Chapters 3–8 (Merged Notes)

Compiled from Existing Chapters
Format unified based on Chapter 7

Chapter 3: Generalized Least Squares (GLS)

1 Motivation: When OLS Assumptions Fail

In our previous discussions of Ordinary Least Squares (OLS), we relied heavily on the Gauss-Markov assumptions. Specifically, we assumed spherical errors:

$$Var(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2\mathbf{I}$$

This implies two things:

1. **Homoskedasticity:** The variance of the error term is constant across all observations ($Var(\epsilon_i) = \sigma^2$).
2. **No Autocorrelation:** The errors are uncorrelated between observations ($Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$).

However, economic data rarely behaves this nicely.

1.1 Example: The Gravity Model of Trade

Consider the classic Gravity Model of Trade. A theoretical formulation might look like this:

$$T_{ij} = \alpha \frac{Y_i^{\beta_1} Y_j^{\beta_2}}{D_{ij}^{\gamma}} \cdot \eta_{ij}$$

To estimate this, we log-linearize the model (as seen in Silva and Tenreyro, REStat 2006):

$$\ln T_{ij} = \ln \alpha + \beta_1 \ln Y_i + \beta_2 \ln Y_j - \gamma \ln D_{ij} + \ln \eta_{ij}$$

Let $\epsilon_{ij} = \ln \eta_{ij}$. If the original error term η_{ij} follows a Log-Normal distribution, the variance of $\ln \eta_{ij}$ might depend on the scale of trade or other factors, leading to **heteroskedasticity**. Furthermore, trade flows involving the same country might be correlated, leading to spatial **autocorrelation**.

When $Var(\boldsymbol{\epsilon}|\mathbf{X}) \neq \sigma^2\mathbf{I}$, OLS is no longer the Best Linear Unbiased Estimator (BLUE). We need a more general framework.

—

2 Heteroskedasticity

2.1 Definition and Examples

Heteroskedasticity occurs when the variance of the error term depends on the regressors x_i:

$$Var(\epsilon_i|x_i) = \sigma_i^2 \neq \sigma^2$$

Economic Intuition:

- **Engel Curves (Food Expenditure):** As wealth increases, the *variance* in food expenditure increases. Poor families must spend a fixed portion on food (low variance), while wealthy families can spend a little or a lot (high variance).
- **Financial Markets (ARCH):** Volatility often clusters. Current variance may depend on past variance (autoregressive conditional heteroskedasticity).

2.2 Consequences for OLS

If we ignore heteroskedasticity and use OLS:

1. **Unbiasedness:** $\hat{\beta}^{OLS}$ remains unbiased ($E[\hat{\beta}|\mathbf{X}] = \beta$).
2. **Consistency:** $\hat{\beta}^{OLS}$ remains consistent (plim $\hat{\beta} = \beta$).
3. **Inefficiency:** OLS is no longer efficient. There exists a weighted estimator with lower variance.
4. **Inference Failure:** The standard formula for the variance-covariance matrix, $\widehat{Var}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, is **wrong**. Hypothesis testing (t-stats, F-stats) will be invalid.

The true variance of the OLS estimator under general errors is:

$$Var(\hat{\beta}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\epsilon\epsilon'|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

This is often called the "Sandwich Form."

2.3 Testing for Heteroskedasticity

We test the Null Hypothesis $H_0 : \gamma_1 = \cdots = \gamma_p = 0$ (Homoskedasticity).

2.3.1 1. Breusch-Pagan (BP) Test

Run OLS to get residuals $\hat{\epsilon}_i$. We assume the variance is a linear function of variables $\mathbf{Z}$:

$$\hat{\epsilon}_i^2 = \gamma_0 + \gamma_1 z_{i1} + \cdots + \gamma_p z_{ip} + u_i$$

The test statistic is the Lagrange Multiplier (LM) statistic:

$$LM = N \cdot R^2 \sim \chi^2(p)$$

Note: The BP test is sensitive to the assumption of normality of errors.

2.3.2 2. White Test

This is a general test that does not presume a specific form of heteroskedasticity. We regress $\hat{\epsilon}_i^2$ on all explanatory variables x_i, their squares, and their cross-products.

$$\hat{\epsilon}_i^2 = \gamma_0 + x_i\gamma + (\text{squares and cross products}) + u_i$$

Statistic: $N \cdot R^2 \sim \chi^2(\text{df})$. *Drawback:* This uses many degrees of freedom. If N is small, the test has low power.

—

3 Remedies

3.1 Approach 1: Robust Standard Errors (White's SE)

If we want to keep the OLS point estimates but fix the hypothesis testing, we use the Heteroskedasticity-Consistent (HC) covariance matrix estimator (Eicker-Huber-White).

Recall the "Sandwich" form. We estimate the middle term $\mathbf{X'\Sigma X}$ using the squared residuals:

$$\widehat{S} = \sum_{i=1}^{N} \hat{\epsilon}_i^2 x_i' x_i$$

Thus, the robust variance estimator is:

$$\widehat{Var}_{Robust}(\hat{\beta}) = (\mathbf{X'X})^{-1} \left(\sum_{i=1}^{N} \hat{\epsilon}_i^2 x_i' x_i \right) (\mathbf{X'X})^{-1}$$

Pros: Consistent inference even if the form of heteroskedasticity is unknown.
Cons: Does not improve efficiency of coefficients.

3.2 Approach 2: Weighted Least Squares (WLS)

If we know the form of the heteroskedasticity, i.e., $\sigma_i^2 = \sigma^2 f(x_i)$, we can restore efficiency (BLUE) by transforming the model.

Strategy: Give less weight to high variance observations (they contain less information about the mean). Define weight $w_i = \frac{1}{\sqrt{f(x_i)}}$. Divide the entire equation by $\sqrt{f(x_i)}$:

$$\frac{y_i}{\sqrt{f(x_i)}} = \frac{x_i}{\sqrt{f(x_i)}}\beta + \frac{\epsilon_i}{\sqrt{f(x_i)}}$$

Let $y_i^*, x_i^*, \epsilon_i^*$ be the transformed variables.

$$Var(\epsilon_i^*) = Var\left(\frac{\epsilon_i}{\sqrt{f(x_i)}}\right) = \frac{1}{f(x_i)} Var(\epsilon_i) = \frac{\sigma^2 f(x_i)}{f(x_i)} = \sigma^2$$

The transformed error ϵ_i^* is homoskedastic! We can now run OLS on the transformed data.

—

4 Generalized Least Squares (GLS)

4.1 The General Framework

Let us generalize to the case where $\mathbf{\Omega} = Var(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2\mathbf{\Psi}$, where $\mathbf{\Psi}$ is a generic positive definite symmetric (p.d.s.) matrix. This covers both heteroskedasticity and autocorrelation.

Since $\mathbf{\Psi}$ is p.d.s., its inverse $\mathbf{\Psi}^{-1}$ is also p.d.s. We can perform a Cholesky decomposition (or "matrix square root"):

$$\mathbf{\Psi}^{-1} = \mathbf{P}'\mathbf{P}$$

where $\mathbf{P}$ is a non-singular matrix.

4.2 The GLS Transformation

Pre-multiply the original model $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$ by $\mathbf{P}$:

$$\mathbf{Py} = \mathbf{PX}\beta + \mathbf{P}\boldsymbol{\epsilon} \quad \Rightarrow \quad \mathbf{y}^* = \mathbf{X}^*\beta + \boldsymbol{\epsilon}^*$$

Now, analyze the variance of the transformed error $\boldsymbol{\epsilon}^*$:

$$Var(\boldsymbol{\epsilon}^*) = E[\mathbf{P}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{P}'] = \mathbf{P}E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']\mathbf{P}' = \mathbf{P}(\sigma^2\mathbf{\Psi})\mathbf{P}'$$

From our definition, $\mathbf{\Psi} = (\mathbf{P}'\mathbf{P})^{-1} = \mathbf{P}^{-1}(\mathbf{P}')^{-1}$. Substituting this back:

$$Var(\boldsymbol{\epsilon}^*) = \sigma^2\mathbf{P}\left(\mathbf{P}^{-1}(\mathbf{P}')^{-1}\right)\mathbf{P}' = \sigma^2\mathbf{I}$$

The transformed model satisfies the classical OLS assumptions.

4.3 The GLS Estimator

Running OLS on the transformed variables gives the GLS estimator:

$$\hat{\beta}^{GLS} = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^* = (\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{y}$$

Since $\mathbf{P}'\mathbf{P} = \mathbf{\Psi}^{-1}$:

$$\boxed{\hat{\beta}^{GLS} = (\mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Psi}^{-1}\mathbf{y}}$$

4.4 Properties of GLS

1. **Unbiasedness:** $E[\hat{\beta}^{GLS}|\mathbf{X}] = \beta$.
2. **Consistency:** Requires $plim\ \frac{1}{N}\mathbf{X}^{*\prime}\boldsymbol{\epsilon}^* = 0$.

3. **Asymptotic Normality:**

$$\sqrt{N}(\hat{\beta}^{GLS} - \beta) \xrightarrow{d} N(0, \sigma^2 \mathbf{Q}^{*-1})$$

where $\mathbf{Q}^* = E[x_i^{*\prime} x_i^*]$.

4.5 Feasible GLS (FGLS)

In practice, $\mathbf{\Psi}$ is unknown. If we parameterize $\mathbf{\Psi}(\theta)$, we can estimate $\hat{\theta}$ first (e.g., using OLS residuals), construct $\hat{\mathbf{\Psi}}$, and then calculate GLS. This is **Feasible GLS**. *Note: FGLS is asymptotically equivalent to GLS, but finite sample properties can differ.*

5 Seemingly Unrelated Regressions (SURE)

5.1 The System

Consider a system of M equations (e.g., a demand system for Food, Clothing, Rent). For the m-th equation ($m = 1, \ldots, M$) and individual i:

$$y_{im} = x_{im}\beta_m + \epsilon_{im}$$

Stacking observations for equation m:

$$\mathbf{y}_m = \mathbf{X}_m \beta_m + \boldsymbol{\epsilon}_m$$

5.2 The Stacked Model

We can stack all M equations into one "super" model:

$$\begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_M \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{X}_2 & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_M \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_M \end{pmatrix}$$

Or compactly: $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$.

5.3 The Variance Structure (Kronecker Product)

Here is the key: The errors for the *same* individual might be correlated across equations (e.g., an unobserved shock to income affects both food and clothing consumption).

$$Cov(\epsilon_{ip}, \epsilon_{iq}) = \sigma_{pq}$$

However, individuals are independent: $Cov(\epsilon_{ip}, \epsilon_{jq}) = 0$ for $i \neq j$.

The covariance matrix of the stacked error $\boldsymbol{\epsilon}$ is:

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I}_N = \begin{pmatrix} \sigma_{11}\mathbf{I} & \sigma_{12}\mathbf{I} & \cdots \\ \sigma_{21}\mathbf{I} & \sigma_{22}\mathbf{I} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

where $\boldsymbol{\Sigma}$ is the $M \times M$ covariance matrix of errors across equations, and $\otimes$ denotes the Kronecker product.

5.4 The SURE Estimator

Applying the GLS formula to this stacked system:

$$\hat{\beta}^{SURE} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$$

Using the property $(\boldsymbol{\Sigma} \otimes \mathbf{I})^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}$:

$$\boxed{\hat{\beta}^{SURE} = [\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{X}]^{-1}\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y}}$$

5.5 When Does SURE Matter?

There are two special cases where SURE collapses to simple Equation-by-Equation OLS:

1. **Diagonal $\boldsymbol{\Sigma}$:** If $\sigma_{pq} = 0$ for $p \neq q$, the equations are truly unrelated.
2. **Identical Regressors ($X_m = X$):** If every equation has the exact same explanatory variables.

Proof Sketch for Case 2: If $\mathbf{X}_1 = \cdots = \mathbf{X}_M = \mathbf{X}_0$, then $\mathbf{X} = \mathbf{I}_M \otimes \mathbf{X}_0$. The term $\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{X}$ becomes:

$$(\mathbf{I} \otimes \mathbf{X}_0')(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})(\mathbf{I} \otimes \mathbf{X}_0) = \boldsymbol{\Sigma}^{-1} \otimes (\mathbf{X}_0'\mathbf{X}_0)$$

Substituting this into the estimator, the $\boldsymbol{\Sigma}^{-1}$ terms eventually cancel out, leaving the OLS estimator for each equation.

Intuition: SURE gains efficiency by using the correlation across errors to "borrow information" from other equations. If the regressors are identical, the variation in X explains the variation in y in the same "direction" for all equations, so the cross-equation error correlation provides no additional leverage for identifying β.

GLS.

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

SURE 是一个特例. $\Omega^{-1} = (\Sigma \otimes I)^{-1} = \Sigma^{-1} \otimes I$.

$Var(\hat{\beta}^{GLS}|X) = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\underbrace{E[\varepsilon\varepsilon']}_{\equiv I\ (GLS)}\Omega^{-1}X(X'\Omega^{-1}X)^{-1}$

$= (X'\Omega^{-1}X)^{-1}$

A. $Var(\hat{\beta}^{GLS}) = \frac{Q^{-1}}{N}$ ← 随样本 $\sigma^2 = 1$. } 一回事儿.

relationship: $\frac{1}{N}X'\Omega^{-1}X \to Q$ as $N \to \infty$.

GLS 先找 Ω^{-1} 通过 OLS.

SURE 先找 Σ^{-1}

Chapter 4: Diagnostics, Outliers, and Robust Regression

6 Introduction

In standard Ordinary Least Squares (OLS) regression, we often rely on the assumption that the error terms are normally distributed, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, specifically for finite-sample inference (t-tests and F-tests). Furthermore, OLS minimizes the sum of *squared* residuals, which implies that it penalizes large errors heavily. This makes OLS highly sensitive to data anomalies.

In this lecture, we explore how to detect deviations from normality and how to identify and handle observations that exert undue influence on our regression estimates.

7 Non-Normality

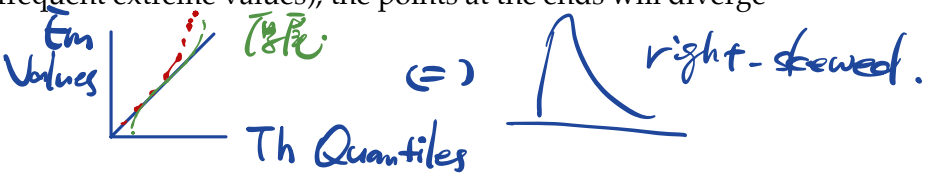
How do we check if our residuals $\hat{\epsilon}_i$ follow a normal distribution?

7.1 Visual Diagnostics

1. **Histogram**: Plot the frequency distribution of the residuals $\hat{\epsilon}_i$. We look for a bell-shaped curve centered at zero.

2. **Quantile-Quantile (Q-Q) Plot**: This plots the *Empirical Quantiles* (from our data) against the *Theoretical Quantiles* (from the Standard Normal distribution).

    - If the data is Normal, the points should fall approximately on a 45-degree line.
    - **Intuition**: Deviations from the line indicate "heavy tails" (outliers) or "light tails". If the tails are "heavy" (frequent extreme values), the points at the ends will diverge from the straight line.



7.2 The Jarque-Bera Test

Visual inspection is subjective. We use the **Jarque-Bera (JB) Test** for a formal statistical conclusion. The normal distribution is defined entirely by its first two moments (mean and variance); its third moment (Skewness) should be 0, and its fourth moment (Kurtosis) should be 3.

The test statistic is constructed as follows:

$$JB = \frac{N}{6}\left(\hat{S}^2 + \frac{1}{4}(\hat{K}-3)^2\right) \sim \chi^2_2 \tag{1}$$

Where:

- N is the sample size.
- $\hat{S}$ is the sample Skewness (measure of asymmetry).
- $\hat{K}$ is the sample Kurtosis (measure of tail thickness).

$$\hat{S} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\sigma^3}$$

$$\hat{K} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^4}{\sigma^4}$$

Intuition: Why does it follow a Chi-squared distribution with 2 degrees of freedom?

- Under the null hypothesis of Normality (H_0), $S = 0$ and $K = 3$.
- We are essentially testing two restrictions simultaneously: Is Skewness close to 0? AND Is Kurtosis close to 3?
- If JB is large (greater than the critical value, e.g., 5.99 at 5% significance), we reject the null hypothesis of normality.

8 Outliers and Leverage

Not all data points are created equal. Some points have a much stronger ability to pull the regression line towards them.

8.1 Univariate Case

Consider the simple regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2}$$

We define the **Leverage** (h_i) of observation i as:

$$h_i = \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{N}(X_j - \bar{X})^2} \tag{3}$$

Economic Intuition (The Fulcrum): Think of the regression line as a seesaw. The sample mean $\bar{X}$ is the fulcrum (pivot point).

- An observation where X_i is close to $\bar{X}$ has low leverage (little torque).
- An observation where X_i is very far from $\bar{X}$ has **high leverage**. Just like sitting at the very end of a seesaw, a small push (change in Y) here moves the whole plank significantly.

Mathematically, the OLS slope $\hat{\beta}_1$ can be viewed as a weighted sum of slopes calculated from individual points relative to the mean:

$$\hat{\beta}^{OLS} \approx \sum_i h_i \left(\frac{y_i - \bar{y}}{x_i - \bar{x}} \right) \tag{4}$$

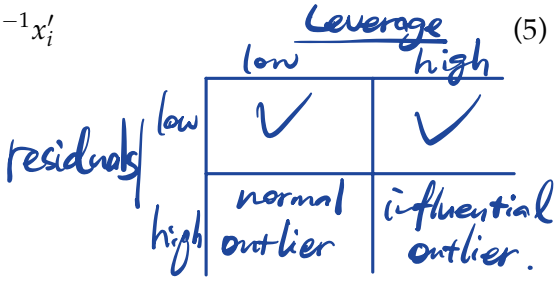
High leverage points (h_i) dominate the determination of $\hat{\beta}$.

8.2 Multivariate Case

In the general model $y_i = x_i\beta + \epsilon_i$ (where x_i is a $1 \times K$ row vector), the leverage is the i-th diagonal element of the "Hat Matrix" ($P = X(X'X)^{-1}X'$):

$$h_{ii} = x_i(X'X)^{-1}x_i' \tag{5}$$

Note: $0 \leq h_{ii} \leq 1$.



9 Detecting Influential Observations

A point can be an outlier in the X-direction (High Leverage) or an outlier in the Y-direction (Large Residual). A point is truly problematic, or **Influential**, if it has *both*.

9.1 Cook's Distance

We use **Cook's Distance (D_i)** to summarize the influence of observation i:

$$D_i = \underbrace{\frac{\hat{\epsilon}_i^2}{K\hat{\sigma}^2}}_{\text{Outlier in } Y} \cdot \underbrace{\frac{h_{ii}}{(1-h_{ii})^2}}_{\text{Outlier in } X \text{ (Leverage)}} \tag{6}$$

Intuition:

- The first term measures how well the model fits observation i. If the residual $\hat{\epsilon}_i$ is huge, the model "missed" this point.
- The second term measures how far x_i is from the center of the data.
- **Interpretation**: If D_i is large (rule of thumb: $D_i > 1$ or sometimes $4/N$), removing this single observation would significantly change the estimated coefficients $\hat{\beta}$.

10 Remedies

If we find significant non-normality or influential outliers, what can we do?

10.1 Method 1: Exclusion

The simplest approach is to simply **drop** the observations identified as outliers.

- *Caution:* This should be done only if there is a valid reason (e.g., data entry error). Dropping valid data just to make the model "fit better" introduces bias.

10.2 Method 2: Winsorization

Instead of deleting data, we transform extreme values.

- **Procedure**: Replace values below the 2.5th percentile with the 2.5th percentile value, and values above the 97.5th percentile with the 97.5th percentile value.
- **Example**: If data is $\{1, 2, \ldots, 100, 1000\}$, and we Winsorize the top value, 1000 might become 100.
- *Note:* Named after Charles P. Winsor (often confused with the Royal House of Windsor, but unrelated).

10.3 Method 3: Robust Regression (LAD)

OLS minimizes squared errors ($\sum e_i^2$), which gives disproportionate weight to large errors (outliers). Alternatively, we can use **Least Absolute Deviations (LAD)**:

$$\hat{\beta}^{(LAD)} = \arg\min_{\beta} \sum_{i=1}^{N} |y_i - x_i\beta| \tag{7}$$

Properties of LAD:

- **Robustness**: This is equivalent to **Median Regression**. Just as the median is less sensitive to extreme values than the mean, LAD is less sensitive to outliers than OLS.
- **Computation**: Unlike OLS, there is **no analytical solution** (no closed-form formula like $(X'X)^{-1}X'Y$). We must use numerical optimization methods (e.g., linear programming).

10.4 Quantile Regression

LAD is a specific case of **Quantile Regression**. While OLS estimates the conditional *mean* $E[y|x]$, Quantile Regression allows us to estimate the conditional *quantile* $Q_\tau(y|x)$ for any $\tau \in (0,1)$.

- If $\tau = 0.5$, this is LAD (Median).
- We can estimate how x affects the 10th percentile ($\tau = 0.1$) or 90th percentile of y.
- **Intuition**: This allows us to see if the effect of a policy (or variable) is different for the "poor" (low quantile) versus the "rich" (high quantile), providing a much richer picture than simple OLS.

OLS $E[y|x]$

LAD conditional median $M(y|x)$

Quantile reg. $Q_\tau(y|x)$ - 加权的 LAD.

$\tau = 0.9$. => 90% 点在下面. 10% 点在上面.

权重更大.

Chapter 5: Endogeneity and Instrumental Variables

11 Introduction: The Breakdown of OLS

Welcome back. In our previous lectures, we built the Ordinary Least Squares (OLS) estimator under ideal conditions. Today, we look at what happens when the world isn't perfect. We are discussing **Endogeneity**—the central villain in the story of causal inference.

Recall our linear model:

$$y_i = x_i\beta + \epsilon_i \tag{8}$$

where x_i is a row vector of regressors ($1 \times K$) and β is a column vector of coefficients ($K \times 1$).

For OLS to be a consistent estimator of the true causal parameter β, we rely on specific assumptions about the relationship between our regressors x_i and the error term ϵ_i.

11.1 Defining Exogeneity

What does it mean for a variable to be exogenous? We define it through two conditions, arranged by strictness:

1. **Strict Exogeneity:**

$$E[\epsilon_i|x_i] = 0 \tag{9}$$

This is a strong condition stating that the mean of the error term is zero for any value of x_i.

2. **Contemporaneous Uncorrelation (Orthogonality):**

$$E[x_i'\epsilon_i] = 0 \tag{10}$$

This is a weaker condition, requiring only that the regressor and the error term are uncorrelated.

Definition of Endogeneity: If these conditions fail—specifically, if $E[x_i'\epsilon_i] \neq 0$—the variable x_i is **endogenous**.

Economic Intuition: The error term ϵ_i contains all the unobserved factors affecting y_i. If x_i is correlated with ϵ_i, when x_i moves, ϵ_i tends to move with it. OLS, which is essentially a covariance machine, will mistakenly attribute the effect of the unseen ϵ_i to x_i. This is "guilt by association."

Law of Iterated Expectations.

12 I. Causes of Endogeneity

Why does endogeneity arise? We categorize the culprits into four main groups.

12.1 1. Autoregression with Serial Correlation

Consider a dynamic time-series model where the current value depends on the past value:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \epsilon_t \tag{11}$$

Suppose the error term follows an AR(1) process (serial correlation):

$$\epsilon_t = \rho \epsilon_{t-1} + u_t \tag{12}$$

The Logic: Here, y_{t-1} is clearly determined by ϵ_{t-1} (from the previous period's equation). However, because of serial correlation, ϵ_{t-1} is part of ϵ_t. Therefore, the regressor y_{t-1} is correlated with the current error ϵ_t.

12.2 2. Measurement Errors

In economics, we rarely observe the "true" variable perfectly. Let the **True Model** be:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{(where } x_i \text{ is the true value)} \tag{13}$$

However, we do not see x_i. We observe a noisy proxy w_i:

$$w_i = x_i + u_i \implies x_i = w_i - u_i \tag{14}$$

where u_i is the measurement error.

Substituting the observed w_i into the true model:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 (w_i - u_i) + \epsilon_i \\ y_i &= \beta_0 + \beta_1 w_i + (\epsilon_i - \beta_1 u_i) \end{aligned} \tag{15}$$

The new composite error term is $v_i = (\epsilon_i - \beta_1 u_i)$. Since $w_i = x_i + u_i$, the observed regressor w_i is positively correlated with the measurement error u_i. Because u_i is now inside the regression error term (with a negative sign), $Cov(w_i, v_i) \neq 0$.

Result: This usually causes **Attenuation Bias**, driving the estimated coefficient toward zero.

12.3 3. Omitted Variable Bias (OVB)

This is the most common form of endogeneity.

- **True Model:** $y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \epsilon_i$.
- **Empirical (Estimated) Model:** We omit w_i (perhaps data is unavailable) and run:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{16}$$

Here, the new error term is $u_i = \beta_2 w_i + \epsilon_i$. Endogeneity arises if and only if $Cov(x_i, w_i) \neq 0$.

Economic Intuition: If high-ability people (w_i, unobserved) tend to get more education (x_i), and ability also increases wages (y_i), ignoring ability means education gets "credit" for the wage increase that was actually caused by ability.

12.4 4. Simultaneity (Reverse Causality)

This occurs when y and x are jointly determined. Consider a system:

$$h_i = \beta_0 + \beta_1 w_i + \beta_2 z_i + u_i \quad \text{(Eq 1)} \tag{17}$$

$$w_i = \gamma_0 + \gamma_1 h_i + \gamma_2 z_i + \nu_i \quad \text{(Eq 2)} \tag{18}$$

If w_i influences h_i, and h_i simultaneously influences w_i, then a shock to h_i (u_i) affects h_i, which in turn affects w_i. Thus, w_i is correlated with u_i. This requires solving for the **Structural Form** vs. **Reduced Form**.

13 II. Properties of OLS under Endogeneity

If we ignore endogeneity and blindly use OLS, what happens?

13.1 1. Bias (Finite Sample Property)

The OLS estimator is:

$$\hat{\beta}^{ols} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\epsilon \tag{19}$$

Taking expectations conditional on X:

$$E[\hat{\beta}^{ols}|X] = \beta + (X'X)^{-1}X'E[\epsilon|X] \tag{20}$$

If $E[\epsilon|X] \neq 0$, then $E[\hat{\beta}^{ols}] \neq \beta$. The estimator is **biased**.

13.2 2. Inconsistency (Asymptotic Property)

Even with infinite data, OLS does not recover the truth.

$$\text{plim } \hat{\beta}^{ols} = \beta + \text{plim } \left(\frac{1}{N}X'X\right)^{-1} \text{plim } \left(\frac{1}{N}X'\epsilon\right) \tag{21}$$

If x_i and ϵ_i are correlated, plim $(\frac{1}{N}X'\epsilon) \neq 0$. Thus, the estimator is **inconsistent**.

14 III. Instrumental Variables (IV)

How do we fix this? We need an exogenous source of variation to clean up x_i. We call this an **Instrumental Variable (z_i)**.

14.1 1. Validity Conditions

For a variable z_i to be a valid instrument, it must satisfy two strict conditions:

1. **Relevance (First Stage):** The instrument must be correlated with the endogenous regressor.

$$Cov(z_i, x_i) \neq 0 \tag{22}$$

This can be tested empirically.

2. **Exclusion Restriction:** The instrument must be uncorrelated with the error term.

$$Cov(z_i, \epsilon_i) = 0 \tag{23}$$

This implies z_i affects y_i **only through** x_i, not directly. *Note: This condition cannot be tested; it must be argued via economic theory or institutional knowledge.*

14.2 2. The "Art" of IV: Example

Finding a valid IV is an art. Consider the classic **Angrist & Krueger (1991)** study on returns to education.

$$\ln(Wage_i) = \beta_0 + \beta_1 Edu_i + \epsilon_i \tag{24}$$

Edu_i is endogenous (ability bias). They used **Quarter of Birth** as an instrument. The logic is that compulsory schooling laws force those born earlier in the year to stay in school slightly less than those born later, but birth quarter is uncorrelated with innate ability (the error term).

Visualizing the IV Logic (DAG): Imagine a path of causality:

$$\text{Base Station } (Z) \rightarrow \text{Mobile Internet Coverage } (X) \rightarrow \text{Economic Outcome } (Y)$$

Here, the construction of Base Stations (Z) is the instrument. It drives internet coverage (X). We argue that Base Stations affect the outcome (Y) *only* through the internet coverage they provide (Exclusion). If Base Stations were built specifically in rich areas, this exclusion restriction would fail (Confrounders).

15 IV. Two-Stage Least Squares (2SLS)

The standard method to implement IV is 2SLS. As the name suggests, we do this in two steps.

15.1 The Procedure

Step 1 (First Stage): Isolate the exogenous variation in x_i. We regress the endogenous x_i on the instrument z_i (and other exogenous covariates).

$$x_i = \delta_0 + \delta_1 z_i + u_i \tag{25}$$

We obtain the predicted values $\hat{x}_i$:

$$\hat{x}_i = \hat{\delta}_0 + \hat{\delta}_1 z_i \tag{26}$$

Because z_i is uncorrelated with ϵ_i, the predicted part $\hat{x}_i$ is also uncorrelated with ϵ_i. It is the "clean" version of x.

Step 2 (Second Stage): Replace x_i with $\hat{x}_i$ in the original equation.

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \epsilon_i \tag{27}$$

$$= \beta_0 + \beta_1(\hat{\delta}_0 + \hat{\delta}_1 z_i) + \epsilon_i \tag{28}$$

15.2 The IV Estimator Formula

In the simple case with one endogenous regressor and one instrument, the IV estimator is given by the ratio of covariances:

$$\hat{\beta}^{IV} = \frac{Cov(z_i, y_i)}{Cov(z_i, x_i)} \tag{29}$$

Economic Intuition of the Formula: The numerator captures how the instrument affects the outcome (Reduced Form). The denominator captures how the instrument affects the endogenous variable (First Stage). By dividing them, we scale the effect to find the impact of a unit change in x on y, driven purely by the instrument.

$\hat{\beta}^{IV}$: biased., Consistent, asymptotically normal

$\hat{\beta}^{IV} = \beta + \frac{\Sigma(z_i - \bar{z})u_i}{\Sigma(z_i - \bar{z})x_i}$ 、正不能写到分母上.

用 plim 和 Slutsky Thm.

$plim \frac{N}{D} = \frac{plim N}{plim D}$ → $E[x_i u_i] = 0$.

→ LLN.

$A.Var(\hat{\beta}^{2SLS}) = \frac{Avar(\hat{\beta}^{OLS})}{R^2_{x,z}}$ — 比 OLS 大.

Chapter 6: The Generalized Method of Moments (GMM)

Abstract

These notes provide a rigorous yet intuitive overview of the Generalized Method of Moments (GMM). We begin by establishing GMM as a unifying framework that encompasses OLS, IV, 2SLS, and System methods. We explore the geometry of the objective function, the necessity of the weighting matrix, and derive the asymptotic properties of the estimators. Special attention is paid to Simultaneous Equation Systems (SES), 3SLS, and the "J-test" for over-identification.

16 Overview: The Unifying Power of Moments

Welcome. In previous courses, you learned OLS and Maximum Likelihood as separate tools. Today, we unify them. The Generalized Method of Moments (GMM) is not just an estimator; it is a framework. Whether you are dealing with simple linear regression or complex structural models, the fundamental logic remains the same: **Nature provides us with population moment conditions; we try to match them with sample analogues.**

16.1 Moments as Models

In econometrics, a "model" is essentially a set of assumptions about population moments. Consider the linear model:

$$y_i = x_i\beta + \epsilon_i \tag{30}$$

where x_i is a $1 \times K$ row vector. The crucial assumption identifying β is the orthogonality condition:

$$\mathbb{E}[x_i'\epsilon_i] = 0 \tag{31}$$

This states that the regressors are uncorrelated with the error term.

16.2 The General Form

Let us generalize. Let θ be an $M \times 1$ parameter vector. We define a moment function vector $\psi(w_i;\theta)$, where w_i represents our data (y_i, x_i, z_i). The population moment condition is:

A model: $$\mathbb{E}[\psi(w_i;\theta)] = 0 \tag{32}$$

This is a system of M equations. Our goal is to find θ such that this holds.

Since we do not observe the population, we rely on the **Sample Analogue**. By the Weak Law of Large Numbers (WLLN), the sample mean converges to the population mean. Thus, we look at:

$$\psi_N(\theta) = \frac{1}{N}\sum_{i=1}^{N}\psi(w_i;\theta) \tag{33}$$

and we want to set $\psi_N(\theta) = 0$.

17 Identification Strategies

The solvability of $\psi_N(\theta) = 0$ depends on the relationship between the number of moment conditions (M) and the number of parameters to estimate (K).

17.1 Just-Identification ($M = K$)

If the number of equations equals the number of unknowns (e.g., standard OLS where we have K regressors and K orthogonality conditions), there is a unique solution. For the OLS case, the condition $\frac{1}{N}\sum x_i'(y_i - x_i\beta) = 0$ leads directly to:

$$\hat{\beta}^{MM} = (X'X)^{-1}X'y \tag{34}$$

This is the classical Method of Moments estimator.

17.2 Over-Identification ($M > K$)

This is the heart of GMM. Suppose we have more information than parameters. For example, we might have multiple instruments for a single endogenous variable.

$$\mathbb{E}\begin{bmatrix} x_i'\epsilon_i \\ x_i'\epsilon_i^3 \end{bmatrix} = 0 \tag{35}$$

Here, we have a system of equations that is likely inconsistent in finite samples—we cannot set all sample moments to exactly zero simultaneously.

Instead of finding a root, we minimize the "distance" of the moments from zero. We define a **Loss Function** $Q_N(\theta)$:

$$Q_N(\theta) = \psi_N(\theta)'C\psi_N(\theta) \tag{36}$$

where C is a symmetric, positive-definite weighting matrix. The GMM estimator is defined as:

$$\hat{\theta}^{GMM} = \arg\min_{\theta} Q_N(\theta) \tag{37}$$

This minimization tries to satisfy the moment conditions as closely as possible.

18 Deriving the GMM Estimator

18.1 The Geometry of the Weighting Matrix C

Why do we need C? Intuitively, not all moment conditions are created equal. Some moments might have high variance (they are "noisy" rulers), while others are precise.

- **Shift-share IV example:** Think of different instruments as measuring tools. Some are precise; some are not. C assigns weights to these tools.
- If we define $C = \mathbf{I}$, we treat all moments equally (Identity Matrix).

Inverse Variance Weighting : WLS. GMM.

- If we define C to be the inverse of the covariance matrix of the moments, we down-weight noisy moments.

Consider a simplified linear case where $y_i = \theta + \epsilon_i$. We have two moments (conditions) for θ: x_i and y_i both target θ. Let the objective function be:

$$Q_N(\theta) = \begin{bmatrix} \bar{x} - \theta \\ \bar{y} - \theta \end{bmatrix}' \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} \bar{x} - \theta \\ \bar{y} - \theta \end{bmatrix} \tag{38}$$

Minimizing this quadratic form yields a weighted average:

$$\hat{\theta}^{GMM} = \lambda \bar{x} + (1 - \lambda)\bar{y} \tag{39}$$

where λ is determined by the entries of C.

To maximize efficiency (minimize variance), we choose λ^* such that:

$$\lambda^* = \frac{\sigma_{yy} - \sigma_{xy}}{\sigma_{xx} - 2\sigma_{xy} + \sigma_{yy}} \tag{40}$$

This implies that the optimal weighting matrix C relates to the inverse of the variance of the moments.

18.2 Two-Step GMM

Since the optimal weight matrix $C \propto \text{Var}(\psi_i)^{-1}$ depends on the unknown parameters (via the residuals), we cannot calculate it immediately. We use a **Two-Step Procedure**:

1. **Step 1:** Estimate $\hat{\theta}^{(1)}$ using a suboptimal but valid matrix (usually $C = I$ or $(Z'Z)^{-1}$).
2. **Calculate Residuals:** Use $\hat{\theta}^{(1)}$ to form residuals and estimate the moment covariance matrix $\hat{S}$.
3. **Step 2:** Set $\hat{C}^{(2)} = \hat{S}^{-1}$. Re-minimize $Q_N(\theta)$ to get $\hat{\theta}^{(2)}$.

19 IV and 2SLS as GMM

Let's look at Instrumental Variables (IV) through the GMM lens. Model: $y = X\beta + \epsilon$ where $\mathbb{E}[X'\epsilon] \neq 0$ (Endogeneity). Instruments: Z such that $\mathbb{E}[Z'\epsilon] = 0$.

19.1 Just-Identified Case (IV)

If the dimension of Z equals the dimension of X ($M = K$), the sample moment is $\frac{1}{N}Z'(y - X\beta) = 0$. Solving for β:

$$\hat{\beta}^{IV} = (Z'X)^{-1}Z'y \tag{41}$$

Note that $(Z'X)^{-1} = (X'Z)^{-1}$ exists because dimensions match.

19.2 Over-Identified Case (2SLS)

If $M > K$, we cannot set $Z'(y - X\beta)$ to exactly zero. The GMM objective function with $C = (Z'Z)^{-1}$ is:

$$Q_N(\beta) = (y - X\beta)'Z(Z'Z)^{-1}Z'(y - X\beta) \tag{42}$$

Recognize that $P_Z = Z(Z'Z)^{-1}Z'$ is the projection matrix. The objective function becomes minimizing the projected residuals:

$$Q_N(\beta) = (y - X\beta)'P_Z(y - X\beta) \tag{43}$$

The First Order Condition (FOC) with respect to β:

$$-2X'P_Z(y - X\beta) = 0 \implies X'Z(Z'Z)^{-1}Z'y = X'Z(Z'Z)^{-1}Z'X\beta \tag{44}$$

Solving this yields the familiar 2SLS estimator:

$$\hat{\beta}^{2SLS} = (X'P_ZX)^{-1}X'P_Zy = (\hat{X}'\hat{X})^{-1}\hat{X}'y \tag{45}$$

where $\hat{X} = P_ZX$.

19.3 Properties of 2SLS

1. **Biased:** $\mathbb{E}[\hat{\beta}^{2SLS}] \neq \beta$ in finite samples. Why? Because $\hat{X}$ includes estimated parameters from the first stage, and $\mathbb{E}[\epsilon|\hat{X}] \neq 0$ necessarily.
2. **Consistent:** As $N \to \infty$, $\hat{\beta}^{2SLS} \xrightarrow{p} \beta$. The bias vanishes asymptotically.

20 Simultaneous Equation Systems and System GMM

We now extend GMM to systems of equations (Simultaneous Equations Models - SEM) or Panel Data settings. Consider a system of G equations for individual i:

$$y_{gi} = x_{gi}\beta_g + \epsilon_{gi}, \quad g = 1, \ldots, G \tag{46}$$

20.1 Stacked Notation

To handle this efficiently, we use block-diagonal matrix notation. Let:

- $y_i = [y_{1i}, \ldots, y_{Gi}]'$ $(G \times 1)$
- $X_i = \text{diag}(x_{1i}, \ldots, x_{Gi})$ (Block diagonal matrix of regressors)
- $Z_i = \text{diag}(z_{1i}, \ldots, z_{Gi})$ (Instruments for each equation)

The moment condition for the system is:

$$\mathbb{E}[Z_i'\epsilon_i] = \mathbb{E}[Z_i'(y_i - X_i\beta)] = 0 \tag{47}$$

The variance of the moment conditions is given by the matrix Ω:

$$\Omega = \mathbb{E}[Z_i' \epsilon_i \epsilon_i' Z_i] = \mathbb{E}[Z_i' \Sigma Z_i] \tag{48}$$

where $\Sigma = \mathbb{E}[\epsilon_i \epsilon_i' | Z_i]$ is the $G \times G$ covariance matrix of errors across equations.

Assuming homoskedasticity conditional on Z, we can write $\Omega = Z'(\Sigma \otimes I_N)Z$ (in full matrix notation). The **System GMM (SGMM)** estimator utilizes the optimal weight matrix $C_N = \Omega^{-1}$:

$$\hat{\beta}^{SGMM} = (X'ZC_NZ'X)^{-1}X'ZC_NZ'y \tag{49}$$

21 Three-Stage Least Squares (3SLS)

Historically, 3SLS (developed in the 1960s) preceded the formal GMM framework (1980s), but it is essentially a specific case of System GMM.

21.1 The Procedure

3SLS combines 2SLS with Seemingly Unrelated Regressions (SURE).

1. **Stage 1:** Estimate the reduced form equations $X = Z\delta + u$ to get $\hat{X}$.
2. **Stage 2:** Estimate the error covariance matrix Σ (the correlations between equations) using 2SLS residuals.
3. **Stage 3:** Apply GLS (Generalized Least Squares) using $\hat{\Sigma}$.

The estimator is given by:

$$\hat{\beta}^{3SLS} = (\hat{X}'(\Sigma^{-1} \otimes I)\hat{X})^{-1}\hat{X}'(\Sigma^{-1} \otimes I)y \tag{50}$$

21.2 Consistency Requirements

For 3SLS to be consistent, a very strong condition is required: **The instruments must be valid for the entire system.** Specifically, $\mathbb{E}[z_i' \Sigma^{-1} \epsilon_i] = 0$. This implies that an instrument used in equation g must be uncorrelated with the error term in equation h. If one equation is misspecified, the contamination can spread to estimates in other equations via the Σ^{-1} matrix.

22 Hypothesis Testing in GMM

The GMM framework provides natural tests for model validity.

IVs' joint validity.

22.1 The Hansen J-Test (Over-identifying Restrictions)

If $M > K$, the model is over-identified. We can test if the moment conditions are mutually consistent. The test statistic is simply the minimized value of the objective function (scaled):

$$J = N \cdot Q_N(\hat{\theta}^{GMM}) \xrightarrow{d} \chi^2_{M-K} \tag{51}$$

If J is large, the moments are "far" from zero, suggesting the model specification or instruments are invalid. *Note: This requires the use of the optimal weighting matrix.*

22.2 The Sargan-Hansen C-Test (Difference-in-Sargan)

Suppose we suspect a subset of instruments might be invalid. We can partition the moments into M (all) and M_A (valid subset). We calculate two J-statistics:

1. J using all instruments.
2. J_A using only the valid subset.

The difference statistic is:

$$C = J - J_A \xrightarrow{d} \chi^2_{M-M_A} \tag{52}$$

This tests the validity of the suspect subset of instruments.

23 Conclusion

GMM provides a flexible, powerful way to think about estimation. By defining the appropriate moment conditions ($\mathbb{E}[Z'\epsilon] = 0$) and weighting matrix ($C$), we can derive efficient estimators for a vast array of economic models. While OLS and 2SLS are just special cases, the System GMM allows us to handle complex interdependencies in panel data and simultaneous equations.

Endogeneity.

	homoskedasticity	hetero
Single eq.	2SLS	GMM
System of eqs.	3SLS	System GMM

Chapter 7: Maximum Likelihood Estimation (MLE)

24 Introduction: The Philosophy of MLE

Welcome. Up to this point, we have largely relied on the method of least squares. While OLS is intuitive and robust (giving us the Best Linear Unbiased Estimator under Gauss-Markov assumptions), it has limitations when the Data Generating Process (DGP) is complex—such as when dealing with binary choices, counts, or truncated data.

Today, we transition to a more general framework: **Maximum Likelihood Estimation (MLE)**. The philosophy here is fundamentally different. In OLS, we minimize the mistakes (residuals). In MLE, we ask a reverse-engineering question: *Given the data we observed, what parameters would have made this data the most probable outcome?*.

We assume a distribution—a probabilistic engine generating the data—and we tune the knobs (parameters) of that engine until the data generated matches our observation as closely as possible.

25 I. Intuition: The Discrete Case

Let's start with the simplest possible example to build intuition. Imagine an urn containing balls.

25.1 The Urn Experiment

Suppose we have a population of balls, either Red or White. We are interested in one parameter: the proportion of red balls, denoted by p.

- **Data (N):** We draw N balls.
- **Observation (N_1):** We observe N_1 red balls and $N - N_1$ white balls.

This is the classic Bernoulli trial setup leading to a Binomial distribution. The probability of observing exactly N_1 red balls given a parameter p is:

$$P(N_{red} = N_1|p) = C_N^{N_1} p^{N_1}(1-p)^{N-N_1} \tag{53}$$

where $C_N^{N_1}$ is the combinatorial coefficient.

25.2 The Likelihood Maximization

In probability theory, we know p and calculate the chance of data N_1. In econometrics, we observe N_1 and want to find $\hat{p}$.

$$\hat{p} = \arg\max_p P(N_{red}(p) = N_1)$$

To solve this, we maximize the log-likelihood (since log is a monotonic transformation, it preserves the maximum location and turns products into sums, making derivatives easier).

$$\ln \mathcal{L}(p) = \ln \left(C_N^{N_1} \right) + N_1 \ln p + (N - N_1) \ln(1 - p)$$

Taking the First Order Condition (F.O.C) with respect to p:

$$\frac{d \ln p}{dp} = \frac{N_1}{p} - \frac{N - N_1}{1 - p} \equiv 0 \tag{54}$$

Solving for p, we get the MLE estimator:

$$\hat{p}^{MLE} = \frac{N_1}{N} \tag{55}$$

.

Intuition: The estimator is simply the sample proportion. This matches our "Classical" intuition, but we arrived here by formally maximizing the probability of the observed sample space. Note that we must be careful with the definition of the probability space to avoid logical traps like *Bertrand's Paradox*, ensuring our density is well-defined.

26 II. The Continuous Case: Linear Regression

Now, let's apply this to the bread and butter of economics: the linear regression model.

$$y_i = \beta_0 + x_i \beta_1 + \epsilon_i$$

. Note that here x_i represents the row vector of regressors for observation i.

26.1 Assumptions and Density

We assume the errors are independent and identically distributed (i.i.d.) normal: $\epsilon_i \sim N(0, \sigma^2)$. The probability density function (PDF) for a single observation y_i conditional on x_i is:

$$f(y_i | x_i; \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y_i - \beta_0 - x_i \beta_1)^2}{2\sigma^2} \right\} \tag{56}$$

.

26.2 The Log-Likelihood Function

Since observations are independent, the joint density is the product of individual densities. The log-likelihood is the sum of log-densities:

$$\ln \mathcal{L}(\beta, \sigma^2) = \sum_{i=1}^{N} \left[-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(y_i - \beta x_i)^2}{2\sigma^2} \right] \tag{57}$$

.

26.3 Estimating β

To find $\hat{\beta}^{MLE}$, we maximize $\ln \mathcal{L}$. Notice that the term $-\frac{1}{2}\ln(2\pi\sigma^2)$ does not depend on β. We are left with maximizing:

$$-\sum(y_i - x_i\beta)^2$$

Maximizing the negative sum of squared errors is identical to **minimizing the sum of squared errors**. Therefore:

$$\hat{\beta}^{MLE} \equiv \hat{\beta}^{OLS} \tag{58}$$

. Under normality assumptions, MLE and OLS coefficients are identical.

26.4 Estimating σ^2: A Tale of Bias

Now, differentiate with respect to σ^2 (treat it as a single parameter):

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{\sum \hat{\epsilon}_i^2}{2\sigma^4} \equiv 0 \tag{59}$$

. Solving this yields:

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} \hat{\epsilon}_i^2 \tag{60}$$

Crucial Observation: Recall that the OLS unbiased estimator for variance is $s^2 = \frac{1}{N-k}\sum \hat{\epsilon}^2$ (where k is the number of parameters). The MLE estimator divides by N, not $N-k$.

- $\hat{\sigma}^2_{MLE}$ is **biased** in finite samples (it underestimates variance).
- However, as $N \to \infty$, $1/N \approx 1/(N-k)$, so it is **consistent**.

Generally, MLE estimators are not necessarily unbiased, but we prize them for their asymptotic properties (consistency and efficiency).

27 III. Formal Framework and Identification

Let's formalize the structure.

27.1 Likelihood and Parameter Space

We define a parameter space Θ. The likelihood function is a mapping $L_N : \Theta \times \mathcal{S} \to \mathbb{R}$.

$$L_N(\theta; \{x_i\}) = \prod_i f_x(x_i; \theta)$$

And the log-likelihood:

$$\mathcal{L}_N(\theta) = \sum_i \ln f_x(x_i; \theta)$$

.

MLE 就是爬山.

27.2 Identification

A model is **identified** if distinct parameters yield distinct likelihood values (in expectation).

$$\text{If } \theta \neq \theta_0 \implies \mathcal{L}_N(\theta, y) \neq \mathcal{L}_N(\theta_0, y)$$

. Without identification, we cannot recover the unique true parameters from the data.

27.3 Example: Probit Model (Binary Choice)

DGP.

Consider a latent variable model, often used in labor economics or utility theory:

$$y_i^* = x_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0,1)$$

We do not observe y_i^*. We observe a binary choice y_i:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

. This is the **Probit Model**.

The probability of observing $y_i = 1$ is:

$$\begin{aligned} P(y_i = 1|x_i) &= P(y_i^* > 0|x_i) = P(x_i\beta + \epsilon_i > 0) \\ &= P(\epsilon_i > -x_i\beta) \\ &= P(\epsilon_i < x_i\beta) \quad \text{(by symmetry of Normal dist.)} \\ &= \Phi(x_i\beta) \end{aligned}$$

. Here, Φ is the Cumulative Distribution Function (CDF) of the standard normal. This illustrates how MLE allows us to estimate parameters even when the dependent variable is not continuous, by mapping the latent utility to a probability space.

28 IV. Optimization: Score and Hessian

Finding the maximum involves calculus on the likelihood surface.

28.1 The Score Vector (Gradient)

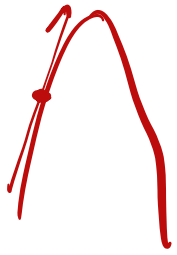
Notice! 不是 $\mathcal{L}$, 是 $\ell = \ln \mathcal{L}$.

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

$$S_N(\theta; X) \equiv \frac{\partial \mathcal{L}_N(\theta; X)}{\partial \theta} \tag{61}$$

. At the maximum likelihood estimate $\hat{\theta}_{MLE}$, the score must be zero (F.O.C.):

$$\left.\frac{\partial \mathcal{L}_N}{\partial \theta}\right|_{\hat{\theta}_{MLE}} = 0$$



.

28.2 The Hessian Matrix (Curvature)

The **Hessian** is the matrix of second derivatives:

$$H_N(\theta; X) \equiv \frac{\partial^2 \mathcal{L}_N(\theta; X)}{\partial \theta \partial \theta'} \tag{62}$$

. For $\hat{\theta}$ to be a maximum, the Hessian must be **negative definite** (the surface must be concave shaped like a hill, not a valley).

Example: Normal Mean and Variance For the Normal distribution $N(\mu, \sigma^2)$, the log-likelihood is:

$$\mathcal{L} = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum(y_i - \mu)^2$$

. The Hessian with respect to μ is:

$$\frac{\partial^2 \mathcal{L}}{\partial \mu^2} = -\frac{N}{\sigma^2} < 0$$

And with respect to σ^2:

$$\frac{\partial^2 \mathcal{L}}{\partial (\sigma^2)^2} = \frac{N}{2\sigma^4} - \frac{\sum(y_i - \mu)^2}{\sigma^6}$$

Evaluated at the optimum $\hat{\sigma}^2$, this simplifies to $-\frac{N}{2\sigma^4} < 0$. Since the diagonal elements are negative and the determinant conditions hold, the function is concave, ensuring a unique maximum.

29 V. Properties of MLE: Invariance

A powerful property of MLE is **Invariance**. If $\hat{\theta}$ is the MLE for θ, then for any one-to-one function $g(\cdot)$, the MLE for $g(\theta)$ is simply $g(\hat{\theta})$.

$$\widehat{g(\theta)}^{MLE} \equiv g(\hat{\theta}^{MLE}) \tag{63}$$

.

Example: Precision Parameter If we estimate variance σ^2, but we care about precision $\lambda^2 = 1/\sigma^2$. We do not need to re-derive the likelihood for λ. We simply compute:

$$\hat{\lambda}^2 = \frac{1}{\hat{\sigma}^2_{MLE}}$$

.

30 VI. The Trinity: Score, Hessian, and Fisher Information

This section contains the theoretical core of MLE inference. We need to understand the statistical properties of the Score and Hessian to derive standard errors.

30.1 Lemma 1: The Score has Zero Expectation

$$\mathbb{E}[S_N(\theta_0; X)] = 0 \tag{64}$$

. **Proof:**

$$\mathbb{E}[S] = \int \frac{\partial \ln f(x)}{\partial \theta} f(x) dx = \int \frac{1}{f(x)} \frac{\partial f(x)}{\partial \theta} f(x) dx = \int \frac{\partial f(x)}{\partial \theta} dx$$

By Leibniz rule (interchanging integration and differentiation), this equals

$$\frac{\partial}{\partial \theta} \int f(x) dx = \frac{\partial}{\partial \theta}(1) = 0$$

. This relies on the "regularity conditions" (bounds of integration do not depend on θ).

30.2 Fisher Information Matrix

The Fisher Information $\mathcal{I}(\theta)$ tells us how much information the data contains about the parameter. It is defined as the variance of the Score:

$$\mathcal{I}_N(\theta) = \text{Var}[S_N(\theta; X)] = \mathbb{E}[S_N S_N'] \tag{65}$$

(Since mean score is 0).

30.3 Lemma 2: Information Equality

Under regularity conditions, the Fisher Information is also the expected value of the negative Hessian:

$$\mathcal{I}_N(\theta_0) = -\mathbb{E}[H_N(\theta_0; X)] \tag{66}$$

. **Intuition:** The Hessian measures curvature. A very "pointy" likelihood peak (high negative curvature) means the variance of the estimator is low (we are very sure about the parameter). Thus, high curvature = high information.

Proof Sketch: Differentiate the identity $\int f(x; \theta) dx = 1$ twice. First derivative gave $\mathbb{E}[S] = 0$. Differentiating again yields the relationship:

$$\int \frac{\partial^2 \ln f}{\partial \theta \partial \theta'} f dx + \int \left(\frac{\partial \ln f}{\partial \theta}\right)\left(\frac{\partial \ln f}{\partial \theta'}\right) f dx = 0$$

$$\mathbb{E}[H] + \mathbb{E}[SS'] = 0 \implies \mathcal{I}(\theta) = -\mathbb{E}[H]$$

For i.i.d samples, information is additive: $\mathcal{I}_N(\theta) = N \times \mathcal{I}_1(\theta)$.

31 VII. Asymptotic Properties

We rarely know the finite-sample distribution of MLE estimates (except in simple cases like the Normal). Instead, we rely on large-sample (asymptotic) theory.

31.1 1. Consistency

Does $\hat{\theta} \xrightarrow{p} \theta_0$ as $N \to \infty$?

To prove this, we view MLE as an **M-Estimator**. We maximize an objective function $Q_N(\theta) = \frac{1}{N}\mathcal{L}_N(\theta)$. As $N \to \infty$, $Q_N(\theta)$ converges to its expectation $Q^*(\theta) = \mathbb{E}[\ln f(x;\theta)]$.

Kullback-Leibler Inequality: By Jensen's Inequality, the expected log-likelihood is maximized at the true parameter θ_0:

$$\mathbb{E}\left[\ln \frac{f(x;\tilde{\theta})}{f(x;\theta_0)}\right] \leq \ln \mathbb{E}\left[\frac{f(x;\tilde{\theta})}{f(x;\theta_0)}\right] = \ln(1) = 0$$

Thus $\mathbb{E}[\ln f(x;\tilde{\theta})] \leq \mathbb{E}[\ln f(x;\theta_0)]$. This ensures that the population objective function $Q^*(\theta)$ has a unique maximum at θ_0.

Conditions for Consistency:

1. **Compactness:** Parameter space Θ is compact.
2. **Continuity:** $\ln f(x;\theta)$ is continuous.
3. **Uniform Convergence:** $Q_N(\theta)$ converges uniformly to $Q^*(\theta)$.

If these hold, $\hat{\theta}_{MLE}$ converges in probability to θ_0.

31.2 2. Asymptotic Normality

This is the result we use for hypothesis testing (t-stats).

$$\sqrt{N}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, V) \tag{67}$$

We derive this using a **Mean Value Expansion** (Taylor Series) of the Score function around the true parameter θ_0.

$$S_N(\hat{\theta}) = S_N(\theta_0) + H_N(\tilde{\theta})(\hat{\theta} - \theta_0)$$

where $\tilde{\theta}$ is between $\hat{\theta}$ and θ_0. Since $\hat{\theta}$ is the maximizer, $S_N(\hat{\theta}) = 0$. Thus:

$$\sqrt{N}(\hat{\theta} - \theta_0) = -\left[\frac{1}{N}H_N(\tilde{\theta})\right]^{-1}\left[\frac{1}{\sqrt{N}}S_N(\theta_0)\right] \tag{68}$$

.

Now apply our limit theorems:

1. **CLT on Score:** $\frac{1}{\sqrt{N}}S_N(\theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0))$ (Since $\mathbb{E}[S] = 0, \text{Var}[S] = \mathcal{I}$).
2. **WLLN on Hessian:** $\frac{1}{N}H_N(\tilde{\theta}) \xrightarrow{p} \mathbb{E}[H(\theta_0)] = -\mathcal{I}(\theta_0)$ (By Information Equality and Consistency).

Combining these using Slutsky's Theorem:

$$\text{Asy. Var} = [-\mathcal{I}]^{-1}\mathcal{I}[-\mathcal{I}]^{-1} = \mathcal{I}(\theta_0)^{-1}$$

.

Result: The MLE is asymptotically normal with variance equal to the inverse of the Fisher Information.

$$\hat{\theta}_{MLE} \overset{a}{\sim} N\left(\theta_0, \frac{1}{N}\mathcal{I}^{-1}(\theta_0)\right)$$

. This implies that the "sharper" the curvature of the likelihood function (larger $\mathcal{I}$), the smaller the variance of our estimator.

32 Summary

In this chapter, we have:

1. Defined MLE as maximizing the probability of observing the sample data.
2. Shown it encompasses OLS in linear normal models but extends to broader applications (Probit).
3. Established the "Trinity" of MLE inference: Score (zero mean), Hessian (negative definite), and Fisher Information (variance of score).
4. Proved that under regularity conditions, MLE is Consistent and Asymptotically Normal with variance $\mathcal{I}^{-1}$.

Chapter 8: Discrete Choice Models

33 Introduction

Welcome to the fascinating world of Microeconometrics. Up to this point, we have primarily dealt with continuous dependent variables. However, in the real world, many economic outcomes are discrete. An individual decides whether to work or not ($y = 0, 1$), whether to buy a car or not, or which mode of transportation to take to work (bus, car, train).

In this chapter, we focus on models where the dependent variable y_i takes on a finite number of values. We start with the fundamental binary choice case:

$$y_i = \begin{cases} 1 & \text{if the event occurs} \\ 0 & \text{otherwise} \end{cases}$$

34 The Linear Probability Model (LPM)

Let us begin with the simplest possible approach. Why not simply run an OLS regression? This is known as the Linear Probability Model.

34.1 Model Setup

Consider the standard linear specification:

$$y_i = x_i\beta + \epsilon_i \tag{69}$$

where x_i is a $1 \times K$ **row vector** of regressors, and β is a $K \times 1$ vector of coefficients.

What is the interpretation of the conditional expectation here? Since y_i is Bernoulli distributed (taking values 0 or 1), its expectation is simply the probability of success:

$$\begin{aligned} \mathbb{E}[y_i|x_i] &= 1 \cdot \text{Prob}(y_i = 1|x_i) + 0 \cdot \text{Prob}(y_i = 0|x_i) \\ &= \text{Prob}(y_i = 1|x_i) \end{aligned}$$

Assuming zero conditional mean of the error term $\mathbb{E}[\epsilon_i|x_i] = 0$, we have:

$$\text{Prob}(y_i = 1|x_i) = x_i\beta$$

Consequently, the probability of failure is $1 - x_i\beta$.

34.2 Advantages and Disadvantages

Why do we still use the LPM?

1. **Simplicity:** It is computationally trivial and coefficients are directly interpretable as marginal effects.

2. **Fixed Effects:** It is very easy to include fixed effects (e.g., individual or time dummies) to account for unobserved heterogeneity. This is computationally difficult in non-linear models like Logit or Probit (the incidental parameters problem).

However, there are fundamental flaws:

1. **Unbounded Probabilities:** The predicted values $\hat{y}_i = x_i\hat{\beta}$ can be less than 0 or greater than 1, which makes no sense as probabilities.
2. **Heteroskedasticity:** The error term ϵ_i is inherently heteroskedastic.

$$\begin{aligned}\text{Var}(\epsilon_i|x_i) &= \text{Var}(y_i|x_i) \\ &= \mathbb{E}[y_i^2|x_i] - (\mathbb{E}[y_i|x_i])^2 \\ &= x_i\beta - (x_i\beta)^2 \\ &= x_i\beta(1 - x_i\beta)\end{aligned}$$

Since the variance depends on x_i, we must use **Heteroskedasticity-Robust Standard Errors** when using LPM.

35 Latent Variable Models and CDFs

To solve the problem of unbounded probabilities, we need a function $F(\cdot)$ that maps the linear index $x_i\beta$ into the interval $(0, 1)$.

$$\text{Prob}(y_i = 1|x_i) = F(x_i\beta)$$

Usually, $F(\cdot)$ is chosen to be a Cumulative Distribution Function (CDF).

35.1 Micro-foundation: The Utility Maximization Framework

Where does this come from? It's not just mathematical convenience; it has a deep economic structure based on utility maximization.

Assume there is an unobserved, **latent variable** y_i^* representing the net utility or propensity to take an action:

$$y_i^* = x_i\beta + \epsilon_i \tag{70}$$

We do not observe y_i^*, but we observe the decision y_i based on a threshold (usually normalized to 0):

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Now we can derive the probability:

$$\begin{aligned}\text{Prob}(y_i = 1|x_i) &= \text{Prob}(y_i^* > 0|x_i) \\ &= \text{Prob}(x_i\beta + \epsilon_i > 0|x_i) \\ &= \text{Prob}(\epsilon_i > -x_i\beta|x_i)\end{aligned}$$

If the distribution of ϵ_i is symmetric around 0 (which is true for Normal and Logistic distributions), then $\text{Prob}(\epsilon_i > -z) = \text{Prob}(\epsilon_i < z)$. Thus:

$$\text{Prob}(y_i = 1|x_i) = F_\epsilon(x_i\beta)$$

where F_ϵ is the CDF of the error term.

35.2 Probit and Logit

The choice of the error distribution dictates the model:

1. **Probit Model:** Assume $\epsilon_i \sim N(0, 1)$ (Standard Normal).

$$F(x_i\beta) = \Phi(x_i\beta) = \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

2. **Logit Model:** Assume ϵ_i follows a Logistic distribution.

$$F(x_i\beta) = \Lambda(x_i\beta) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}$$

   The Logistic distribution has "fatter tails" than the Normal distribution, but empirically the results are often very similar (after scaling).

Note on Identification: We assume $\text{Var}(\epsilon_i) = 1$ (for Probit) or $\pi^2/3$ (for Logit). We cannot estimate the variance of ϵ_i because it is not identified separately from β (if we double β and σ, the probability remains the same). 我们真正估计的是 $\frac{\beta}{\sigma}$. 人为规定 $\sigma^2 = \begin{cases} 1 & \text{for probit} \\ \pi^2/3 & \text{for logit} \end{cases}$ 确定分母.

36 Marginal Effects Delta Method !

In the linear model, $\beta_k = \frac{\partial \mathbb{E}[y|x]}{\partial x_k}$. This is constant for everyone. In non-linear models, the effect of a change in x depends on where you start (the current level of $x_i\beta$).

The Marginal Effect (ME) for a continuous regressor x_{ik} is:

$$ME_{ik} = \frac{\partial \text{Prob}(y_i = 1|x_i)}{\partial x_{ik}} = \frac{\partial F(x_i\beta)}{\partial (x_i\beta)} \cdot \frac{\partial (x_i\beta)}{\partial x_{ik}} = f(x_i\beta) \cdot \beta_k$$

where $f(\cdot)$ is the Probability Density Function (PDF).

- **Probit:** $ME_{ik} = \phi(x_i\beta)\beta_k$, where $\phi(\cdot)$ is the standard normal PDF.
- **Logit:** $ME_{ik} = \Lambda(x_i\beta)(1 - \Lambda(x_i\beta))\beta_k$.

Intuition: The marginal effect is largest when $f(x_i\beta)$ is largest, which is when $x_i\beta \approx 0$ (i.e., when the probability is close to 0.5). If an individual is already almost certain to choose 1 ($x_i\beta$ is very large), a small change in x won't matter much.

切换临界点 ME最大.

37 Estimation: Maximum Likelihood (MLE)

Since the error terms are not observed, we cannot use OLS. We rely on Maximum Likelihood Estimation.

37.1 The Likelihood Function

The data follows a Bernoulli distribution. The probability mass function for a single observation is:

$$P(y_i|x_i) = [F(x_i\beta)]^{y_i}[1 - F(x_i\beta)]^{1-y_i}$$

The Log-Likelihood function for the sample size N is:

$$\ln \mathcal{L}(\beta) = \sum_{i=1}^{N} [y_i \ln F(x_i\beta) + (1 - y_i) \ln(1 - F(x_i\beta))]$$

37.2 First Order Conditions (Score)

We maximize $\ln \mathcal{L}(\beta)$ with respect to β. The FOC is:

$$\frac{\partial \ln \mathcal{L}}{\partial \beta} = \sum_{i=1}^{N} \left[\frac{y_i}{F_i} f_i - \frac{1 - y_i}{1 - F_i} f_i \right] x_i' = 0$$

where $F_i = F(x_i\beta)$ and $f_i = F'(x_i\beta)$.

37.2.1 Logit Case

For Logit, a beautiful simplification occurs because $f(z) = \Lambda(z)(1 - \Lambda(z))$.

$$\frac{f_i}{F_i(1 - F_i)} = \frac{\Lambda_i(1 - \Lambda_i)}{\Lambda_i(1 - \Lambda_i)} = 1$$

Thus, the FOC simplifies to:

$$\sum_{i=1}^{N} (y_i - \Lambda(x_i\beta))x_i' = 0$$

This looks remarkably like the OLS normal equations (residual $\times$ regressor $= 0$), but it is non-linear in β.

37.2.2 Probit Case

For Probit, we define generalized residuals using the inverse Mills ratio. Let $q_i = 2y_i - 1$.

$$\frac{\partial \ln \mathcal{L}}{\partial \beta} = \sum_{i=1}^{N} \lambda_i x_i' = 0$$

where $\lambda_i = \frac{q_i \phi(q_i x_i \beta)}{\Phi(q_i x_i \beta)}$.

37.3 Hessian and Asymptotic Variance

To ensure we found a maximum and to calculate standard errors, we look at the second derivative (Hessian matrix H). For Logit:

$$H = \frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \beta'} = -\sum_{i=1}^{N} \Lambda_i(1-\Lambda_i) x_i' x_i$$

Notice that $\Lambda_i(1-\Lambda_i)$ is always positive. Thus, the Hessian is negative definite globally. This means the Logit log-likelihood is globally concave, ensuring a unique maximum!

The Asymptotic Variance of $\hat{\beta}$ is given by the inverse of the Information Matrix:

$$A.\text{Var}(\hat{\beta}) = -\mathbb{E}[H]^{-1} = \left(\sum_{i=1}^{N} f_i^2 x_i' (\text{Var}(y_i))^{-1} x_i \right)^{-1}$$

In practice, we use the observed Hessian.

38 Hypothesis Testing

We have three main ways to test hypotheses (e.g., $H_0 : \beta_j = 0$).

1. **Wald Test:** Uses the estimate $\hat{\beta}$ and its standard error. It asks: "Is $\hat{\beta}$ far from 0?"
$$W = (\hat{\beta} - \beta_0)' [\text{Var}(\hat{\beta})]^{-1} (\hat{\beta} - \beta_0) \sim \chi^2$$

2. **Likelihood Ratio (LR) Test:** Requires estimating both the Unrestricted model (L_U) and the Restricted model (L_R). It asks: "Does the likelihood drop significantly when we impose the restriction?"
$$LR = -2(\ln L_R - \ln L_U) \sim \chi^2_J$$
where J is the number of restrictions. This is often preferred in non-linear models.

3. **Lagrange Multiplier (LM) Test:** Requires estimating only the Restricted model. It asks: "If we start at the restricted estimate, is the gradient (score) steep?" This is computationally cheapest if the unrestricted model is hard to estimate.

39 Endogeneity in Binary Choice Models

This is a critical topic. What if x_i is correlated with ϵ_i? In OLS, we use IV (2SLS). In Binary Choice, **2SLS is inconsistent**. We need a structural approach.

39.1 The Problem

Structural equation:

$$y_i^* = x_i\beta + \gamma w_i + \epsilon_i$$

where w_i is endogenous ($Cov(w_i, \epsilon_i) \neq 0$). First stage for w_i:

$$w_i = z_i\alpha + u_i$$

where z_i are exogenous instruments.

Assume (ϵ_i, u_i) follow a **Bivariate Normal (BVN)** distribution:

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_u \\ \rho\sigma_u & \sigma_u^2 \end{pmatrix} \right)$$

39.2 The Control Function Approach (Intuition)

We can decompose the error ϵ_i into a part correlated with u_i and a pure random part. Recall from properties of BVN:

$$\mathbb{E}[\epsilon_i | u_i] = \rho \frac{u_i}{\sigma_u}$$

So we can write $\epsilon_i = \rho \frac{u_i}{\sigma_u} + \nu_i$, where ν_i is independent of u_i (and thus w_i).

Substitute this back into the structural equation:

$$\begin{aligned} y_i^* &= x_i\beta + \gamma w_i + \left(\rho \frac{u_i}{\sigma_u} + \nu_i \right) \\ &= x_i\beta + \gamma w_i + \lambda \hat{u}_i + \nu_i \end{aligned}$$

This suggests a two-step procedure (Control Function):

1. Regress w_i on z_i to get residuals $\hat{u}_i$.
2. Run Probit of y_i on x_i, w_i, **and** $\hat{u}_i$.

Testing if the coefficient on $\hat{u}_i$ is zero is a test for exogeneity!

39.3 Full Information MLE

While the two-step method gives intuition, we usually estimate everything jointly using FIMLE to get correct standard errors. The joint density is $f(y_i, w_i) = f(y_i|w_i)f(w_i)$.

$$\text{Prob}(y_i = 1 | w_i, z_i) = \Phi\left(\frac{x_i\beta + \gamma w_i + (\rho/\sigma_u)u_i}{\sqrt{1-\rho^2}} \right)$$

We maximize the joint likelihood over all parameters.

Chapter 8 & 9

Gemini 3 Pro

Fall 2025

Chapter 8 (cont.) Multinomial Choice Models

1 Introduction

Welcome. Today, we depart from the continuous world of OLS and enter the discrete world of **Multinomial Choice Models**. In many economic situations, agents are not choosing a continuous quantity (like how much rice to buy) but rather making a discrete choice among mutually exclusive alternatives (e.g., Which brand of car to buy? Which mode of transport to take to work?).

We will build the theoretical foundation using the Random Utility Model (RUM), explore the specific distributional assumptions that lead to the Logit family, and rigorously derive the Maximum Likelihood Estimation (MLE) properties.

2 The General Framework

2.1 Random Utility Model (RUM)

Let us consider an individual i facing $J+1$ alternatives, indexed by $j = 0, 1, \ldots, J$. We posit that the utility individual i derives from choice j is composed of a deterministic component (observable to the econometrician) and a random component (unobservable).

$$U_{ij} = z_{ij}\theta + \epsilon_{ij}$$

Here, z_{ij} captures the characteristics affecting utility, and ϵ_{ij} represents the idiosyncratic shock.

The individual acts rationally, choosing the option that maximizes their utility. Therefore, the probability that individual i chooses alternative j is:

$$P(y_i = j|z_{ij}) = P(U_{ij} > U_{ik}, \forall k \neq j)$$

2.2 Ordered Choice (Brief Note)

Before we dive into unordered choices, note that if the alternatives represent a natural ranking (e.g., Education Level: Primary $\rightarrow$ Secondary $\rightarrow$ Graduate), we would use an **Ordered Probit**

or Logit.

Scale: $1, 2, 3, 4, 5$

In such cases, we model a latent variable crossing specific thresholds. However, our focus today is on choices where no natural ordering exists (e.g., Bus vs. Car vs. Train).

3 McFadden's Formulation (1974)

To make the probability statement $P(U_{ij} > U_{ik})$ operational, we must assume a distribution for the error terms ϵ_{ij}.

3.1 The Type I Extreme Value Distribution

Daniel McFadden, in his Nobel-winning work, utilized the **Type I Extreme Value (Gumbel) distribution**.

$$F(\epsilon) = \exp(-\exp(-\epsilon))$$

Economic Intuition: Why this specific distribution? It is not arbitrary. Just as the Normal distribution is the limit of *sums* of random variables (Central Limit Theorem), the Gumbel distribution is the limiting distribution of the **maximum** of a sequence of random variables (analogous to extreme value theory). If we view the error term as the maximum of many small, unobserved factors affecting utility, this assumption is structurally sound.

Under this assumption, the probability of choosing j takes the elegant closed-form known as the **Multinomial Logit (MNL)**:

$$P(y_i = j|z) = \frac{\exp(z_{ij}\theta)}{\sum_{l=0}^{J} \exp(z_{il}\theta)}$$

4 Variable Decomposition

The variable vector z_{ij} often contains two distinct types of data. We can decompose z_{il} into:

$$z_{il} = (\mathbf{x}_{il}, \mathbf{w}_i)$$

1. **Choice-variant data ($\mathbf{x}_{il}$):** Attributes that vary across alternatives (e.g., the price of the bus vs. the price of the car).
2. **Individual-variant data ($\mathbf{w}_i$):** Attributes of the individual that are constant across alternatives (e.g., the individual's income, age). Note that $\mathbf{w}_i$ is a **row vector**.

Combining these, the parameter vector splits into $\theta = (\beta', \alpha')'$. The general probability becomes:

$$P(y_i = j) = \frac{\exp(\mathbf{x}_{ij}\beta + \mathbf{w}_i\alpha_j)}{\sum_{k=0}^{J} \exp(\mathbf{x}_{ik}\beta + \mathbf{w}_i\alpha_k)}$$

Notice that β is constant (the marginal utility of price is the same), but α_j is indexed by j (income affects the utility of buying a Porsche differently than buying a Toyota).

5 Multinomial Logit (MNL)

We typically use the term "Multinomial Logit" when we focus solely on individual characteristics $\mathbf{w}_i$.

$$P_{ij} = P(y_i = j|\mathbf{w}_i) = \frac{\exp(\mathbf{w}_i\alpha_j)}{\sum_{l=0}^{J}\exp(\mathbf{w}_i\alpha_l)}$$

5.1 Identification and Normalization

If we add a constant vector δ to all α_j, the probabilities remain unchanged because the terms cancel out in the numerator and denominator. To identify the model, we must select a base outcome (numéraire).

$$\text{Set } \alpha_0 = 0$$

Thus, the probability for choice j simplifies to:

$$P_{ij} = \frac{\exp(\mathbf{w}_i\alpha_j)}{1 + \sum_{k=1}^{J}\exp(\mathbf{w}_i\alpha_k)}$$

5.2 Log-Odds Ratio and IIC

A crucial property of the MNL is the **Log-Odds Ratio**. Let us look at the ratio of probabilities between two choices j and k:

$$\frac{P_{ij}}{P_{ik}} = \frac{\exp(\mathbf{w}_i\alpha_j)}{\exp(\mathbf{w}_i\alpha_k)} = \exp(\mathbf{w}_i(\alpha_j - \alpha_k))$$

Taking the natural log:

$$\ln\left(\frac{P_{ij}}{P_{ik}}\right) = \mathbf{w}_i(\alpha_j - \alpha_k)$$

Interpretation: The relative odds of choosing j over k depend **only** on the parameters of j and k. They do **not** depend on the existence or attributes of other alternatives.

This property is known as **IIC (Independence of Irrelevant Choices)**, often referred to in literature as IIA. While computationally convenient, it imposes a strong restriction: removing one option (e.g., a red bus) assumes that its probability mass is redistributed proportionally among the remaining options (car, blue bus), which may not be realistic in all contexts (the "Red Bus/Blue Bus" paradox).

6 Maximum Likelihood Estimation (MLE)

Since y_i is discrete, we cannot use OLS. We use Full Information MLE.

6.1 The Likelihood Function

We define an indicator variable d_{ij}:

$$d_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{if } y_i \neq j \end{cases}$$

The log-likelihood function for the sample is:

$$\ln L = \sum_{i=1}^{N} \sum_{j=0}^{J} d_{ij} \ln P_{ij}$$

6.2 Derivatives and Score Function

To find the optimal parameters α_k, we take the derivative. This requires the derivative of the probability function P_{ij} with respect to the parameter vector α_k.

Step 1: The derivative of P_{ij} w.r.t α_k Recall $P_{ij} = \frac{e^{\mathbf{w}_i \alpha_j}}{\sum_l e^{\mathbf{w}_i \alpha_l}}$.

Case 1: $k = j$.

$$\frac{\partial P_{ij}}{\partial \alpha_j} = \frac{e^{\mathbf{w}_i \alpha_j}(\sum e^{\cdots}) - e^{\mathbf{w}_i \alpha_j} e^{\mathbf{w}_i \alpha_j}}{(\sum e^{\cdots})^2} \mathbf{w}_i' = (P_{ij} - P_{ij}^2)\mathbf{w}_i' = P_{ij}(1 - P_{ij})\mathbf{w}_i'$$

(Note: Since $\mathbf{w}_i$ is a row vector, the gradient w.r.t column α involves the transpose $\mathbf{w}_i'$).

Case 2: $k \neq j$.

$$\frac{\partial P_{ij}}{\partial \alpha_k} = \frac{0 - e^{\mathbf{w}_i \alpha_j} e^{\mathbf{w}_i \alpha_k}}{(\sum e^{\cdots})^2} \mathbf{w}_i' = -P_{ij} P_{ik} \mathbf{w}_i'$$

We can summarize using the Kronecker delta δ_{jk} (where $\delta_{jk} = 1$ if $j = k$, else 0):

$$\frac{\partial P_{ij}}{\partial \alpha_k} = P_{ij}(\delta_{jk} - P_{ik})\mathbf{w}_i'$$

6.3 First Order Condition (FOC)

Now, we differentiate the Log-Likelihood:

$$\frac{\partial \ln L}{\partial \alpha_k} = \sum_i \sum_j \frac{d_{ij}}{P_{ij}} \frac{\partial P_{ij}}{\partial \alpha_k}$$

Substitute the probability derivative derived above:

$$= \sum_i \sum_j \frac{d_{ij}}{P_{ij}} [P_{ij}(\delta_{jk} - P_{ik})\mathbf{w}_i']$$

$$= \sum_i \mathbf{w}_i' \left[\sum_j (d_{ij}\delta_{jk} - d_{ij}P_{ik}) \right]$$

$$= \sum_i \mathbf{w}_i' \left[d_{ik} - P_{ik} \sum_j d_{ij} \right]$$

Since $\sum_j d_{ij} = 1$ (the individual must choose exactly one option):

$$\frac{\partial \ln L}{\partial \alpha_k} = \sum_{i=1}^{N} (d_{ik} - P_{ik}) \mathbf{w}_i' = 0$$

Economic Intuition of FOC: This result is profound yet intuitive. It states that at the maximum likelihood estimate, the weighted sum of the **residuals** $(d_{ik} - P_{ik})$ is zero. The "error" (difference between observed choice and predicted probability) is orthogonal to the regressors $\mathbf{w}_i$. This is the nonlinear analogue to the OLS normal equations.

6.4 Hessian and Marginal Effects

To check second-order conditions or compute standard errors, we need the second derivatives.

$$\frac{\partial^2 \ln L}{\partial \alpha_j \partial \alpha_k'} = -\sum_i \mathbf{w}_i' \frac{\partial P_{ij}}{\partial \alpha_k'} \dots$$

More importantly for interpretation, we care about **Marginal Effects**: How does probability change when $\mathbf{w}_i$ changes? Let $T_{ij} = \frac{\partial P_{ij}}{\partial \mathbf{w}_i}$.

$$T_{ij} = P_{ij}\alpha_j - P_{ij} \sum_l P_{il}\alpha_l = P_{ij}(\alpha_j - \bar{\alpha})$$

where $\bar{\alpha} = \sum_l P_{il}\alpha_l$ is the probability-weighted average of the parameters. This shows that the effect of a variable depends on the distance of that alternative's parameter from the "average" parameter.

7 Conditional Logit

Finally, we consider the case where data varies by alternative, often termed the **Conditional Logit**.

$$P(y_i = j | z_{ij}) = \frac{\exp(\mathbf{x}_{ij}\beta)}{\sum_{l=0}^{J} \exp(\mathbf{x}_{il}\beta)}$$

Here, we do not need to normalize β to zero because the identification comes from the variation in $\mathbf{x}$ across j.

7.1 IIR / IIC Revisited

Similar to MNL, the Conditional Logit exhibits the IIC property:

$$P_{ij} = \frac{e^{\mathbf{x}_{ij}\beta}}{1 + \sum \dots}$$

$$\ln \frac{P_{ij}}{P_{ik}} = (\mathbf{x}_{ij} - \mathbf{x}_{ik})\beta$$

The log-odds ratio depends only on the difference in attributes between option j and option k.

Summary: We have moved from simple utility maximization to a tractable probabilistic model by assuming Type I Extreme Value errors. Whether using Multinomial Logit (individual characteristics) or Conditional Logit (choice attributes), the core mechanics of MLE remain focused on minimizing the gap between observed choices (d_{ij}) and predicted probabilities (P_{ij}).

Chapter 9 Linear Panel Data Models

8 Introduction: The Power of Panel Data

Welcome. Today, we transition from the flat world of cross-sectional data into the rich, multi-dimensional world of **Panel Data**. Broadly speaking, panel data is defined not just by having "data," but by possessing a specific structure: we observe the *same* units (individuals, firms, countries) repeatedly over time.

Let us denote our outcome variable as y_{it}, where $i = 1, \ldots, N$ represents the individual dimension and $t = 1, \ldots, T$ represents the time dimension.

Why do we care about panel data? The notes highlight two primary advantages:

1. **Increased Sample Size ($N \times T$):** We simply have more data points, improving the precision of our estimates.
2. **Causal Inference:** This is the "killer app" of panel data. It allows us to control for unobserved heterogeneity—those pesky omitted variables that are specific to an individual but constant over time (like "ability" in a wage equation or "culture" in a trade model).

9 Model Specification and Pooled OLS

9.1 The General Framework

A general linear specification might look like this:

$$y_{it} = x_{it}\beta_{it} + \epsilon_{it} \tag{1}$$

Here, x_{it} is a $1 \times K$ row vector of regressors. However, estimating a unique β_{it} for every observation is impossible (we would have more parameters than data). We must impose structure to make the heterogeneity manageable:

1. **Pooled ($\beta_{it} = \beta$):** We assume parameters are constant across everyone and every time period. This ignores the panel structure entirely (treating it as a 3D object flattened into 2D).
2. **Time Series ($\beta_{it} = \beta_i$):** We estimate separate betas for each individual. This requires large T.
3. **Repeated Cross-Section ($\beta_{it} = \beta_t$):** We allow parameters to change over time but force them to be the same across individuals.

9.2 Pooled OLS

If we assume complete homogeneity ($y_{it} = x_{it}\beta + \epsilon_{it}$), we can use **Pooled OLS**. The estimator is:

$$\hat{\beta}^{pool} = (X'X)^{-1}X'y \tag{2}$$

where X is the stacked $NT \times K$ matrix of regressors.

Assumptions for Pooled OLS:

1. **Rank Condition:** $\text{Rank}(E[x_{it}'x_{it}]) = K$. No perfect multicollinearity.
2. **Contemporaneous Exogeneity:** $E[\epsilon_{it}|x_{it}'] = 0$.
3. **Homoskedasticity/No Serial Correlation:** $Var(\epsilon_{it}|x_{it}') = \sigma^2$ and $E[\epsilon_{it}\epsilon_{js}|x_{it}, x_{js}] = 0$ for $i \neq j$ or $t \neq s$.

Critique: Pooled OLS completely ignores individual heterogeneity. If there is an unobserved effect correlated with x_{it}, Pooled OLS is biased and inconsistent.

10 Linear Unobserved Effects Models

This is the core of modern panel data analysis. We specify the model as:

$$y_{it} = c_i + x_{it}\beta + u_{it} \tag{3}$$

Here:

- c_i: The **Unobserved Individual Effect** (or heterogeneity). It captures everything specific to individual i that does not change over time.
- u_{it}: The **Idiosyncratic Error**. This changes over i and t.

10.1 The "Old" View: Fixed vs. Random Effects

Traditionally, the distinction was defined by the nature of c_i:

- **Fixed Effects:** c_i is a fixed parameter to be estimated (dummy variable).
- **Random Effects:** c_i is a random variable drawn from a distribution, part of the compound error term $(c_i + u_{it})$.

10.2 The "Modern" Approach

In modern econometrics (and in our class notes), we view c_i always as a random variable. The critical distinction is based on the correlation between c_i and the regressors x_{it}:

1. **Fixed Effects (FE):** We allow arbitrary correlation between the unobserved effect and the regressors.

$$Cov(x_{it}, c_i) \neq 0 \tag{4}$$

   Because they are correlated, we must "remove" c_i to avoid omitted variable bias.

2. **Random Effects (RE):** We assume the unobserved effect is uncorrelated with the regressors.

$$E[c_i|X_i] = 0 \quad (\text{or } Cov(x_{it}, c_i) = 0) \tag{5}$$

If this holds, we can treat c_i as part of the error term and use Generalized Least Squares (GLS) for efficiency.

10.3 Strict Exogeneity

For FE models to work reliably, we usually require a stronger assumption on the idiosyncratic error than just contemporaneous exogeneity. We need **Strict Exogeneity**:

$$E[u_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}, c_i] = 0 \tag{6}$$

Intuition: The error term u_{it} today is uncorrelated with regressors from the past, present, and future.

Counter-example: Strict exogeneity fails in dynamic panels where $y_{it} = \beta y_{i,t-1} + c_i + u_{it}$. Here, u_{it} affects y_{it}, which becomes $y_{i,t}$ (the regressor) in the next period. Therefore, $E[u_{it}|x_{i,t+1}] \neq 0$.

11 The Fixed Effects Model (FE)

11.1 Setup and LSDV

Recall the model:

$$y_{it} = c_i + x_{it}\beta + u_{it} \tag{7}$$

Or in vector notation for individual i (stacking T observations):

$$y_i = c_i e + X_i \beta + u_i \tag{8}$$

where e is a $T \times 1$ vector of ones: $e = [1, 1, \ldots, 1]'$. X_i is a $T \times K$ matrix.

One way to estimate this is **LSDV (Least Squares Dummy Variables)**. We literally add N dummy variables, one for each individual.

$$\hat{\beta}^{LSDV} \text{ comes from regressing } y \text{ on } X \text{ and } N \text{ dummies.} \tag{9}$$

While valid, if N is large (e.g., millions), inverting the matrix is computationally expensive. We need a trick.

11.2 The Within Transformation (De-meaning)

To eliminate c_i, we can subtract the individual-specific mean from the data. Define the mean over time for variable m_{it} as $\bar{m}_i = \frac{1}{T}\sum_{t=1}^{T} m_{it}$. Note that for the time-invariant effect, $\bar{c}_i = c_i$.

The transformation:

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + (c_i - c_i) + (u_{it} - \bar{u}_i) \tag{10}$$

The c_i disappears! We are left with time-demeaned data:

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{u}_{it} \tag{11}$$

We can now run OLS on this transformed data. This is the **Within Estimator**.

11.3 Matrix Derivation (The Annihilator Matrix Q)

Let's formalize this using matrix algebra as detailed in the notes. We define a projection matrix that creates the means: $P = \frac{1}{T}ee'$. We define the residual-maker (or de-meaning) matrix Q_T:

$$Q = I_T - P = I_T - \frac{1}{T}ee' \tag{12}$$

Structure of Q:

$$Q = \begin{bmatrix} 1 - \frac{1}{T} & -\frac{1}{T} & \dots \\ -\frac{1}{T} & 1 - \frac{1}{T} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{13}$$

Key Property: $Qe = (I - \frac{1}{T}ee')e = e - \frac{1}{T}e(e'e)$. Since $e'e = T$, this becomes $e - e = 0$. The matrix Q annihilates any time-invariant vector.

Applying Q to our equation for individual i:

$$Qy_i = Q(c_i e + X_i\beta + u_i) \tag{14}$$

$$Qy_i = c_i(Qe) + QX_i\beta + Qu_i \tag{15}$$

$$Qy_i = 0 + QX_i\beta + Qu_i \tag{16}$$

The Fixed Effects estimator is:

$$\hat{\beta}^{FE} = \left(\sum_{i=1}^{N} X_i'Q'QX_i\right)^{-1}\left(\sum_{i=1}^{N} X_i'Q'Qy_i\right) \tag{17}$$

Since Q is idempotent ($Q'Q = Q$) and symmetric ($Q' = Q$), this simplifies to:

$$\hat{\beta}^{FE} = \left(\sum_{i=1}^{N} X_i'QX_i\right)^{-1}\left(\sum_{i=1}^{N} X_i'Qy_i\right) \tag{18}$$

11.4 Properties of FE

1. Rank Condition (FE.2): We require $\text{Rank}(E[X_i'QX_i]) = K$. **Implication:** X_i cannot contain time-invariant variables (like gender or race). If a variable doesn't change over time, $x_{it} = \bar{x}_i$, so $\ddot{x}_{it} = 0$. The matrix $X_i'QX_i$ will be singular. Time-invariant variables are "absorbed" by the fixed effects.

2. Unbiasedness: Given Strict Exogeneity ($E[u_i|X_i, c_i] = 0$), implying $E[Qu_i|X_i] = 0$:

$$E[\hat{\beta}^{FE}|X] = \beta + E\left[\left(\sum X_i'QX_i\right)^{-1}\sum X_i'Qu_i\right] = \beta \tag{19}$$

Note: Strict exogeneity (past, present, future) is required because the de-meaning process $\ddot{u}_i$ contains error terms from all periods.

3. Consistency and Asymptotic Normality: As $N \to \infty$ (with T fixed):

$$\hat{\beta} \xrightarrow{p} \beta \text{ as } NT \to \infty \quad \sqrt{N}(\hat{\beta}^{FE} - \beta) \xrightarrow{d} N(0, AVar) \text{ as } N \to +\infty \tag{20}$$

The asymptotic variance is:

$$AVar(\hat{\beta}^{FE}) = \sigma_u^2[E(X_i'QX_i)]^{-1} \tag{21}$$

We estimate σ_u^2 using the residuals, correcting for degrees of freedom (N intercepts estimated):

$$\hat{\sigma}_u^2 = \frac{1}{NT - N - K}\sum_i\sum_t \hat{u}_{it}^2 \tag{22}$$

12 First Differencing (FD)

An alternative way to remove c_i is to subtract the previous period's equation from the current one.

$$y_{it} = c_i + x_{it}\beta + u_{it} \tag{23}$$

$$y_{i,t-1} = c_i + x_{i,t-1}\beta + u_{i,t-1} \tag{24}$$

Subtracting gives:

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it} \tag{25}$$

where $\Delta z_{it} = z_{it} - z_{i,t-1}$.

12.1 FD vs. FE

When $T = 2$, the FD and FE estimators are numerically identical. When $T > 2$, they differ based on assumptions about the error term u_{it}.

- **Efficiency:** If u_{it} is serially uncorrelated (white noise), FE is more efficient (BLUE).
- **Random Walk:** If u_{it} follows a random walk ($u_{it} = u_{i,t-1} + \epsilon_{it}$), then Δu_{it} is white noise. In this case, FD is more efficient.

13 Policy Estimation and DID

Panel data is widely used for evaluating policy interventions (Treatments).

13.1 Difference-in-Differences (DID)

Consider a simple setup with two periods ($t = 1, 2$) and two groups (Control, Treated).

$$y_{it} = c_i + \beta_2 D2_t + \beta_{DID}(Treat_i \times D2_t) + u_{it} \tag{26}$$

This is the standard DID specification.

- c_i absorbs the group difference (Treat vs Control).
- $D2_t$ (time dummy) absorbs the common trend.
- β_{DID} captures the treatment effect.

The estimator is the famous "double difference":

$$\hat{\beta}_{DID} = (\bar{y}_{Treat,2} - \bar{y}_{Treat,1}) - (\bar{y}_{Control,2} - \bar{y}_{Control,1}) \tag{27}$$

13.2 DID with Fixed Effects

With more periods, we generalize this:

$$y_{it} = c_i + \delta_t + \beta_{DID} D_{it} + u_{it} \tag{28}$$

where $D_{it} = 1$ if unit i is treated at time t. Note that we do not need the main terms $Treat_i$ or Post-period dummy because c_i absorbs time-invariant group status and δ_t absorbs time-specific shocks.

13.3 Event Study (Parallel Trends)

The crucial assumption for DID is **Parallel Trends**: in the absence of treatment, the treated and control groups would have moved in parallel. To test this, we use an Event Study specification:

$$y_{it} = c_i + \delta_t + \sum_{\tau=-K, \tau \neq -1}^{L} \beta_\tau D_{it}^\tau + u_{it} \tag{29}$$

Here, D_{it}^τ is a dummy indicating that observation it is τ periods away from the treatment start.

- We omit $\tau = -1$ as the reference category.
- **Pre-trends:** We check if $\hat{\beta}_\tau \approx 0$ for $\tau < 0$. If they are significantly different from zero, the parallel trends assumption likely fails (the groups were already diverging).
- **Dynamic Effects:** The coefficients for $\tau > 0$ show how the treatment effect evolves over time.

14 Random Effects Model (RE)

14.1 Assumptions

We return to $y_{it} = c_i + x_{it}\beta + u_{it}$, but now we assume:

1. **Strict Exogeneity:** $E[u_{it}|X_i, c_i] = 0$.
2. **Orthogonality:** $E[c_i|X_i] = 0$ (or constant). The unobserved effect is *uncorrelated* with regressors.

If assumption (2) holds, Pooled OLS is consistent, but it is inefficient because the composite error $v_{it} = c_i + u_{it}$ is serially correlated. For the same individual, v_{it} and v_{is} both contain c_i.

14.2 The GLS Structure

We treat $v_i = c_i e + u_i$ as the error vector for individual i. The variance-covariance matrix of the error $\Omega = E[v_i v_i']$ is:

$$\Omega = E[(c_i e + u_i)(c_i e + u_i)'] = \sigma_c^2 ee' + \sigma_u^2 I_T \tag{30}$$

In matrix form:

$$\Omega = \begin{bmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ \sigma_c^2 & \dots & \dots & \sigma_c^2 + \sigma_u^2 \end{bmatrix} \tag{31}$$

14.3 Deriving the GLS Estimator

To use GLS, we need Ω^{-1}. The notes provide a beautiful derivation using projection matrices P and Q. Recall $P = \frac{1}{T}ee'$ and $Q = I_T - P$. We can write Ω as a linear combination of P and Q:

$$\Omega = \sigma_u^2 I_T + T\sigma_c^2 (\frac{1}{T} ee') \tag{32}$$

$$\Omega = \sigma_u^2 (P + Q) + T\sigma_c^2 P \tag{33}$$

$$\Omega = (\sigma_u^2 + T\sigma_c^2) P + \sigma_u^2 Q \tag{34}$$

Using the algebra of idempotent matrices (where P^{-1} in the spectral sense works on eigenvalues), the inverse is:

$$\Omega^{-1} = \frac{1}{\sigma_u^2 + T\sigma_c^2} P + \frac{1}{\sigma_u^2} Q \tag{35}$$

Simplifying this expression leads to the transformation parameter ψ (or θ in some texts).

$$\Omega^{-1} = \frac{1}{\sigma_u^2} \left[Q + \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_c^2} P \right] \tag{36}$$

Let $\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_c^2}$.

The GLS estimator is effectively OLS on transformed data:

$$y_{it} - (1 - \sqrt{\psi})\bar{y}_i = (x_{it} - (1 - \sqrt{\psi})\bar{x}_i)\beta + \text{error} \tag{37}$$

This highlights that RE is a **weighted average** of Pooled OLS and Fixed Effects.

- If $\sigma_c^2 = 0$, then $\psi = 1$. The transformation vanishes. We get **Pooled OLS**.
- If $\sigma_c^2 \to \infty$ (huge heterogeneity), $\psi \to 0$. The transformation becomes $y_{it} - \bar{y}_i$. We get **Fixed Effects**.

14.4 Feasible GLS (FGLS)

Since σ_u^2 and σ_c^2 are unknown, we estimate them (FGLS).

1. Run Pooled OLS or FE to get residuals.
2. Estimate $\hat{\sigma}_{total}^2$ and use covariance of residuals to separate $\hat{\sigma}_c^2$ and $\hat{\sigma}_u^2$ (as detailed in Page 14 of the notes).

15 Testing and Extensions

15.1 Hausman Test

How do we choose between FE and RE?

- $H_0 : E[c_i|X_i] = 0$. (RE is consistent and efficient; FE is consistent but inefficient).
- $H_A : E[c_i|X_i] \neq 0$. (RE is inconsistent; FE is consistent).

The test statistic compares the beta estimates:

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[\widehat{Var}(\hat{\beta}_{FE}) - \widehat{Var}(\hat{\beta}_{RE})]^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \tag{38}$$

Under H_0, $H \sim \chi_K^2$. If H is large, we reject the RE assumptions and stick with Fixed Effects.

15.2 Correlated Random Effects (Mundlak/Chamberlain)

This is a middle ground. Instead of assuming c_i is uncorrelated with x (RE) or arbitrarily correlated (FE), we model the correlation explicitly:

$$c_i = \bar{x}_i\gamma + a_i \tag{39}$$

where a_i is uncorrelated with x. Substituting this into the main equation allows us to use RE estimation while controlling for the correlation via the time-averages $\bar{x}_i$.

This concludes the lecture on Linear Panel Models. Remember: The choice between FE and RE is often a trade-off between robustness (FE) and efficiency (RE), but in modern applied economics, we prioritize robustness against omitted variable bias, making Fixed Effects (and DID) the dominant tool.