Assumption 3′ is an optional parametric model of causal effects, which posits that $\tau = \Gamma\theta$, where θ is an $(N_1 - M) \times 1$ vector of unknown parameters and Γ is a known $N_1 \times (N_1 - M)$ matrix of full column rank. This assumption imposes a linear structure on treatment effects

so that they can be efficiently estimated and related to each other.

Assumption 4 concerns spherical errors and states that $\mathbb{E}[\varepsilon\varepsilon'] = \sigma^2 I_N$. While this assumption is useful for the main theorem, it can be relaxed in practice.

8.4 The Underidentification Problem

One of the key insights of the paper is the formal identification of an underidentification problem in fully dynamic specifications. The paper presents the following proposition.

Proposition 8.1: Underidentification

If there are no never-treated units, the path of $\{\tau_h\}_{h\neq -1}$ coefficients is not point-identified in the fully dynamic specification. In particular, for any $\kappa \in \mathbb{R}$, the path $\{\tau_h + \kappa(h+1)\}$ fits the data equally well, with the fixed-effect coefficients appropriately modified.

Proof. In the absence of never-treated units and defining $\tau_{-1} = 0$, we can write

$$\sum_{h\neq -1} \tau_h \mathbb{1}\{K_{it} = h\} = \tau_{K_{it}}.$$

Now consider some collection of $\{\tau_h\}$ (with $\tau_{-1} = 0$) and fixed effects $\tilde{\alpha}_i$ and $\tilde{\beta}_t$. For any $\kappa \in \mathbb{R}$, let

$$\tau_h^\star = \tau_h + \kappa(h+1), \qquad \tilde{\alpha}_i^\star = \tilde{\alpha}_i + \kappa(E_i - 1), \qquad \tilde{\beta}_t^\star = \tilde{\beta}_t - \kappa t.$$

Then for any observation (i,t),

$$\begin{aligned}\tilde{\alpha}_i^\star + \tilde{\beta}_t^\star + \tau_{K_{it}}^\star &= \tilde{\alpha}_i + \kappa(E_i - 1) + \tilde{\beta}_t - \kappa t + \tau_{K_{it}} + \kappa(K_{it} + 1) \\ &= \tilde{\alpha}_i + \tilde{\beta}_t + \tau_{K_{it}} + \kappa(E_i - 1 - t + K_{it} + 1).\end{aligned}$$

Using $K_{it} = t - E_i$, we have $E_i - 1 - t + K_{it} + 1 = 0$, so

$$\tilde{\alpha}_i^\star + \tilde{\beta}_t^\star + \tau_{K_{it}}^\star = \tilde{\alpha}_i + \tilde{\beta}_t + \tau_{K_{it}}.$$

Thus equation (2) has exactly the same fit under the original and modified fixed effects and $\{\tau_h\}$ coefficients, indicating perfect collinearity. □

The fundamental identity underlying this problem is that $K_{it} = t - E_i$, which implies that event time is perfectly collinear with unit and time fixed effects. The conclusion is striking: without restrictions, the model cannot distinguish τ_k from $\tau_k + h \cdot k$. Dynamic treatment effects are identified only up to an arbitrary linear trend.

To understand this result graphically, consider a simple example with two units where one appears to have a trend and level difference relative to the other. There are two possible interpretations of this pattern. First, the treatment might have no impact on the outcome,

with trends merely reflecting features of the environment and the level difference arising from unit fixed effects. Second, the outcome could be entirely driven by causal effects plus anticipation. Without additional assumptions, these interpretations are observationally equivalent.

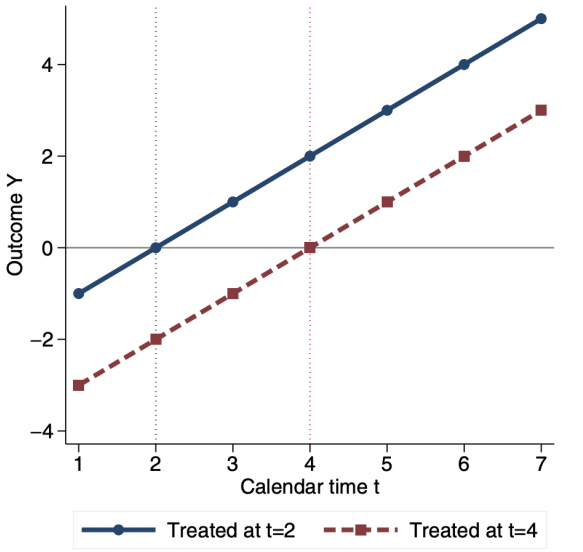

Figure 3: Underidentification graphically

The paper proposes two solutions to this underidentification problem. The first solution is to impose a no-anticipation effect and estimate a semi-dynamic model. However, this requires a strong assumption on anticipation, which must be motivated by an a priori argument. This approach separates estimation and pre-trend testing through an F-test on treatment leads, arbitrarily dropping two leads. The second solution is to introduce a control group of units that are never treated whenever possible, though one should not allow the control group to be on its own trend.

8.5 Negative Weights Problem

The paper also highlights the problem of negative weights that can arise in standard two-way fixed effects (TWFE) estimation. To illustrate this, consider a simple example with two units (A and B) and three periods. Unit A is treated starting in period 2 ($E_A = 2$), while unit B is treated starting in period 3 ($E_B = 3$), as described in the Table 8.5.

Suppose we estimate the TWFE model:

$$Y_{it} = \alpha_i + \beta_t + \tau^{static} D_{it} + \varepsilon_{it}$$

An admissible comparison would be

$$(Y_{A2} - Y_{A1}) - (Y_{B2} - Y_{B1}) = \tau_{A2}$$

$\mathbb{E}[Y_{it}]$	$i = A$	$i = B$
$t = 1$	α_A	α_B
$t = 2$	$\alpha_A + \beta_2 + \tau_{A2}$	$\alpha_B + \beta_2$
$t = 3$	$\alpha_A + \beta_3 + \tau_{A3}$	$\alpha_B + \beta_3 + \tau_{B3}$
Event date	$E_i = 2$	$E_i = 3$

Table 1: Two-Unit, Three-Period Example

which uses unit B as a control for unit A. However, a forbidden comparison would be

$$(Y_{B3} - Y_{B2}) - (Y_{A3} - Y_{A2}) = \tau_{B3} + \tau_{A2} - \tau_{A3}$$

which compares already-treated unit A to newly-treated unit B.

The key conclusion is that if τ_{A3} goes up, the apparent treatment effect on B at time 3 goes down. This means that τ_{A3} receives a negative weight because OLS imposes treatment effect homogeneity. For large enough relative time periods, when there are no never-control units, only forbidden comparisons exist, making estimation of long-run causal effects incorrect.

8.6 The Efficient Estimator

8.6.1 The general theorem

The paper's main theoretical contribution is an efficient estimator theorem.

Theorem 8.1: Efficient Estimator

Suppose Assumptions 1, 2, 3′, and 4 hold. Among all linear unbiased estimators of τ_w, the unique efficient estimator $\hat{\tau}_w^*$ is obtained via:

1. **Estimate**: Within the untreated observations only $(i, t) \in \Omega_0$, estimate the α_i and β_t by OLS in
$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}.$$
2. **Extrapolate**: Set $\hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\beta}_t$ and
$$\hat{\tau}_{it}^* = Y_{it} - \hat{Y}_{it}(0) \text{ for each treated observation } (i, t) \in \Omega_1$$
3. **Take Averages**: Estimate the target τ_w by a weighted sum,
$$\hat{\tau}_w^* = \sum_{(i,t)\in\Omega_1} w_{it}\, \hat{\tau}_{it}^*.$$

This is a powerful result because it shows that efficiency can be achieved while relaxing some of the more stringent assumptions.

8.6.2 The Imputation Estimator

The paper provides an imputation representation of the efficient estimator that is particularly useful for practitioners.

Theorem 8.2: Imputation representation of the efficient estimator

With a null Assumption 3′ (that is, if $\Gamma = I_{N_1}$), the unique efficient linear unbiased estimator $\hat{\tau}_w^*$ of τ_w from Theorem 0.1 can be obtained via an imputation procedure:

1. **Estimate**: Within the untreated observations only $(i,t) \in \Omega_0$, estimate the α_i and β_t by OLS in
$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}.$$
2. **Extrapolate**: Set $\hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\beta}_t$ and
$$\hat{\tau}_{it}^* = Y_{it} - \hat{Y}_{it}(0) \text{ for each treated observation } (i,t) \in \Omega_1$$
3. **Take Averages**: Estimate the target τ_w by a weighted sum,
$$\hat{\tau}_w^* = \sum_{(i,t)\in\Omega_1} w_{it}\, \hat{\tau}_{it}^*.$$

8.7 Pre-trend Testing

One of the key insights of the paper is the separation of estimation of treatment effects from pre-trends testing. To test for pre-trends in the imputation framework, use the sample of untreated or not yet treated observations to estimate

$$Y_{it} = \alpha_i + \beta_t + \sum_{k=E_i-R}^{E_i-1} \tau_k \mathbb{1}\{K_{it} = k\} + \varepsilon_{it}$$

for $(i,t) \in \Omega_0$ Then perform an F-test of joint significance, testing

$$\mathcal{H}_0 : \tau_{E_i-R} = \cdots = \tau_{E_i-1} = 0$$

This approach ensures that the pre-trend test is not contaminated by the estimation of treatment effects.

8.8 Application: Marginal Propensity to Consume Out of Tax Rebates

The paper provides an application studying the Economic Stimulus Act of 2008, a 100 billion dollar programme that sent tax rebates to approximately 130 million tax filers. The goal is to estimate the marginal propensity to consume (MPC) out of the tax rebate, following Broda and Parker, 2014.

The timing of the tax rebate (measured in weeks) was determined by the last two digits of the taxpayer's Social Security number, providing plausibly exogenous variation in treatment timing. The researchers use the following specification:

$$Y_{it} = \alpha_i + \beta_t + \sum_{h=-a}^{b} \tau_h \mathbb{1}\{K_{it} = h\} + \varepsilon_{it}$$

where Y_{it} is the dollar amount of spending in calendar week t for household i.

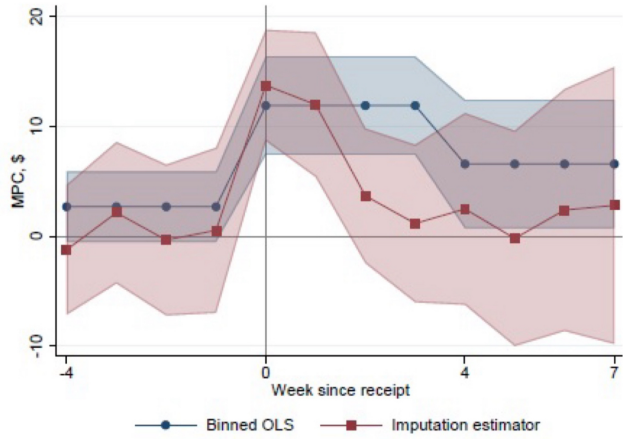

Figure 4: Binned OLS vs. Imputation Estimator

Figure 8.8 shows the difference between OLS and the imputation estimator. The results comparing binned OLS versus the imputation estimator are striking. The imputation estimator shows much less persistence than binned OLS. The two-month MPCs are very different: \$73.8 for binned OLS versus \$38 for the imputation estimator. This difference arises because binned OLS uses forbidden comparisons that inflate long-run effects.

The reason for this discrepancy can be understood by considering how OLS estimates treatment effects at longer horizons. Tax rebates were sent out in a seven-week window. Information on treatment effects eight weeks after treatment comes from late periods in the sample and from individuals who were treated early. Crucially, we can never observe τ_8 and τ_{-1} at the same time: if anyone is observed eight weeks after treatment, everyone else must already be treated. Therefore, OLS performs extrapolations by observing τ_8 and τ_3 at the same time, as well as τ_3 and τ_{-1}, and inferring something about the difference between τ_8 and τ_{-1}. This works if treatment effects are homogeneous, but not otherwise, especially if people treated earlier have larger treatment effects, which would be a forbidden extrapolation.

8.9 Asymptotic Properties

The paper also establishes the asymptotic properties of the imputation estimator. Assumption 5 concerns clustered error terms and requires that errors ε_{it} are independent across

units i and have bounded variance, with $Var(\varepsilon_{it}) \leq \bar{\sigma}^2$ for all $(i,t) \in \Omega$. Assumption 6 is the Herfindahl condition, which requires that along the asymptotic sequence, the Herfindahl norm of the weights

$$\|v\|_H^2 = \sum_i \left(\sum_{t:(i,t)\in\Omega} |v_{it}| \right)^2$$

converges to zero, where v_{it} are the weights in the unbiased linear estimator

$$\hat{\tau}_w = \sum_{(i,t)\in\Omega} v_{it} Y_{it}$$

Under Assumptions 1′, 2, 3′, 5, and 6, the paper establishes consistency:

$$\hat{\tau}_w - \tau_w \xrightarrow{L^2} 0$$

for any unbiased linear estimator $\hat{\tau}_w$ of τ_w, such as $\hat{\tau}_w^*$ from the main theorem.

The paper also establishes asymptotic normality. If the assumptions of the consistency proposition hold, and there exists $\kappa > 0$ such that $\mathbb{E}[|\varepsilon_{it}|^{2+\kappa}]$ is uniformly bounded, the weights are not overly concentrated (formally, $\sum_i \left(\sqrt{n_H} \sum_{t:(i,t)\in\Omega} |v_{it}| \right)^{2+\kappa} \longrightarrow 0$), and the variance does not vanish ($\liminf n_H \sigma_w^2 > 0$), then with $\sigma_w^2 = Var[\hat{\tau}_w]$, we have $\sigma_w^{-1}(\hat{\tau}_w - \tau_w) \xrightarrow{d} \mathcal{N}(0,1)$.

8.10 Key Takeaways

The paper provides several important lessons for practitioners. First, researchers should use difference-in-differences if they have an ex-ante reason to believe that the parallel trends and no anticipation assumptions hold. Second, it is crucial to separate pre-trends testing from the estimation of treatment effects to avoid contamination. Third, researchers should use a valid control group even in the absence of never-treated units, and the imputation estimator provides the efficient way to do so.

References

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, *91*(6), 3253–3285.

Broda, C., & Parker, J. A. (2014). The economic stimulus payments of 2008 and the aggregate demand for consumption. *Journal of Monetary Economics*, *68*, S20–S36.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, *69*(5), 1127–1160.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. (2022). Valid t-ratio inference for iv. *American Economic Review*, *112*(10), 3260–3290.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

Staiger, D. O., & Stock, J. H. (1994). Instrumental variables regression with weak instruments.

A Appendix

A.1 Asymptotic Tools for i.n.i.d. Data

When the error covariance is non-spherical, the sequence of terms used in consistency proofs, such as $\mathbf{x}_i\varepsilon_i$, is often **independent but not identically distributed (i.n.i.d.)**. In this case, standard i.i.d. Laws of Large Numbers (LLN) cannot be directly applied.
To ensure the Weak Law of Large Numbers (WLLN) holds for a sequence of independent random variables $\{X_i\}_{i=1}^n$ (where $X_i = \mathbf{x}_i\varepsilon_i$ in the context of OLS with non-spherical errors), the key requirement is often the condition of **Uniform Integrability (UI)**.

Definition A.1: Uniform Integrability

A sequence of random variables $\{X_i, i \geq 1\}$ is uniformly integrable if

$$\lim_{\delta\to\infty} \sup_{i\geq 1} \mathbb{E}[|X_i| \cdot \mathbb{1}\{|X_i| > \delta\}] = 0$$

The core idea behind uniform integrability is controlling the probability mass in the tails of the distribution. For convergence theorems like the WLLN, we need to ensure that the tail behavior (where $|X_i|$ is large, i.e., beyond δ) does not become overwhelmingly large for any single variable X_i in the sequence.

The condition requires that the expected value of the tail (the portion of the distribution where $|X_i| > \delta$) goes to zero **uniformly** for all variables X_i in the sequence as δ increases. This ensures that no individual variable "carries too much weight" in the limit, which allows the Law of Large Numbers to hold even when the individual variances (second moments) are not bounded. We now introduce sufficient and necessary conditons for uniform integrability.

Theorem A.1: Necessary Conditions for UI

The following condition is necessary for a sequence of random variables $\{X_i, i \geq 1\}$ to be uniformly integrable (UI):

$$\max_{i\geq 1} \mathbb{E}(|X_i|) < \infty$$

Theorem A.2: Sufficient Conditions for UI

The following conditions are sufficient for a sequence of random variables $\{X_i, i \geq 1\}$ to be uniformly integrable (UI):

1. **Existence of a Moment Higher than 1:** If $\sup_{i\geq 1} \mathbb{E}\{|X_i|^{1+\eta}\} < \infty$ for some $\eta > 0$, then $\{X_i, i \geq 1\}$ is UI.
2. **Identically Distributed:** If $\{X_i, i \geq 1\}$ is identically distributed (I.D.) and

$\mathbb{E}|X_i| < \infty$, then $\{X_i, i \geq 1\}$ is UI.

The proof of these theorems are fairly easy, but are omitted. What is interesting in our application is mostly sufficient condition 1. Indeed, it is enough to have finiteness of a moment slighlty above one to have uniform integrability, which we will see below will allow us to derive another version of the WLLN.

Theorem A.3: WLLN for Independent, Uniformly Integrable Sequences

Let $\{X_i\}_{i=1}^n$ be a sequence of independent and uniformly integrable random variables. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \xrightarrow{p} 0.$$

Application to OLS Consistency: This WLLN is vital in showing the consistency of the OLS estimator ($\hat{\beta} \xrightarrow{p} \beta$) when the data is independent but not identically distributed (i.n.i.d.) due to heteroskedasticity. Specifically, it ensures that the critical term $\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i$ converges in probability to 0, provided the sequence $\{x_i \varepsilon_i\}$ is uniformly integrable.

A.2 Additional derivations

A.2.1 Restricted Least Squares Derivation

Consider the linear regression model

$$y = X\beta + u$$

and the linear restrictions

$$R\beta = c.$$

The restricted least squares (RLS) estimator is obtained by minimizing

$$(y - X\beta)'(y - X\beta)$$

subject to the constraint $R\beta = c$.
The Lagrangian is

$$\mathcal{Q}(\beta, \lambda) = (y - X\beta)'(y - X\beta) - 2\lambda'(R\beta - c).$$

Differentiating with respect to β and λ gives

$$\frac{\partial \mathcal{Q}}{\partial \beta} = -2X'y + 2X'X\beta - 2R'\lambda = 0, \tag{A.1}$$

$$\frac{\partial \mathcal{Q}}{\partial \lambda} = R\beta - c = 0. \tag{A.2}$$

From (A.1) we obtain

$$X'X\beta - R'\lambda = X'y,$$

which implies

$$\beta = (X'X)^{-1}X'y + (X'X)^{-1}R'\lambda.$$

Define the unrestricted OLS estimator

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Then

$$\beta = \hat{\beta} + (X'X)^{-1}R'\lambda.$$

Substitute into (A.2):

$$R\hat{\beta} + R(X'X)^{-1}R'\lambda = c.$$

Rearranging,

$$R(X'X)^{-1}R'\lambda = c - R\hat{\beta}.$$

Assuming $R(X'X)^{-1}R'$ is nonsingular,

$$\lambda = \left[R(X'X)^{-1}R'\right]^{-1}(c - R\hat{\beta}).$$

Substituting back,

$$\tilde{\beta} = \hat{\beta} + (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(c - R\hat{\beta}).$$

This is the restricted least squares estimator.