Lecture Notes:
Econometrics II

Based on lectures by **Marko Mlikota** in Spring semester, 2025

Draft updated on April 6, 2025

These lecture notes were taken in the course *Econometrics II* taught by **Marko Mlikota** at Graduate Institute of International and Developoment Studies, Geneva as part of the International Economics program (Semester II, 2024).

Currently, these are just drafts of the lecture notes. There can be typos and mistakes anywhere. So, if you find anything that needs to be corrected or improved, please inform at `jingle.fu@graduateinstitute.ch`.

Contents

Lecture 1.

Review of Econometrics I

1.1 Basic assumptions

Firstly, we recall the basic assumptions of the linear regression model.

Assumption 1.1.1 (Basic Assumptions).

A0. (Correct Specification). Model is correctly specified: $y_i = x_i'\beta + u_i$

A1. (Independent Sampling). Observations $z_i = \{y_t, x_i\}_{i=1}^n$ are independent across i.

A2. (Full rank). The matrix $X'X = \sum x_i x_i'$ is of full rank.

A3. (Conditional Independence). $\mathbb{E}[u_i|x_i] = 0$.

A4. (Homoskedasticity). $\mathbb{V}[u_i|x_i] = \sigma^2$ for all i.
$\mathbb{V}[y_i] = \mathbb{V}[x_i'\beta + u_i|x_i] = \sigma^2$

Under these four basic assumptions, and that x_i is exogenous, giving $\mathbb{E}[x_i u_i] = 0$, then

$$\hat{\beta} = (X'X)^{-1}X'Y \xrightarrow{p} \beta.$$

1.2 Frisch-Waugh-Lovell Theorem

Definition 1.2.1 (Partitioned regression).

We consider a normal linear regression model $Y = X\beta + U$. Let X be partitioned as:

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

where X is $n \times K$, X_1 is $n \times K_1$ and X_2 is $n \times K_2$. And we partition the parameter vector β accordingly:

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

where β_1 is $K_1 \times 1$ and β_2 is $K_2 \times 1$. Thus, the model can be written as:

$$Y = X_1\beta_1 + X_2\beta_2 + U$$

where U is the error term.

Also take the following notation:

$$P_1 = X_1(X_1'X_1)^{-1}X_1', \quad M_1 = I - P_1, \quad \tilde{X}_2 = M_1X_2, \quad \tilde{U} = M_1Y$$

thus $\tilde{U}$ is the residual vector from the regression of Y on X_1, and the k-th column of $\tilde{X}_2$ is the residual vector from the regression of the corresponding k-th column of X_2 on X_1.

The OLS estimator $\beta = (\beta_1, \beta_2)$ can be obtained by regression of Y on $X = [X_1, X_2]$, and can be written as:

$$Y = X\hat{\beta} + \hat{U} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{U}$$

We are interested in algebraic expressions for $\hat{\beta}_1$ and $\hat{\beta}_2$.

Let's first focus on $\hat{\beta}_1$. The least squares estimator $\hat{\beta}_1$ is found by the joint minimization:

$$(\hat{\beta}_1, \hat{\beta}_2) = \arg\min_{\beta_1,\beta_2} (Y - X_1\beta_1 - X_2\beta_2)' (Y - X_1\beta_1 - X_2\beta_2)$$

Denote $(Y - X_1\beta_1 - X_2\beta_2)' (Y - X_1\beta_1 - X_2\beta_2)$ as $\text{SSE}(\beta_1, \beta_2)$.

By nested minimization, we can rewrite the above as:

$$\hat{\beta}_1 = \arg\min_{\beta_1} \left(\min_{\beta_2} \text{SSE}(\beta_1, \beta_2) \right)$$

For the inner minimization problem: $\min_{\beta_2} \text{SSE}(\beta_1, \beta_2)$, this is simply the regression of $Y - X_1\beta_1$ on X_2, with the solution:

$$\arg\min_{\beta_2} \text{SSE}(\beta_1, \beta_2) = (X_2'X_2)^{-1}X_2'(Y - X_1\beta_1)$$

with residuals:

$$Y - X_1\beta_1 - X_2(X_2'X_2)^{-1}X_2'(Y - X_1\beta_1) = (M_2Y - M_2X_1\beta_1) = M_2(Y - X_1\beta_1)$$

where $M_2 = I - X_2(X_2'X_2)^{-1}X_2'$ is the annihilator matrix for X_2.

So the inner minimization problem has minimized value:

$$\min_{\beta_2} \text{SSE}(\beta_1, \beta_2) = (Y - X_1\beta_1)'M_2'M_2(Y - X_1\beta_1) = (Y - X_1\beta_1)'M_2(Y - X_1\beta_1)$$

Substituting this into the outer minimization problem, we have:

$$\begin{aligned}
\hat{\beta}_1 &= \arg\min_{\beta_1} (Y - X_1\beta_1)'M_2(Y - X_1\beta_1) \\
&= (X_1'M_2X_1)^{-1} (X_1'M_2Y)
\end{aligned}$$

By a similar argument we find

$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1} (X_2'M_1Y)$$

By the previous notation and the fact that M_1 and M_2 are idempotent matrices, we can have:

$$\begin{aligned}
\hat{\beta}_2 &= (X_2'M_1X_2)^{-1} (X_2'M_1Y) \\
&= (X_2'M_1M_1X_2)^{-1} (X_2'M_1M_1Y) \\
&= \left(\tilde{X}_2'\tilde{X}_2\right)^{-1} \left(\tilde{X}_2\tilde{U}\right)
\end{aligned}$$

Thus the coefficient estimator $\hat{\beta}_2$ is algebraically equivalent to the OLS estimator of the regression of $\tilde{U}$ on $\tilde{X}_2$. Notice that these two are M_1Y and M_1X_2, and we know that pre-multiplication by M_1 creates least squares residuals. Therefoe, $\tilde{U}$ is simply the least squares residual from a regression of Y on X_1, and the columns of $\tilde{X}_2$ are the least squares residuals from a regression of the columns of X_2 on X_1. From the above steps, we have proven the following theorem.

Theorem 1.2.1 (Frisch-Waugh-Lovell (FWL) theorem)**.**

In the model $Y = X_1\beta_1 + X_2\beta_2 + U$, the OLS estimator of β_2 and the OLS residuals $\hat{U}$ may be computed via the following two steps:

1. Regress Y on X_1 and obtain the residuals $\hat{U}_1 = Y - X_1\hat{\beta}_1$.
2. Regress the columns of X_2 on X_1 and obtain the residuals $\tilde{X}_2 = M_1 X_2$.
3. Regress $\hat{U}$ on $\tilde{X}_2$: $\hat{U} = \tilde{X}_2 b + V$ and obtain the OLS estimator $\hat{\beta}_2 = \hat{b}$ and the residual $\hat{U} = \hat{V}$.

1.3 Endogeneity

We say that there's endogeneity in the linear model

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

if β is the parameter of interest and

$$\mathbb{E}[x_i u_i] \neq 0.$$

This is a core problem in econometrics and largely differentiates the field from statistics.

Endogeneity implies that the least squares estimator is inconsistent for the structural parameter. Indeed, under i.i.d. sampling, least squares is consistent for the projection coefficient.

$$\hat{\beta} \xrightarrow{p} \beta + \left(\mathbb{E}[XX']\right)^{-1}\mathbb{E}[Xu] \neq \beta$$

The inconsistency of least squares is typically referred to as **endogeneity bias** or **estimation bias** due to endogeneity.

Commonly, there are three reasons for endogeneity:

1. Measurement error: x_i is measured with error.

   Suppose our true Regression is: $y_i = x_i^{*\prime}\beta + \varepsilon_i$, $\mathbb{E}[x_i^*\varepsilon_i] = 0$, β is the structural parameter. But, $x_i^{*\prime}$ is not observed. Instead, we observe: $x_i = x_i^* + v_i$, where v_i is the measurement error, independent of x_i^* and ε_i: $\mathbb{E}[x_i^* v_i'] = 0$, $\mathbb{E}[v_i\varepsilon_i] = 0$ [1].

   The model $x_i = x_i^* + v_i$ with x_i^* and v_i uncorrelated, and $\mathbb{E}[v_i] = 0$ is known as the **classical measurement error model**. This means that x_i is a noisy but unbiased estimate of x_i^*. By substitution we can express y_i as a function of the observed variable x_i.

   $$y_i = x_i^{*\prime}\beta + \varepsilon_i = (x_i - v_i)'\beta + \varepsilon_i = x_i'\beta + u_i$$

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

   This means that (y_i, x_i) satisfy the linear equation $y_i = x_i'\beta + u_i$ with an error u_i. But this error is not a projection error.

   $$\begin{aligned}\mathbb{E}[x_i u_i] &= \underbrace{\mathbb{E}[x_i\varepsilon_i]}_{0} - \mathbb{E}[x_i v_i']\beta \\ &= -\mathbb{E}[(x_i^* + v_i)v_i']\beta \\ &= -\underbrace{\mathbb{E}[x_i^* v_i']}_{0}\beta - \mathbb{E}[v_i v_i']\beta \\ &= -\mathbb{E}[v_i v_i']\beta \neq 0\end{aligned}$$

   if $\mathbb{E}[v_i v_i'] \neq 0$ and $\beta \neq 0$.

[1] This is an example of a latent variable model, where "latent" refers to an unobserved structural variable.

Remark (Measurement Error Bias).

Let's rewrite in matrix form: $Y = X^{*\prime}\beta + \varepsilon$, $X = X^* + v$, $\mathbb{E}[X^* v] = 0$, $\mathbb{E}[\varepsilon v] = 0$, v is a $k \times 1$ error. We can write $Y = (X - v)'\beta + \varepsilon = X'\beta + u$, where $u = \varepsilon - v'\beta$. And we have: $\mathbb{E}[Xu] = \mathbb{E}[(X^* + v)(\varepsilon - v'\beta)] = -\mathbb{E}[vv']\beta \neq 0$ if $\beta \neq 0$ and $\mathbb{E}[vv'] \neq 0$.

We can calculate the form of the projection coefficient (which is consistently estimated by least squares). For simplicity suppose that $k = 1$, we find:

$$\beta^* = \beta + \frac{\mathbb{E}[Xu]}{\mathbb{E}[X^2]} = \beta\left(1 - \frac{\mathbb{E}[v^2]}{\mathbb{E}[X^2]}\right)$$

Since $\frac{\mathbb{E}[v^2]}{\mathbb{E}[X^2]} < 1$, the projection coefficient shrinks the structural parameter β towards zero. This is called **measurement error bias** or **attenuation bias**.

2. Simultaneity (Reverse causality): Simultaneity arises when at least one of the explanatory variables is determined simultaneously along with y. If, say, x_i is determined partly by y, and x_i and u_i are generally correlated.

$$y_i = x_i'\beta + u_i = x_{i1}^*\beta_1 + x_{i2}\beta_2 + u_i, \quad x_i = z_i'\gamma + y_i\delta + v_i.$$

Example 1 (Supply and Demand).

The variables Q and P (quantity and price) are determined jointly by the demand equation:

$$Q = -\beta_1 P + u_1$$

and supply function:

$$Q = \beta_2 P + u_2$$

where u_1 and u_2 are the demand and supply shocks, respectively. Assume that $u = (u_1, u_2)$ satisfy that $\mathbb{E}[u] = 0$ and $\mathbb{E}[uu'] = I_2$ (for simplicity). The question is: If we regress Q on P, what will happen? Let's solve P and Q in error terms:

$$\begin{aligned}
\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}\begin{bmatrix} Q \\ P \end{bmatrix} &= \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\
\begin{bmatrix} Q \\ P \end{bmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\
&= \frac{1}{\beta_1 + \beta_2}\begin{bmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\
&= \begin{bmatrix} \frac{\beta_2 u_1 + \beta_1 u_2}{\beta_1 + \beta_2} \\ \frac{u_1 - u_2}{\beta_1 + \beta_2} \end{bmatrix}
\end{aligned}$$

The projection of Q on P yields $Q = \beta^* P + u^*$, where $\mathbb{E}[Pu^*] = 0$ and the projection coefficient is

$$\beta^* = \frac{\mathbb{E}[PQ]}{\mathbb{E}[P^2]} = \frac{\beta_2 - \beta_1}{2}.$$

The OLS estimator satisfies $\hat{\beta} \xrightarrow[p]{} \beta^*$ and the limit does not equal either β_1 or β_2. This is called **simultaneity bias** or **simultaneous equation bias**.

This occurs generally when Y and X are jointly determined, as in a market equilibrium. Generally, when both the dependent variable and a regressor are simultaneously determined then the regressor should be treated as endogenous.

3. Omitted variables: The most prominent cause of endogeneity are omitted variables (OVs).

   Suppose the true regression is: $y_i = x_i'\beta + w_i'\delta + \varepsilon_i$, where exogeneity holds: $\mathbb{E}[x_i\varepsilon_i] = 0$, $\mathbb{E}[w_i\varepsilon_i] = 0$. If we omit w_i and instead estimates:

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

   where $u_i = w_i'\delta + \varepsilon_i$, then in this misspecified model, exogeneity is only given if x_i and w_i are uncorrelated, since:

$$\begin{aligned}\mathbb{E}[x_i u_i] &= \mathbb{E}[x_i(w_i'\delta + \varepsilon_i)] \\ &= \mathbb{E}[x_i w_i']\delta + \underbrace{\mathbb{E}[x_i\varepsilon_i]}_{0}\end{aligned}$$

   Since $\hat{\beta} - \beta \xrightarrow{p} \mathbb{E}[x_i x_i']^{-1}\mathbb{E}[x_i u_i]$, we can assess the sign and size of the asymptotic bias based ont the signs of correlation between x_i and w_i.

For our general regression model $y_i = x_i'\beta + u_i$, we have $\mathbb{E}[x_i u_i] \neq 0$, thus $\hat{\beta}_{OLS} \xrightarrow{p} \beta$ doesn't hold.

To consistently estimate β, we require additional assumptions. One type of information which is commonly used in economics is the **instruments**.

Definition 1.3.1 (Instrumental Variable).

We take $z_i \in \mathbb{R}^r$ as an instrumental variable if:

$$\begin{aligned}\mathbb{E}[z_i u_i] &= 0 \\ \mathbb{E}[z_i x_i] &\neq 0 \\ \mathbb{E}[z_i z_i'] &> 0 \\ \text{rank}\Big(\mathbb{E}[z_i x_i']\Big) &= k \leq r \quad {}^{2}\end{aligned}$$

[2] We say that the model is just-identified if $k = r$ and over-identified if $k < r$.

1.3.1 Instrumental Variables and 2SLS

Then, we have the 2SLS method:

Definition 1.3.2 (2SLS Method).

1. Estimate: $x_i = z_i'\gamma + e_i \Rightarrow \hat{\gamma} = (Z'Z)^{-1}Z'X \Rightarrow \hat{X} = Z'\hat{\gamma} = P_Z X$;

2. Estimate: $y_i = \hat{x}_i'\beta + u_i^*$.

$$\begin{aligned}\hat{\beta}_{2SLS} &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y \\ &= \left((P_Z X)'P_Z X\right)^{-1}(P_Z X)'Y \\ &= (X'P_Z X)^{-1}X'P_Z Y \\ &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}X'Z(Z'Z)^{-1}Z'Y \\ &= \beta + \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}X'Z(Z'Z)^{-1}Z'u \\ &= \beta + (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'u \quad {}^{3} \\ &= \beta + (Z'X)^{-1}Z'u \\ &\xrightarrow{p} \beta.\end{aligned}$$

To compute $\hat{\beta}_{2SLS}$, we need $Z'Z$ to be full rank, which requires us to have more observations than IVs.

Ideally, z_i should be as highly correlated with x_i as possible, but uncorrelated with u_i. To see this, we find hte variance of $\hat{\beta}_{2SLS}$

$$\begin{aligned}
\mathbb{V}[\hat{\beta}_{2SLS}|X,Z] &= \mathbb{V}\left[(X'P_ZX)^{-1}X'P_ZU|X,Z\right] \\
&= (X'P_ZX)^{-1}\,\mathbb{V}\left[X'P_ZU|X,Z\right](X'P_ZX)^{-1} \\
&= (X'P_ZX)^{-1}X'P_Z\mathbb{E}[UU'|X,Z]P_ZX\,(X'P_ZX)^{-1} \\
&= (X'P_ZX)^{-1}\sigma^2
\end{aligned}$$

which holds under homoskedasticity. As we know $\mathbb{V}[\hat{\beta}_{OLS}] = (X'X)^{-1}\sigma^2$,

$$\begin{aligned}
\mathbb{V}\left[\hat{\beta}_{OLS}\right]^{-1} - \mathbb{V}\left[\hat{\beta}_{2SLS}\right]^{-1} &= (\sigma^2)^{-1}X'X - (\sigma^2)^{-1}X'P_ZX \\
&= (\sigma^2)^{-1}X'(I-P_Z)X \\
&= (\sigma^{-2})X'M_ZX \\
&= \sigma^{-2}(\underbrace{M_ZX}_{\hat{E}})'M_ZX \\
&= \sigma^{-2}SSR_{1SLS} > 0.
\end{aligned}$$

This means that the variance of 2SLS estimator is larger than that of the OLS.

By the usual arguments, the asymptotic analysis reveals that:

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

where

$$V_{2SLS} = Q_{XZ}^{-1}X'Z\,(Z'Z)^{-1}\,Z'UU'Z\,(Z'Z)^{-1}\,(X'Z)'\,Q_{XZ}^{-1}$$

where $Q_{XZ} = (Z'X)\,(Z'Z)^{-1}\,(X'Z)$

As usual, we can estimate it by replacing u_i with $\hat{u}_i$ and expectation operators with population means. Thereby, it's important to note that $u_i \neq u_i^*$, and to obtain $\hat{u}_i$, we don't use $\hat{x}_i$,but x_i:

$$\hat{u}_i = y_i - x_i'\hat{\beta}_{2SLS}$$

Under homoskedasticity, $V_{2SLS} = \sigma^2 Q_{XZ}^{-1}$, which we estimate using $\hat{\sigma}^2 = \frac{1}{n}\sum_i u_i^2$.

1.3.2 Weak Identification in IV Models

If the the correlation between x_i and z_i is weak, then we say it's a **weak instrument**. Under weak IVs, the finite sample distribution of $\hat{\beta}_{2SLS}$ may not assemble the asymptotic property.

In absence of an asymptotic distribution, we can conduct inference using its numerical approximation via bootstrapping. Or alternatively, we can construct a confidence set for β using the Following the procedure of Anderson and Rubin (1949).

The method is based on the idea that, for $\beta = \beta_0$, the auxiliary regression $y_i - x_i'\beta = \delta z_i + v_i$ should yield $\delta = 0$, because $y_i - x_i'\beta_0 = u_i$ and u_i is uncorrelated with z_i.

Theorem 1.3.1 (Anderson-Rubin Method).

For a given β_0, we get:

$$\sqrt{n}\hat{\delta}(\beta_0) = \sqrt{n}\left(Z'Z\right)^{-1} Z'(Y - X\beta_0) = \left(Z'Z\right)^{-1} \sqrt{n}Z'U \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_u^2}{\mathbb{E}(z_i^2)}\right)$$

which allows us to test $\mathcal{H}_0 : \delta = 0$. For many βs, test: $\mathcal{H}_0 : \delta(\beta) = 0$, e.g. using t-test.

$$T_t = \frac{\hat{\delta}(\beta_0)}{se(\hat{\delta}(\beta_0))} = \frac{\hat{\delta}_0}{\sqrt{\hat{\sigma}_u^2/Z'Z}} \xrightarrow{d} \mathcal{N}(0, 1)$$

The 90% CI for β is the set of βs at which $\delta(\beta) = 0$ cannot be rejected at 90% confidence level. A confidence set for β is given by taking all β_0 such that $\mathcal{H}_0 : \delta = 0$ cannot be rejected.

Remark (About Anderson-Rubin (AR) Test). [a]

Consider our model

$$\begin{aligned} y &= X\beta + u, \\ X &= Z\Pi + v, \end{aligned}$$

where X is one-dimensional and test for hypothesis $H_0 : \beta = \beta_0$. Under the null, vector $y - X\beta$ is equal to the error u_t and is uncorrelated with Z (due to exogeneity of instruments). The suggested statistics is:

$$AR(\beta_0) = \frac{(y - X\beta)' P_Z (y - X\beta)}{(y - X\beta)' M_Z (y - X\beta)/(T - k)}.$$

here $P_Z = Z(Z'Z)^{-1}Z', M_Z = I - P_z$.

The distribution of AR does not depend on μ asymptotically $AR \to \chi_k^2/k$. The formula may remind you of the J-test for over-identifying restrictions. It would be a J-test if one were to plugs in $\hat{\beta}_{TSLS}$. In a more general situation of more than one endogenous variable and/or included exogenous regressors AR statistic is F-statistic testing that all coeffcients on Z are zero in the regression of $y - \beta_0 X$ on Z and W.

Note, that one tests all coefficients β simultaneously (as a set) in a case of more than one endogenous regressor. AR confidence set One can construct a confidence set robust towards weak instruments based on the AR test by inverting it. That is, by finding all β which are not rejected by the data. In this case, it is the set :

$$CI = \{\beta_0 : AR(\beta_0) < \chi_{k,1-\alpha}^2\}.$$

The nice thing about this procedure is that solving for the confidence set is equivalent to solving a quadratic inequality. This confidence set can be empty with positive probability (caution!).

[a] Retrieved from MIT14.384 Time Series Analysis, Fall 2007 Professor Anna Mikusheva, Lecture 7-8, https://ocw.mit.edu/courses/14-384-time-series-analysis-fall-2013/365cba34145fa204731e9df202d4771e_MIT14_384F13_lec7and8.pdf

Lecture 2.

Causal Inference

Rubin (1975[11]) and Holland (1986[12]) made up the aphorism[1]:

> *"No causation without manipulation"*

Not everybody agrees with this point of view. In our lecture, we'll define causal effects using the potential outcomes framework (Neyman, 1923[13]; Rubin, 1974[14]).

2.1 Potential Outcomes Framework

In this framework, an experiment, or at least a thought experiment, has a treatment, and we are interested in its effect on an outcome or multiple outcomes. Sometimes, the treatment is also called an intervention or a manipulation.

Firstly, we consider an experiment with n units indexed by $i = 1, 2, \cdots, n$. We focus on a treatment with two levels:

$$d_i = \begin{cases} 0 & \text{control} \\ 1 & \text{treatment} \end{cases}$$

We seek to identify the causal effect of treatment d_i on some outcome y_i. For each i, the outcome od interest y_i has two versions:

$$y_i = \begin{cases} y_{0i} & d_i = 0 \\ y_{1i} & d_i = 1 \end{cases}$$

This notation emphasizes that y_{di} is the realization of the outcome y_i that would materialize if unit i received treatment $d_i = d$.

Neyman (1923[13]) first used this notation. It seems intuitive but has some hidden assumptions. Rubin (1980[15]) made the following clarifications on the hidden assumptions.

Assumption 2.1.1 (No interference).

Unit i's potential outcomes do not depend on other units' treatments. This is sometimes called the no-interference assumption.

Assumption 2.1.2 (Consistency).

There are no other versions of the treatment. Equivalently, we require that the treatment levels be well-defined, or have no ambiguity at least for the outcome of interest. This is sometimes called the consistency assumption.

The causal effect of the treatment on the i-th unit is then defined as:

$$\Delta_i = y_{1i} - y_{0i}$$

These potential outcomes are constants at the level of unit i.

Remark (Problem of causal inference).

The fundamental problem in causal inference is that only one treatment can be assigned to a given individual, and so only one of y_{0i} and y_{1i} can be observed. Thus Δ_i can never be observed.

Definition 2.1.1 (Stable Unit Treatment Value Assumption (SUTVA)).

Rubin (1980[15]) called the Assumptions 2.1.1 and 2.1.2 above together the *Stable Unit Treatment Value Assumption (SUTVA).*

The observed outcome of unit i is a function of the potential outcomes and the treatment indicator, we can write:

$$y_i = d_i y_{1i} + (1 - d_i) y_{0i}$$

In principle, by virtue of being (discrete) RVs, both d_i and y_i each have a distribution function, which, together with their possible realizations, defines various moments. However, their unconditional probabilities and moments at the level of unit i is not of interest. Only the conditional probabilities of y_i given d_i is of interest.

Remark (Rubin (2005[16])).

Under SUTVA, Rubin (2005) called the $n \times 2$ matrix of potential outcomes the Science Table:

i	y_{1i}	y_{0i}
1	y_{11}	y_{01}
2	y_{12}	y_{02}
$\vdots$	$\vdots$	$\vdots$
n	y_{1n}	y_{0n}

Due to the fundamental contributions of Neyman and Rubin to statistical causal inference, the potential outcomes framework is sometimes referred to as the Neyman Model, the Neyman-Rubin Model, or the Rubin Causal Model. Causal effects are functions of the Science Table. Inferring individual causal effects

$$\tau_i = y_{1i} - y_{0i}, \quad (i = 1, \ldots, n)$$

is fundamentally challenging because we can only observe either y_{1i} or y_{0i}, for each unit i, that is, we can observe only half of the Science Table.

SUTVA(2.1.1) ensures that the individual treatment effect is well defined.

Now, although Δ_i itself is unobservable, we can (perhaps remarkably) use randomized experiments to learn certain properties of it. The expectations $\mathbb{E}[y_{0i}]$ and $\mathbb{E}[y_{1i}]$ denote the average potential outcomes across unit i in population.

In particular, large randomized experiments let us recover the **Average Treatment Effect (ATE)**:

$$\text{ATE} = \mathbb{E}[y_{1i} - y_{0i}] = \mathbb{E}[y_{1i}] - \mathbb{E}[y_{0i}]$$

For a population, we can define the treatment conditional expectations:

$$\mathbb{E}[y_i|d_i = 1], \mathbb{E}[y_{0i}|d_i = 1], \mathbb{E}[y_{1i}|d_i = 1] = \mathbb{E}[y_i|d_i = 1]$$

that denote the averages of the outcome y_i.

Analogously, we can define the control conditional expectations:

$$\mathbb{E}[y_i|d_i = 0], \mathbb{E}[y_{0i}|d_i = 0] = \mathbb{E}[y_i|d_i = 0], \mathbb{E}[y_{1i}|d_i = 0]$$

for the non-treated subpopulation.

Similar to ATE, we candefine the Average Treatment Effect for the Treatment-Group (ATT) and the Average Treatment Effect for the Control-Group (ATC) as distinct objects:

$$\begin{aligned}
\text{ATT} &= \mathbb{E}[y_{1i} - y_{0i}|d_i = 1] \\
\text{ATC} &= \mathbb{E}[y_{1i} - y_{0i}|d_i = 0] \\
\mathbb{E}[z] = \mathbb{E}[z|d = 1]\mathbb{P}[d = 1] &+ \mathbb{E}[z|d = 0]\mathbb{P}[d = 0] = \mathbb{E}[\mathbb{E}[z|d]]
\end{aligned}$$

2.1.1 Identification of Causal Effects

Now, suppose we observe treatments and outcomes over a random sample n from the overall population, $\{d_i, y_i\}_{i=1}^n = \{d_i, y_{d_i i}\}_{i=1}^n$, as either $y_I = y_{0i}$, or $y_i = y_{1i}$.

Let $n_w = |\{i : d_i = w\}|$ be the size of sets of units in our sample who received and did not receive treatment, respectively. This means that: while we observe a sample of size n of d_i and y_i from the overall population, we are observing a sample of size n_0 of realizations of y_{0i} from the non-treated subpopulation and a sample of size n_1 of realizations of y_{1i} from the treated subpopulation.

$N = \{i = 1, 2, \cdots, n\}$, $N_1 = \{i \in N : d_i = 1\} \leftarrow n_1 = |N_1|$, $N_0 = \{i : d_i = 0\} \leftarrow n_0 = |N_0|$.

Based on this data, we can use the analogy principle to consistently estimate the first term in the ATT formula and the second term in the ATC formula:

$$\begin{aligned}
\frac{1}{n_1}\sum_{i \in N_1} y_i = \frac{1}{n_1}\sum_{i \in N_1} y_{1i} \xrightarrow{p} \mathbb{E}[y_{1i}|d_i = 1] = \mathbb{E}[y_i|d_i = 1] \\
\frac{1}{n_0}\sum_{i \in N_0} y_i = \frac{1}{n_0}\sum_{i \in N_0} y_{0i} \xrightarrow{p} \mathbb{E}[y_{0i}|d_i = 0] = \mathbb{E}[y_i|d_i = 0]
\end{aligned}$$

Without further assumptions, we cannot identify the remaining terms. Firstly, we cannot observe $\mathbb{E}[y_{0i}|d_i = 1]$ and $\mathbb{E}[y_{1i}|d_i = 0]$ because we do not observe y_{0i} for treated units, and we do not observe y_{1i} for non-treated units. Secondly, we can not observe $\mathbb{E}[y_{1i}]$ and $\mathbb{E}[y_{0i}]$ because both N_1 and N_0 are random samples from the overall population. As a result, the ATE is in general not identified from our data!

We can define the the difference-in-means estimator as:

$$\hat{\tau}_{DM} = \frac{1}{n_1}\sum_{i \in N_1} y_i - \frac{1}{n_0}\sum_{i \in N_0} y_i \xrightarrow{p} \mathbb{E}[y_{1i}|d_i = 1] - \mathbb{E}[y_{0i}|d_i = 0] = \text{ATE} = \text{ATT} = \text{ATC}.$$

We define the difference of treated and non-treated as: *Naive Difference.*

$$\begin{aligned}
\text{ND} &= \mathbb{E}[y_{1i}|d_i = 1] - \mathbb{E}[y_{0i}|d_i = 0] \\
&= \mathbb{E}[y_{1i}|d_i = 1] - \mathbb{E}[y_{0i}|d_i = 1] + \mathbb{E}[y_{0i}|d_i = 1] - \mathbb{E}[y_{0i}|d_i = 0] \\
&= ATT + \mathbb{E}[y_{0i}|d_i = 1] - \mathbb{E}[y_{0i}|d_i = 0]
\end{aligned}$$

For LRM, $y_i = \beta_0 + \beta_1 d_i + u_i$,

$$\begin{aligned}
\text{ND} &= \mathbb{E}[y_i|d_i = 1] - \mathbb{E}[y_i|d_i = 0] \\
&= \mathbb{E}[\beta_0 + \beta_1 + u_i|d_i = 1] - \mathbb{E}[\beta_0 + u_i|d_i = 0] \\
&= \beta_1 + \mathbb{E}[u_i|d_i = 1] - \mathbb{E}[u_i|d_i = 0]
\end{aligned}$$

Lecture 3.

Panel Data Analysis

Economists traditionally use the term **panel data** to refer to data structures consisting of observations on individuals for multiple time periods. There are several distinct advantages of panel data relative to cross-section data:

1. Possibility of controlling for unobserved time-invariant endogeneity without the use of instrumental variables
2. Possibility of allowing for broader forms of heterogeneity
3. Modeling dynamic relationships and effects

It's typical to index observations by both the individual i and the time period, t, thus y_{it} denotes a variable for individual i in time t, where $n = 1, \cdots, N$, $t = 1, \cdots, T$.

Definition 3.0.1 (Balanced and Unbalanced Panel Data[2]).

When observations are available on all individuals for the same time periods we say that the panel is **balanced**. In this case there are an equal number T of observations for each individual and the total number of observations is $n = NT$.

When different time periods are available for the individuals in the sample we say that the panel is **unbalanced**. This is the most common type of panel data set. It does not pose a problem for applications but does make the notation cumbersome and also complicates computer programming.

3.1 Incidental Parameters Problem

3.1.1 Pooled OLS Estimation

Suppose we are estimating the following panel data regression:

$$y_{it} = \alpha + x_{it}'\beta + u_{it}, \quad \mathbb{E}[u_{it}x_{it}] = 0, \quad \mathbb{V}[u_{it}|x_{it}] = \sigma^2$$

Omitting the distinction between intercept and slope, we can write the model as:

$$y_{it} = \tilde{x}_{it}'\tilde{\beta} + u_{it}$$
$$\tilde{x}_{it} = \begin{bmatrix} 1 \\ x_{it} \end{bmatrix}, \quad \tilde{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

where $i = 1 :_N$, $T = 1 : t$.

Or, we can write the model as:

$$\underset{T\times 1}{y_i} = \underset{T\times K}{\tilde{X}_i} \underset{K\times 1}{\tilde{\beta}} + \underset{T\times 1}{u_i}$$

Using OLS method to estimate $\tilde{\beta}$, we have:

$$\min_{\tilde{\beta}} \sum_i \sum_t u_{it}^2 = \min_{\tilde{\beta}} \sum_i u_i'u_i = \min_{\tilde{\beta}} (y_i - \tilde{X}_i\tilde{\beta})'(y_i - \tilde{X}_i\tilde{\beta})$$

The FOC of this equation is:

$$
\begin{aligned}
\sum_i -\tilde{X}_i'(y_i - \tilde{X}_i\tilde{\beta}) = 0 \\
\left(\sum_i \tilde{X}_i'\tilde{X}_i\right)\tilde{\beta} = \sum_i \tilde{X}_i' y_i \\
\hat{\tilde{\beta}}_{POLS} &= \left(\sum_i \tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_i \tilde{X}_i' y_i \\
&= \left(\sum_i\sum_t \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\left(\sum_i\sum_t \tilde{x}_{it}y_{it}\right) \\
&= \tilde{\beta} + \left(\frac{1}{NT}\sum_i\sum_t \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\frac{1}{NT}\left(\sum_i\sum_t \tilde{x}_{it}u_{it}\right) \\
&\xrightarrow{p} \tilde{\beta} + \mathbb{E}\left[\sum_t \tilde{x}_{it}\tilde{x}_{it}'\right]\mathbb{E}\left[\sum_t \tilde{x}_{it}u_{it}\right] \\
&= \tilde{\beta}
\end{aligned}
$$

Hence $\hat{\beta}_{OLS}$ is consistent provided that x_{it} and u_{it} are contemporaneously uncorrelated, as $\mathbb{E}[x_{it}u_{it}] = 0, \forall t$. The regressors are allowed to be correlated with the past, and future u_{it}. This occurs when there's feedback loop by which $y_{i,t-1}$ affects x_{it}.

In this proof, we show that either $N \to \infty$ or $T \to \infty$ is sufficient for consistency of $\hat{\beta}_{POLS}$. However, most panel data applications have a large N and small T dimension, so standard panel data features T fixed and $n \to \infty$.

3.1.2 Asymptotic Normality

From the analysis of consistency, we know that:

$$
\hat{\tilde{\beta}}_{POLS} = \left(\sum_i \tilde{X}_i'\tilde{X}_i\right)^{-1}\sum_i \tilde{X}_i' y_i
$$

Hence:

$$
\begin{aligned}
\sqrt{n}(\hat{\tilde{\beta}}_{POLS} - \tilde{\beta}) &= \left(\frac{1}{n}\sum_i \tilde{X}_i'\tilde{X}_i\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_i \tilde{X}_i' u_i\right) \\
&\xrightarrow{p} \mathbb{E}[\tilde{X}_i'\tilde{X}_i]^{-1} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\left(\tilde{X}_i'u_i\right)\left(\tilde{X}_i'u_i\right)'\right]\right) \\
&\xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\tilde{X}_i'\tilde{X}_i\right]^{-1}\mathbb{E}\left[\tilde{X}_i'u_iu_i'\tilde{X}_i\right]\mathbb{E}\left[\tilde{X}_i'\tilde{X}_i\right]^{-1}\right)
\end{aligned}
$$

The above model is homogeneous, which is unattractive, as the data generating process would differ across i, with some units having a higher level of the outcome variable y_{it} than others, regardless of covariates x_{it}(with a higher intercept α) or a stronger effect of some covariates $x_{it,k}$ on y_{it} than others.

At the other extreme, we assume the fully heterogenous estimation:

$$
y_{it} = \alpha_i + x_{it}'\beta + u_{it}, \quad \mathbb{E}[u_{it}x_{it}] = 0, \quad \mathbb{V}[u_{it}|x_{it}] = \sigma_i^2.
$$

Under $T = 1$, we run $y_i = \beta_0 + x_i'\beta + v_i$, where $v_i = u_i + \underbrace{\alpha_i - \beta_0}_{\tilde{\alpha}_i}$ and $\mathbb{E}[v_i] = 0$.

Under $T > 1$, we run:

$$\begin{aligned} y_i &= x_i'\beta + \sum_{j=1}^{n} \alpha_j \mathbf{1}\{i = j\} + u_{it} \\ &= \tilde{x}_{it}'\tilde{\beta} + u_{it} \end{aligned}$$

$$\tilde{x}_{it} = \begin{bmatrix} x_{it} \\ \mathbf{1}\{i=1\} \\ \mathbf{1}\{i=2\} \\ \vdots \\ \mathbf{1}\{i=n\} \end{bmatrix}, \quad \tilde{\beta} = \begin{bmatrix} \beta \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}$$

In a similar way, we can write the regression as

$$y_i = \tilde{X}_i \tilde{\beta}_i + u_i$$

with $\tilde{\beta}_i$ is specific for each i. We have n separate time series regressions, one for each unit i.

Following the same analyzing process, we can get:

$$\hat{\tilde{\beta}}_{i,OLS} = \left(\sum_i \tilde{X}_i' \tilde{X}_i\right) \sum_i \tilde{X}_i' y_i = \left(\sum_t \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1} \left(\sum_t \tilde{x}_{it} y_{it}\right),$$

which obviously shows that $\hat{\tilde{\beta}}$ is consistent $\Leftrightarrow T \to \infty$.

3.1.3 One-way error component model

With the fully homogeneous specification unattractive and the fully heterogeneous specifi- cation infeasible, researchers usually go for a compromise and let intercepts (and error term variances) be unit-specific.

Definition 3.1.1 (One-way error component model)**.**

$$y_{it} = \alpha_i + x_{it}' + u_{it}, \quad \mathbb{E}[u_{it} x_{it}] = 0, \quad \mathbb{V}[u_{it}|x_{it}] = \sigma^2, \tag{3.1}$$

where α_i is an individual-specific effect, and u_{it} are idiosyncratic(i.i.d.) errors.

In any case, the equation above makes clear that α_i contains all factors that affect y_{it}, that are not included in x_{it} and that are fixed over time (the time-varying factors are in u_{it}).

Suppose the model is correctly specified, and we have a cross-sectional dataset available, i.e. $T = 1$. Then, we would estimate:

$$y_{it} = \beta_0 + x_{it}'\beta + v_i, \text{ for } t = 1,$$

where $v_i = \alpha_i + u_{it} - \beta_0$.

If the unobserved heterogeneity α_i is correlated with the covariate x_{it}, our standard OLS estimator is biased and inconsistent.

If we have a panel dataset, i.e. $T > 1$, we can write the above model into a regression of $k + n$ regressors:

$$y_{it} = x_{it}'\beta + \sum_{j=1}^{N} \mathbf{1}\{i = j\}\alpha_j + u_{it} = x_{it}^{*'}\beta^* + u_{it},$$

where $x_{it}^* = \left(x_{it}', \mathbf{1}\{i=1\}, \cdots, \mathbf{1}\{i=N\}\right)'$, and $\beta^* = \left(\beta', \alpha_1, \cdots, \alpha_N\right)'$.

This leads to the pooled OLS estimator for β^*:

$$\hat{\beta}^* = \left(\sum_i \sum_t x^*_{it} x^{*'}_{it}\right) \sum_i \sum_t x^*_{it} y_{it}.$$

However, the estimator suffers from the so-called **IPP problem**, as the number of parameters increase with $N \to \infty$, the limit of $\frac{1}{N}\sum_i x^*_{it} x^{*'}_{it}$ is not well-defined and as a result, we can't establish consistency of $\hat{\beta}_{OLS}$.

3.2 Random Effects

As with pooled OLS, a random effects analysis puts α_i into the error term. In fact, random effects analysis imposes more assumptions than those needed for pooled OLS: **strict exogeneity** in addition to orthogonality between α_i and x_{it}.

3.2.1 Basic Assumptions and POLS

Stating the assumption in terms of conditional means, we have:

Assumption 3.2.1 (Random Effect).

(a) $\mathbb{E}[u_{it}|\tilde{X}_i, \tilde{\alpha}_i] = 0, \forall t.$

(b) $\mathbb{E}[\tilde{\alpha}_i|\tilde{X}_i] = \mathbb{E}[\tilde{\alpha}_i] = 0.$

where $\tilde{X}_i = (x_{i1}, \cdots, x_{iT})$.

Assumption 3.2.1(a) is the strict exogeneity condition and Assumption 3.2.1(b) is is how we will state the orthogonality.

Remark (Why Strict Ecogeneity?[3]).

Why do we maintain Assumption 3.2.1(a) when it is more restrictive than needed for a pooled OLS analysis? Because the random effects approach exploits the serial correlation in the composite error, $v_{it} = \alpha_i + u_{it}$, in a generalized least squares (GLS) framework. In order to ensure that feasible GLS is consistent, we need some form of strict exogeneity between the explanatory variables and the composite error.

Under this assumption, we can write:

$$\begin{aligned} y_{it} &= x'_{it}\beta + v_{it} \\ \mathbb{E}[v_{it}|X_i] &= 0, t = 1, \cdots, T \end{aligned}$$

The conditions shows that our model satisfies the GLS assumption, which confirms that we can apply GLS methods that account for the particular error structure $v_{it} = \alpha_i + u_{it}$.

By defining $v_{it} = u_{it} + \alpha_i - \beta_0$, we can transform the random effect model to the following:

$$\begin{aligned} y_{it} &= \alpha_i + x'_{it}\beta + u_{it} \\ &= \underbrace{\beta_0 + x'_{it}\beta}_{\tilde{x}'_{it}\tilde{\beta}} + \underbrace{u_{it} + \alpha_i - \beta_0}_{\equiv v_{it}} \end{aligned}$$

Defining again $\tilde{x}_{it} = (1, x'_{it})'$, $\tilde{\beta} = (\beta_0, \beta')'$, we can rewrite the model as:

$$
\begin{aligned}
y_{it} = \tilde{x}'_{it}\tilde{\beta} + v_{it} &\Leftrightarrow y_i = \tilde{X}'_i\tilde{\beta} + v_i \\
\rightarrow \hat{\tilde{\beta}} &= \left(\sum_i \tilde{X}'_i\tilde{X}_i\right)^{-1} \sum_i \tilde{X}'_i y_i
\end{aligned}
$$

With this intercept β_0, $\mathbb{E}[v_i] = 0$ is guaranteed to hold. Define $\tilde{\alpha}_i = \alpha_i - \beta_0$ as the mean-zero unit-specific heterogeneity so that $v_i = u_i + \tilde{\alpha}_i$.

> **Note** (POLS).
>
> Homogenous spec: $y_{it} = \alpha + x'_{it}\beta + u_{it} = \tilde{x}'_{it}\tilde{\beta} + v_{it}$. $\hat{\tilde{\beta}}$ is consistent if $\mathbb{E}[v_{it}x_{it}] = 0, \forall t$.

Using pooled OLS to estimate $\hat{\tilde{\beta}}$,

$$
\begin{aligned}
\hat{\tilde{\beta}}_{RE-OLS/POLS} &= \left(\frac{1}{n}\sum_i \tilde{X}'_i\tilde{X}_i\right)^{-1} \frac{1}{n}\sum_i \tilde{X}'_i y_i \\
&= \tilde{\beta} + \left(\frac{1}{n}\sum_i \tilde{X}'_i\tilde{X}_i\right)^{-1} \frac{1}{n}\sum_i \tilde{X}'_i v_i \\
&\xrightarrow{p} \tilde{\beta} + \mathbb{E}[\tilde{X}'_i\tilde{X}_i]^{-1}\mathbb{E}[\tilde{X}'_i v_i] \\
\text{where} \quad \mathbb{E}[\tilde{X}'_i v_i] &= \mathbb{E}\left[\sum_t \tilde{x}'_{it}v_{it}\right] \\
&= \sum_t \mathbb{E}\left[\tilde{x}'_{it}v_{it}\right] \\
&= \sum_t \mathbb{E}\left[\tilde{x}_{it}(u_{it} + \alpha_i - \beta_0)\right]
\end{aligned}
$$

Here, the error term v_i is not equal to the original error term u_{it}.

> **Note.**
>
> Under the random effect, you have to use the heteroskedasticity-robust methods. Because even if we assume u_{it} to be homoskedastic, v_{it} is not, as it includes also the unit-specific heterogeneity α_i.

3.2.2 From POLS to GLS

So, to obtain consistency, we need to assume that:

> **Assumption 3.2.2** (Random Effect Independence). (a) $\mathbb{E}[u_{it}|\tilde{x}_{it}, \tilde{\alpha}_i] = 0, \forall t$.
>
> (b) $\mathbb{E}[\tilde{\alpha}_i|\tilde{x}_{it}] = 0, \forall t$.

And, we are also obliged to use HAC-robust standard error because:

$$
\Omega \equiv \mathbb{E}[v_i v'_i|\tilde{X}_i] = \mathbb{E}[(\alpha_i\mathbf{1}_i + u_i)(\tilde{\alpha}_i\mathbf{1}_i + u_i)'|\tilde{X}_i] = \mathbb{E}[\tilde{\alpha}_i^2\mathbf{1}_i\mathbf{1}'_i|\tilde{X}_i] + \mathbb{E}[u_i u'_i|\tilde{X}_i]
$$

is not diagonal.

> **Assumption 3.2.3** (Random Effect Rank).
>
> $\operatorname{rank}\mathbb{E}\left[X'_i\Omega^{-1}X_i\right] = K$

We know that both GLS and feasible GLS estimator would be consistent under Assumption 3.2.1 and 3.2.3. A general FGLS analysis, using an unrestricted variance estimator Ω, is consistent and asymptotically normal as $N \to \infty$.

But, we won't exploit the unobserved effects structure v_{it}. A standard random effects analysis adds assumptions on the idiosyncratic errors that give Ω a special form. The first assumption is that the idiosyncratic errors u_{it} have a constant unconditional variance across t:

Assumption 3.2.4 (RE-Homoskedasticity).
$\mathbb{E}[u_{it}^2] = \sigma_u^2, \forall t$

The second assumption is that the idiosyncratic errors are serially uncorrelated:

Assumption 3.2.5 (RE-Serial Uncorrelated).
$\mathbb{E}[u_{it}u_{is}] = 0, \forall t \neq s$

Under these two assumptions, we can derive the variances and covariances of the elements of v_i. Given the error structure the natural estimator for β is GLS. The GLS eimator for β is:

$$\hat{\tilde{\beta}}_{RE-GLS} = \left(\sum_i \tilde{X}_i' \Omega^{-1} \tilde{X}_i\right)^{-1} \sum_i \tilde{X}_i' \Omega^{-1} y_i$$

where $\Omega^{-\frac{1}{2}} y_i = \Omega^{-\frac{1}{2}} \tilde{X}_i' \tilde{\beta} + \Omega^{-\frac{1}{2}} v_i$.

$$\begin{aligned}
\Omega = \mathbb{E}[v_i v_i' | \tilde{X}_i] &= \mathbb{E}\left[\begin{bmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{iT} \end{bmatrix} \begin{bmatrix} v_{i1} & v_{i2} & \cdots & v_{iT} \end{bmatrix} | \tilde{X}_i\right] \\
&= \mathbb{E}\begin{bmatrix} \mathbb{E}[v_{i1}^2|\tilde{X}_i] & \mathbb{E}[v_{i1}v_{i2}|\tilde{X}_i] & \cdots & \mathbb{E}[v_{i1}v_{iT}|\tilde{X}_i] \\ \mathbb{E}[v_{i2}v_{i1}|\tilde{X}_i] & \mathbb{E}[v_{i2}^2|\tilde{X}_i] & \cdots & \mathbb{E}[v_{i2}v_{iT}|\tilde{X}_i] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[v_{iT}v_{i1}|\tilde{X}_i] & \mathbb{E}[v_{iT}v_{i2}|\tilde{X}_i] & \cdots & \mathbb{E}[v_{iT}^2|\tilde{X}_i] \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{E}\left[\alpha_i^2|\tilde{X}_i\right] + \mathbb{E}[u_{i1}^2|\tilde{X}_i] & \mathbb{E}\left[\alpha_i^2|\tilde{X}_i\right] + \mathbb{E}[u_{i1}u_{i2}|\tilde{X}_i] & \cdots & \mathbb{E}\left[\alpha_i^2|\tilde{X}_i\right] + \mathbb{E}[u_{i1}u_{iT}|\tilde{X}_i] \\ \mathbb{E}\left[\alpha_i^2|\tilde{X}_i\right] + \mathbb{E}[u_{i2}u_{i1}|\tilde{X}_i] & \mathbb{E}\left[\alpha_i^2|\tilde{X}_i\right] + \mathbb{E}[u_{i2}^2|\tilde{X}_i] & \cdots & \mathbb{E}\left[\alpha_i^2|\tilde{X}_i\right] + \mathbb{E}[u_{i2}u_{iT}|\tilde{X}_i] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left[\alpha_i^2|\tilde{X}_i\right] + \mathbb{E}[u_{iT}u_{i1}|\tilde{X}_i] & \mathbb{E}\left[\alpha_i^2|\tilde{X}_i\right] + \mathbb{E}[u_{iT}u_{i2}|\tilde{X}_i] & \cdots & \mathbb{E}\left[\alpha_i^2|\tilde{X}_i\right] + \mathbb{E}[u_{iT}^2|\tilde{X}_i] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_u^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_u^2 + \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_u^2 + \sigma_\alpha^2 \end{bmatrix} \\
&= \sigma_\alpha^2 \mathbf{1}_i \mathbf{1}_i' + \sigma_u^2 I \\
\text{beacuse } \mathbb{V}[\tilde{\alpha}_i|\tilde{X}_i] &= \sigma_{\alpha_i}^2 = \sigma_\alpha^2 \\
\mathbb{V}[u_{it}|\tilde{X}_i] &= \sigma_u^2, \forall i.
\end{aligned}$$

where I is an identity matrix of dimention T_i. Under the assumption $\mathbb{E}[u_{it}x_{is}] = 0$, we now describe some statistical properties of $\hat{\tilde{\beta}}_{RE-GLS}$.

RE Consistency

$$\begin{aligned}
\hat{\tilde{\beta}}_{RE-GLS} - \tilde{\beta} &= \Big(\sum_i \tilde{X}_i'\Omega^{-1}\tilde{X}_i\Big)^{-1}\Big(\sum_i \tilde{X}_i'\Omega^{-1}v_i\Big) \\
&\to \mathbb{E}\Big[\sum_i \tilde{X}_i'\Omega^{-1}\tilde{X}_i\Big]\mathbb{E}\Big[\sum_i \tilde{X}_i'\Omega^{-1}v_i\Big] \\
\text{where } \mathbb{E}\Big[\sum_i \tilde{X}_i'\Omega^{-1}v_i\Big] &= \sum_i \mathbb{E}\Big[\tilde{X}_i'\Omega^{-1}v_i\Big] \\
&= \sum_i \tilde{X}_i'\Omega^{-1}\mathbb{E}[v_i|\tilde{X}_i] \\
&= \sum_i \tilde{X}_i'\Omega^{-1}\mathbb{E}[u_i + \tilde{\alpha}_i|\tilde{X}_i] \\
&= 0
\end{aligned}$$

Thus, $\hat{\tilde{\beta}}_{RE-GLS}$ is conditionally unbiased for $\tilde{\beta}$. The conditional variance of $\hat{\tilde{\beta}}_{RE-GLS}$ is:

$$\mathbb{V}\Big[\hat{\tilde{\beta}}_{RE-GLS}\Big] = \Big(\sum_i \tilde{X}_i'\Omega^{-1}\tilde{X}_i\Big)^{-1}\sigma_u^2$$

RE Asymptotic Distribution

The asymptotic variance of $\hat{\tilde{\beta}}_{RE-GLS}$ is:

$$\begin{aligned}
\sqrt{n}\left(\hat{\tilde{\beta}}_{RE-GLS} - \tilde{\beta}\right) &\xrightarrow{d} \mathcal{N}(0, V) \\
\text{where } V_{GLS} &= \mathbb{E}\left[\tilde{X}_i'\Omega^{-1}\tilde{X}_i\right]^{-1}\mathbb{E}\left[\tilde{X}_i'\Omega^{-1}v_iv_i'\Omega^{-1}\tilde{X}_i\right]\mathbb{E}\left[\tilde{X}_i'\Omega^{-1}\tilde{X}_i\right]^{-1} \\
&= \mathbb{E}\left[\tilde{X}_i'\Omega^{-1}\tilde{X}_i\right]^{-1}\underbrace{\mathbb{E}[v_iv_i'|\tilde{X}_i]}_{\equiv\Omega}
\end{aligned}$$

Because we do not know Ω, the RE-GLS estimator is infeasible.

The previous assumptions 3.2.4 and 3.2.5 are special to random eﬀects. For efficiency of feasible GLS, we assume that the variance matrix of v_i conditional on $\tilde{X}_i$ is constant:

$$\mathbb{E}[v_iv_i'|\tilde{X}_i] = \mathbb{E}[v_iv_i'].$$

The two conditions are also implied by the following stronger version of assumption:

Assumption 3.2.6 (RE General Homoskedasticity and Serial Uncorrelation)**.**

(a) $\mathbb{E}[u_iu_i'|\tilde{X}_i, \tilde{\alpha}_i] = 0$

(b) $\mathbb{E}[\tilde{\alpha}_i^2|\tilde{X}_i] = \sigma_\alpha^2$

Under assumption 3.2.6(a), $\mathbb{E}[u_{it}^2|\tilde{X}_i, \tilde{\alpha}_i] = \sigma_u^2$, which implies assumption 3.2.4, and $\mathbb{E}[u_{it}u_{is}|\tilde{X}_i, \tilde{\alpha}_i] = 0$ which implies assumption 3.2.5. But this new assumption is stronger because it implies that the conditional variances are constant and the conditional covariances are zero.

Together with assumption 3.2.2(b), assumption 3.2.6(b) is the same as: $\mathbb{V}[\tilde{\alpha}_i|\tilde{X}_i] = \mathbb{V}[\tilde{\alpha}_i]$, which is a homoskedasticity assumption for the unobserved effects $\tilde{\alpha}_i$.

A feasible version replaces Ω with an estimator $\hat{\Omega}$. To implement an FGLS procedure, define:

$$\sigma_v^2 = \sigma_u^2 + \sigma_\alpha^2$$

then we can obtain: $\hat{\Omega} = \hat{\sigma}_\alpha^2 \mathbf{1}_i \mathbf{1}_i' + \hat{\sigma}_u^2 I_T$, a $T \times T$ matrix that we assume to be positive definite. In a panel data context, the FGLS estimator that uses this variance matrix is what is known as the **random effects estimator**.

$$\hat{\beta}_{RE} = \left(\sum_i \tilde{X}_i' \hat{\Omega}^{-1} \tilde{X}_i\right)^{-1} \sum_i \tilde{X}_i' \hat{\Omega}^{-1} y_i$$

Hence, the motivation for using GLS is different than under a cross-sectional regression with heteroskedasticity. We use GLS because of the autocorrelation in v_{it} induced by the presence of time variant α_i.

3.2.3 Comparing POLS and GLS

Now, let's compare the $\hat{\beta}_{RE-GLS}$ with the pooled estimator $\hat{\beta}_{POLS}$.

Under the assumptions of the random effects model, POLS estimator is also unbiased for β and has conditional variance:

$$V_{POLS} = \left(\sum_i X_i' X_i\right)^{-1} \left(\sum_i X_i' \Omega X_i\right) \left(\sum_i X_i' X_i\right)^{-1}$$

Using the algebra of the Gauss-Markov Theorem we deduce that:

$$V_{RE-GLS} \leq V_{POLS}$$

and thus the random effects estimator $\hat{\beta}_{RE-GLS}$ is more efficient than $\hat{\beta}_{POLS}$ under the strict exogeneity assumption 3.2.1. The two variance matrices are identical when there is no individual-specific effect $\sigma_\alpha^2 = 0$ for then $V_{RE-GLS} = V_{POLS} = (X'X)^{-1} \sigma_u^2$.

Under the assumption that the random effects model is a useful approximation but not literally true, we may use the cluster-robust covariance matrix estimator such as:

$$\hat{V}_{RE-GLS} = \left(\sum_i X_i' \Omega^{-1} X_i\right)^{-1} \left(\sum_i X_i' \Omega^{-1} \hat{v}_i \hat{v}_i' \Omega^{-1} X_i\right) \left(\sum_i X_i' \Omega^{-1} X_i\right)^{-1}$$

where $\hat{v}_i = y_i - X_i \hat{\beta}_{RE-GLS}$, This may be re-scaled by a degree of freedom adjustment if desired.

3.3 Fixed Effects

In the econometrics literature if the stochastic structure of α_i is treated as unknown and possibly correlated with x_{it}, then α_i is called a **fixed effect**.

Correlation between α_i and x_{it} will cause both pooled and random effect estimators to be biased. This is due to the classic problems of omitted variables bias and endogeneity.

The presence of the unstructured individual effect α_i means that it is not possible to identify β under a simple projection assumption such as $\mathbb{E}[u_{it} x_{it}] = 0$. It turns out that a sufficient condition for identification is the following.

Definition 3.3.1 (Strictly exogeneity).

A regressor x_{it} is said to be strictly exogeneity if $\mathbb{E}[x_{it} u_{is}] = 0, \forall t, s = 1, \cdots, T$.

Strict exogeneity is a strong projection condition, meaning that is a $X_{is}, s \neq t$ is added into the regression model, it would have a zero coefficient. Strict exogeneity is a projection analog of the **strict mean independence**:

$$\mathbb{E}[u_{it}|X_i] = 0$$

which implies the strict exogeneity but not vice versa.

The strict exogeneity assumption 3.3.1 is sufficient for identification and asymptotic theory, we'll also use the strict mean independence assumption for finite sample analysis.

> **Remark** (About strict exogeneity[2]).
>
> Strict ecogeneity(assumption 3.3.1) is typically inappropriate in dynamic models.

3.3.1 Within Transformation

In previous steps, we showed that if x_{it} and α_i are correlated, then pooled OLS and RE-GLS estimator would be biased and inconsistent. If we leave the relationship between α_i and x_{it} fully unstructured, then the only way to consistently estimate the coefficient β is by an estimator which is invariant to α_i.

The first fixed effects (FE) assumption is strict exogeneity of the explanatory variables conditional on α_i:

> **Assumption 3.3.1** (FE Strict Exogeneity).
>
> $\mathbb{E}[u_{it}|X_i, \alpha_i] = 0, \forall t = 1, \cdots, T$

This assumption is identical to the assumption 3.2.1(a), we maintain strict exogeneity of $x_{it}, t = 1, \cdots, T$ conditional on the unobserved effect. The key difference is that *we do not assume assumption 3.2.1(b), which means that, for FE analysis, $\mathbb{E}[\alpha_i|X_i]$ can be any function of X_i.*

By relaxing assumption 3.2.1(b), we can conssitently estimate partial effects in the presence of time-consistent omitted variables that can be arbitrarily related to unobserved variables x_{it}. *Therefore, FE analysis is more robust than RE analysis.*

But this robustness has a cost: we can not include any time-constant variables in x_{it} without further assumptions. The reason is simple: if α_i can be arbitrarily correlated with each element of x_{it}, then there's no way to distinguish the effect of time-constant observables from the time-constant unobservable α_i.

We transform the equation to get rid of α_i: $y_{it} = \alpha_i + x_{it}'\beta + u_{it}$. The first transformation is the **within transformation**. Define the mean of a variable for a given individual as

$$\overline{y}_i = \frac{1}{T}\sum_t y_{it}$$
$$\overline{x}_i = \frac{1}{T}\sum_t x_{it}$$
$$\overline{u}_i = \frac{1}{T}\sum_t u_{it}$$

We call this the **individual-specific mean** since it is the mean of a given individual. [1]

Then, subtracting the individual-specific mean from the variable we obtain the deviations:

$$(y_{it} - \overline{y}_i) = (x_{it} - \overline{x}_i)'\beta + (u_{it} - \overline{u}_i) + (\alpha_i - \alpha_i)$$
$$\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{u}_{it} \text{ or at individual level: } \ddot{y}_i = \ddot{X}_i'\beta + \ddot{u}_i$$

[1]Some authors call this the **time-average** or **time-mean** since it is the average over the time periods.

This is the **within transformation**. We also refer to $\ddot{y}_{it}$ as the **demanded values** or **deviation from individual means**. *What is important is that the demeaning has occured at the individual level.*

Denote the time-averages method by $\hat{\beta}_{FE-W}$, in order to ensure that the FE estimator is consistent and well behaved asymptotically, we need a standard rank condition on the matrix of time-demeaned explanatory variables:

Assumption 3.3.2 (FE full rank).
$\text{rank} \sum_t \mathbb{E}[\ddot{x}_{it}' \ddot{x}_{it}] = \text{rank}\, \mathbb{E}[\ddot{X}_i' \ddot{X}_i] = K$

FE Consistency

$$
\begin{aligned}
\hat{\beta}_{FE-W} &= \left(\sum_i \ddot{X}_i' \ddot{X}_i\right)^{-1} \left(\sum_i \ddot{X}_i' \ddot{y}_i\right) \\
&= \left(\sum_i \sum_t \ddot{x}_{it} \ddot{x}_{it}'\right)^{-1} \sum_i \sum_t \ddot{x}_{it} \ddot{y}_{it} \\
&= \beta + \left(\sum_i \sum_t \ddot{x}_{it} \ddot{x}_{it}'\right)^{-1} \sum_i \sum_t \ddot{x}_{it} \ddot{u}_{it} \\
&\xrightarrow{p} \beta + \mathbb{E}\left[\sum_t \ddot{x}_{it} \ddot{x}_{it}'\right]^{-1} \mathbb{E}\left[\sum_t \ddot{x}_{it} \ddot{u}_{it}\right] \\
\text{where } \mathbb{E}\left[\sum_t \ddot{x}_{it} \ddot{u}_{it}\right] &= \sum_t \mathbb{E}\left[\ddot{x}_{it} \ddot{u}_{it}\right] \\
\mathbb{E}\left[\ddot{x}_{it} \ddot{u}_{it}\right] &= \mathbb{E}\left[\left(x_{it} - \frac{1}{T}\sum_t x_{it}\right)\left(u_{it} - \frac{1}{T}\sum_t u_{it}\right)'\right] \\
&= 0 \quad \text{if } u_{it} \perp\!\!\!\perp x_{is}, \forall t, s = 1, \cdots, T.
\end{aligned}
$$

Then, let $\Sigma_i = \mathbb{E}[u_i u_i' | X_i]$ denote the $T_i \times T_i$ covariance matrix of the idiosyncratic errors. The variance of $\hat{\beta}_{FE-W}$ is:

$$
V_{FE-W} = \mathbb{V}[\hat{\beta}_{FE-W} | X_i] = \left(\sum_i \ddot{X}_i' \ddot{X}_i\right)^{-1} \left(\sum_i \ddot{X}_i' \Sigma_i \ddot{X}_i\right) \left(\sum_i \ddot{X}_i' \ddot{X}_i\right)^{-1}
$$

This expression simplifies when the idiosyncratic errors are homoskedastic and serially uncorrelated:

Assumption 3.3.3 (FE homoskedasticity and Serial Uncorrelation).

(a) $\mathbb{E}[u_{it}^2 | X_i] = \sigma_u^2$

(b) $\mathbb{E}[u_{it} u_{is} | X_i] = 0, \forall s \neq t.$

FE Asymptotic Distribution

In this case, $\Sigma_i = \sigma_u^2 I_i$ and V_{FE-W} simplifies to:

$$
V_{FE-W}^0 = \sigma_u^2 \left(\sum_i \ddot{X}_i' \ddot{X}_i\right)^{-1}
$$

We can also write the asymptotic distribution as below

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{FE-W} - \beta) &= \left(\frac{1}{N}\sum_i \ddot{X}_i'\ddot{X}_i\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_i \ddot{X}_i'\ddot{u}_i\right) \\
&= \left(\frac{1}{N}\sum_i \ddot{X}_i'\ddot{X}_i\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_i \ddot{X}_i'u_i\right)^{2} \\
&\to \mathbb{E}[\ddot{X}_i'\ddot{X}_i]^{-1}\cdot\mathcal{N}(0,\mathbb{V}[\ddot{X}_i'u_i]) \\
&\sim \mathcal{N}\left(0, V_{FE-W}\right) \\
\text{where } V_{FE-W} &= \sigma_u^2\left(\sum_i \ddot{X}_i'\ddot{X}_i\right)^{-1}
\end{aligned}$$

Remark (About FE Asymptotic Distribution).

Actually, the asymptotic distribution of $\hat{\beta}_{FE-W}$ is not as obvious as it seems. We have to restress a few assumptions to guarantee its validity:

$$\frac{1}{\sqrt{N}}\sum_i \ddot{X}_i'u_i \xrightarrow{d} \mathcal{N}(0,\mathbb{V}[\ddot{X}_i'u_i]).$$

Assumption 3.3.4.

(1) Variables $(u_i, X_i), i = 1, \cdots, N$ are independent and identically dis- tributed.

(2) $\mathbb{E}[X_{it}u_{is}] = 0, \forall t = 1, \cdots, T$.

(3) $Q_T = \mathbb{E}[\ddot{X}_i'\ddot{X}_i] > 0$.

(4) $\mathbb{E}[u_{it}^4] < \infty$.

(5) $\mathbb{E}||X_{it}||^4 < \infty$.

Assumption 3.3.4(2) implies that:

$$\mathbb{E}[\ddot{X}_i u_i] = \sum_t \mathbb{E}[\ddot{X}_{it}u_{it}] = \sum_t \mathbb{E}[X_{it}u_{it}] - \sum_t\sum_{s=1}^{T}\mathbb{E}[X_{is}u_{it}] = 0$$

so they are mean zero. Assumption 3.3.4(4) and (5) imply that $\ddot{X}_i u_i$ has a finite covariance matrix Ω. The assumptions for the CLT hold, thus we have the result.

Remark (FE VS. POLS).

[2] From the regression model $\ddot{y}_i = \ddot{X}_i\beta + \ddot{u}_i$, where $\ddot{y}_i$ is $T\times 1$, $\ddot{X}_i$ is $T\times K$, and $\ddot{u}_i$ is $T\times 1$, We can write the individual-specific mean as $\bar{y}_i = (\mathbf{1}_i'\mathbf{1}_i)^{-1}\mathbf{1}_i y_i$. Then, we can define a **individual-specific demeaning operator**:

$$M_i = I_i - \mathbf{1}_i(\mathbf{1}_i'\mathbf{1}_i)^{-1}\mathbf{1}_i',$$

giving that

$$\ddot{y}_i = y_i - \mathbf{1}_i\bar{y}_i = y_i - \mathbf{1}_i(\mathbf{1}_i'\mathbf{1}_i)^{-1}\mathbf{1}_i y_i = M_i y_i.$$

Notice that M_i is idempotent ($M_iM_i = M_i, M_i' = M_i$). Similarly for $\ddot{X}_i$ and $\ddot{u}_i$. Thus, we have:

$$\ddot{X}_i'\ddot{u}_i = X_i'M_iM_iu_i = X_i'M_iu_i = \ddot{X}_i'u_i.$$

It is instructive to compare the variances of the fixed-effects and pooled estimators under

$$\mathbb{E}[u_{it}^2|X_i] = \sigma_u^2$$
$$\mathbb{E}[u_{it}u_{is}|X_i] = 0, \forall s \neq t.$$

and the assumption that there is no individual-specific effect, $\alpha_i = 0$. In this case, we can see that:

$$V_{FE-W}^0 = \sigma_u^2\Big(\sum_i \ddot{X}_i'\ddot{X}_i\Big)^{-1} \geq \sigma_u^2\Big(\sum_i X_i'X_i\Big)^{-1} = V_{POLS}.$$

The inequality holds since the demeaned variables $\ddot{X}_i$ have reduced variation compared to the original observations X_i.

This shows the cost of using fixed effects relative to pooled estimation. The estimation variance increases due to reduced variation in the regressors. This reduction in efficiency is a necessary by-product of the robustness of the estimator to the individual effects α_i.

3.3.2 First Difference Transformation

Another important transformation which does the same as within transformation is **first-differencing**. *This can be applied to all but the first observation (which is essentially lost).*

$$y_{it} - y_{i,t-1} = (x_{it} - x_{i,t-1})'\beta + (u_{it} - u_{i,t-1})$$
$$\Delta y_{it} = \Delta x_{it}'\beta + \Delta u_{it} \text{ or at individual level } \Delta y_i = \Delta X_i\beta + \Delta u_i, \quad i = 1\cdots N, t = 2\cdots T$$

We can see that the individual effect α_i has been eliminated.

Denote the first difference method by $\hat{\beta}_{FE-FD}$, the fixed effect estimator is consistent and asymptotically normal based on two assumptions.

Assumption 3.3.5 (FD Strict exogeneity).

It's the same as FE's assumption 3.3.1.

Assumption 3.3.6 (FD Full rank).

rank $\sum_{t=2}^T \mathbb{E}[\Delta x_{it}'\Delta x_{it}] = K$

FE-FD Consistency

$$\begin{aligned}
\hat{\beta}_{FE-FD} &= \left(\sum_i\sum_t \Delta x_{it}\Delta x_{it}'\right)^{-1} \sum_i\sum_t \Delta x_{it}\Delta y_{it} \\
&= \beta + \left(\frac{1}{NT}\sum_i\sum_t \Delta x_{it}\Delta x_{it}'\right)^{-1} \frac{1}{NT}\sum_i\sum_t \Delta x_{it}\Delta u_{it} \\
&\xrightarrow{p} \beta + \mathbb{E}\left[\sum_t \Delta x_{it}\Delta x_{it}'\right]^{-1} \mathbb{E}\left[\sum_t \Delta x_{it}\Delta u_{it}\right]
\end{aligned}$$

$$\text{where } \mathbb{E}\left[\sum_t \Delta x_{it}\Delta u_{it}\right] = \sum_t \mathbb{E}\left[\Delta x_{it}\Delta u_{it}\right]$$
$$\begin{aligned}
\mathbb{E}\left[\Delta x_{it}\Delta u_{it}\right] &= \mathbb{E}\left[(x_{it} - x_{i,t-1})(u_{it} - u_{i,t-1})'\right] \\
&= 0 \quad \text{if } x_{it} \perp\!\!\!\perp (u_{it}, u_{i,t-1}), \forall t.
\end{aligned}$$

For $T = 2$, $\hat{\beta}_{FE-FD} = \hat{\beta}_{FE-W}$, equals the fixed effects estimator and they differ however, for $T > 2$ (See Hanse, 2022[2]).

FE-FD Asymptotic Distribution

We just use the standard calculation:

$$\sqrt{n}(\hat{\beta}_{FE-FD} - \beta) \xrightarrow{d} \mathcal{N}(0, V_{FE-FD})$$

where

$$V_{FE-FD} = \mathbb{E}\Big[\sum_{t=2}^{T} \Delta x_{it} \Delta x'_{it}\Big]^{-1} \mathbb{E}\Big[\Big(\sum_{t=2}^{T} \Delta x_{it} \Delta u_{it}\Big)\Big(\sum_{s=2}^{T} \Delta x_{is} \Delta u_{is}\Big)'\Big] \mathbb{E}\Big[\sum_{t=2}^{T} \Delta x_{it} \Delta x'_{it}\Big]^{-1}$$

If we still assume that the first-difference error term Δu_{it} is homoskedastic:

Assumption 3.3.7 (FD homoskedasticity).

Denote $e_{it} \equiv \Delta u_{it}$, e_i is the stack of e_{it} for $t = 2, \cdots, T$. $\mathbb{E}\left[e_i e'_i | X_i, \alpha_i\right] = \sigma_e^2 I$.

then, we can write:

$$A\mathbb{V}[\hat{\beta}_{FE-FD}] = \hat{\sigma}_e^2 \Big(\sum_i \Delta X'_i \Delta X_i\Big)^{-1}$$

where $\hat{\sigma}_e^2$ is a consistent estimator of σ_e^2, and the simplest estimator is obtained by com- puting the OLS residuals:

$$\hat{e}_{it} = \Delta y_{it} - \Delta x_{it} \hat{\beta}_{FE-FD}$$

from the pooled regression.

If the assumption 3.3.7 is violated, replacing expectations with sample means and $\Delta u_{it}(e_{it})$ with $\widehat{\Delta u_{it}}(\hat{e}_{it})$ yields the HAC-robust variance estimator $\hat{V}_{FE-FD}$.

$$\hat{V}_{FE-FD} = \Big(\sum_i \Delta X'_i \Delta X_i\Big)^{-1} \Big(\sum_i \Delta X'_i \hat{e}_{it} \hat{e}'_{it} \Delta X_i\Big) \Big(\sum_i \Delta X'_i \Delta X_i\Big)^{-1}$$

Remark (About FE-W and FE-FD (Hansen, 2022 [2])).

The FD method is not as strong as the within method, because it only requires that the variable is uncorrelated with the error term in the same period and the previous period.

If there is a correlation between the error term in current period and two periods ago, there is a problem of feedback loop, which we will imply the correlated random effect model.

Matrix Notation for FE-FD

The first-differencing transformation is $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$. This can be applied to all but the first observation (which is essentially lost). At the level of the individual this can be written as:

$$\Delta Y_i = D_i Y_i$$

where D_i is the $(T_{i-1}) \times T_i$ difference matrix differencing operator:

$$D_i = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

Applying the transformation Δ to $Y_{it} = \alpha_i + X_{it}\beta + u_{it}$ or

$$\Delta Y_i = \Delta X_i \beta + \Delta u_i$$

We can see that the individual effect u_i has been eliminated.

Least squares applied to the differenced equation gives:

$$\begin{aligned} \hat{\beta}_\Delta &= \Big(\sum_i \sum_{t=2}^{T} \Delta X_{it}' \Delta X_{it}\Big)^{-1} \Big(\sum_i \sum_{t=2}^{T} \Delta X_{it}' \Delta Y_{it}\Big) \\ &= \Big(\sum_i \Delta X_i' \Delta X_i\Big)^{-1} \Big(\sum_i \Delta X_i' \Delta Y_i\Big) \\ &= \Big(\sum_i X_i' D_i' D_i X_i\Big)^{-1} \Big(\sum_i X_i' D_i' D_i Y_i\Big) \end{aligned}$$

is called the differenced estimator.

When the errors u_{it} are serially uncorrelated and homoskedastic, then the error $\Delta u_i = D_i u_i$ has the covariance matrix $H\sigma_u^2$, where

$$H = D_i D_i' = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & \ddots & 0 \\ 0 & \ddots & \ddots & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

We can reduce estimation variance by using GLS. When errors are i.i.d. (serially uncorrelated and homoskedastic), this is:

$$\begin{aligned} \tilde{\beta}_\Delta &= \Big(\sum_i \Delta X_i' H^{-1} \Delta X_i\Big)^{-1} \Big(\sum_i \Delta X_i' H^{-1} \Delta Y_i\Big) \\ &= \Big(\sum_i X_i' D_i' (D_i D_i') D_i X_i\Big)^{-1} \Big(\sum_i X_i' D_i' (D_i D_i') D_i Y_i\Big) \\ &= \Big(\sum_i X_i' M_i X_i\Big)^{-1} \Big(\sum_i X_i' M_i Y_i\Big) \end{aligned}$$

where $M_i = D_i'(D_i D_i')^{-1} D_i$. Recall that D_i is $(T_i - 1) \times T_i$ with rank $T_i - 1$ and is orthogonal to the vector of ones $\mathbf{1}_i$. This means that M_i projects orthogonally to $\mathbf{1}_i$ and thus equals the within transformation matrix. Hence $\tilde{\beta}_\Delta = \hat{\beta}_{FE-W}$.

What we have shown is that under i.i.d. errors, GLS applied to the first-differenced equation precisely equals the fixed effects estimator.

Since the Gauss-Markov theorem shows that GLS has lower variance than least squares, this means that the fixed effects estimator is more efficient than first differencing under the assumption that u_{it} is i.i.d.

3.3.3 Hausman Test for Random vs. Fixed Effects

Even if strict exogeneity is satisfied, the consistency of FE estimators comes at an efficiency loss compared to the RE-GLS and POLS estimators. This is easiest seen in the FD-transformation, in which we loose the n observations pertaining to the first time period $t = 1$.

The efficiency loss of the Within-estimator is somewhat more subtle. It arises because the Within-estimator only exploits variation across time and disregards the time-constant variation across cross-sectional units.

As a result, if the core RE assumption of X_i and α_i being uncorrelated is indeed satisfied, we prefer the RE-estimators. If instead it is violated, we of course prefer the less efficient but consistent FE estimators.

Theorem 3.3.1 (Hausman-Test).

$\mathcal{H}_0$: $\hat{\beta}_{RE-GLS} - \hat{\beta}_{FE-W} = 0$

We define:

$$T_{Hausman} = n\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)'\left(A\mathbb{V}[\hat{\beta}_{FE}] - A\mathbb{V}[\hat{\beta}_{RE}]\right)^{-1}\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \to \chi^2_k$$

If $\mathcal{H}_0$ is accepted, the difference between $\hat{\beta}_{RE-GLS}$ and $\hat{\beta}_{FE-W}$ is small enough to suggest that both estimators are consistent. that X_i and α_i are indeed uncorrelated, and therefore, we should use the more efficient estimator $\hat{\beta}_{RE-GLS}$.

If the test rejects this is evidence that the individual effect ui is correlated with the regressors so the random effects model is not appropriate.

Remark (Random Effects or Fixed Effects?(Hansen, 2022[2])).

We have presented the random effects and fixed effects estimators of the regression coefficients. Which should be used in practice? How should we view the difference?

The basic distinction is that the random effects estimator requires the individual error $\tilde{\alpha}_i$ to satisfy the conditional mean assumption $\mathbb{E}[\tilde{\alpha}_i|\tilde{X}_i] = 0$. The fixed effects estimator does not require this condition, and is robust to its violation.

In particular, the individual effect $\tilde{\alpha}_i$ can be arbitrarily correlated to the regressors. On the other hand the random effects estimator is efficient under random effects.

Current econometric practice is to prefer robustness over efficiency. Consequently, current practice is (nearly uniformly) to use the fixed effects estimator for linear panel data models. Random effects estimators are only used in contexts where fixed effects estimation is unknown or challenging (which occurs in many nonlinear models).

The labels "random effects" and "fixed effects" are misleading. These are labels which arose in the early literature and we are stuck with these labels today. In a previous era regressors were viewed as "fixed". Viewing the individual effect as an unobserved regressor leads to the label of the individual effect as "fixed". Today, we rarely refer to regressors as "fixed" when dealing with observational data. We view all variables as random. Consequently describing α_i as "fixed" does not make much sense and it is hardly a contrast with the "random effect" label since under either assumption α_i is treated as random. Once again, the labels are unfortunate but the key difference is whether α_i is correlated with the regressors.

3.3.4 FE-IV Estimation

1. Contemperaneous exogeneity: $\mathbb{E}[x_{it}u_{it}] = 0, \forall t$.

2. Strict exogeneity: $\mathbb{E}[x_{it}u_{is}] = 0, \forall t, s$.

3. Sequential exogeneity: $\mathbb{E}[x_{it}u_{is}] = 0, \forall t, s \geq t$.

Definition 3.3.2 (Predetermined variables(Or Sequantial Exogeneity)).

Predetermined variables are variables that were determined prior to the current period. In econometric models this implies that the current period error term is uncorrelated with current and lagged values of the predetermined variable but may be correlated with future values. This is a weaker restriction than strict exogeneity, which requires the variable to be uncorrelated with past, present, and future shocks.

The models we have discussed so far have been static with no dynamic relationships. In many economic contexts it is natural to expect that behavior and decisions are dynamic, explicitly depending on past behavior.

The workhorse dynamamic model in a panel framework is the p-th order autoregression with regressors and a one-way error component structure(see 3.1.3). This is:

$$y_{it} = \alpha_1 y_{i,t-1} + \cdots + \alpha_p y_{i,t-p} + x_{it}'\beta + \alpha_i + u_{it} \tag{3.2}$$

where α_j are the autoregressive coefficients. x_{it} is a k-vector of regressors, α_i is an individual effect and u_{it} is an idiosyncratic error. It's conventional to assume that u_{it} and α_i are mutuallt independent and the u_{it} are serially uncorrelated and mean zero. For the present we will assume that the regressors x_{it} are strictly exgenous(assumption 3.3.1). Currently, we focus on the AR(1) model:

$$y_{it} = \alpha_i + u_{it} + \beta_1 y_{i,t-1} + x_{it}'\beta_{-1}$$

where β_{-1} is a $k-1$ vector of coefficients on all other regressors.

Definition 3.3.3 (Anderson and Hsiao(1981)).

Anderson and Hsiao (1982) made an important breakthrough by showing that a simple instrumental variables estimator is consistent for the parameters of 3.2. he method first eliminates the individual effect α_i by first differencing:

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}'\beta + u_{it} \\ &= \alpha_i + \beta_1 y_{i,t-1} + \tilde{x}_{it}'\beta_{-1} + u_{it} \\ \Rightarrow \Delta y_{it} &= \beta_1 \Delta y_{i,t-1} + \Delta x_{it}'\beta + \Delta u_{it} \end{aligned}$$

The challenge is that first-differencing induces correlation between $\Delta y_{i,t-1}$ and Δu_{it}:

$$\mathbb{E}[\Delta y_{i,t-1}\Delta u_{it}] = \mathbb{E}\left[(y_{i,t-1} - y_{i,t-2})(u_{it} - u_{i,t-1})\right] = -\sigma_u^2.$$

The other regressors are not correlated with Δu_{it}. For $s > 1$, $\mathbb{E}[\Delta y_{i,t-s}\Delta u_{it}] = 0$ and x_{it} is strictly exogenous $\mathbb{E}[\Delta x_{it}\Delta u_{it}] = 0$. The correlation between $\Delta y_{i,t-1}$and Δu_{it} is endogeneity. One solution to endogeneity is to use an instrument. Anderson-Hsiao pointed out that $y_{i,t-2}$ is a valid instrument because it is correlated with $\Delta y_{i,t-1}$ yet uncorrelated with Δu_{it}. Under sequential exogeneity, instrument-exogeneity is satisied: $\mathbb{E}[y_{i,t-2}\Delta u_{it}] = \mathbb{E}[y_{i,t-2}\Delta u_{it}] - \mathbb{E}[y_{i,t-2}\Delta u_{it-1}] = 0$.

This is the IV usign the instruments $(y_{i,t-2}, \cdots, y_{i,t-s-1})$ for $(\Delta y_{i,t-1}, \cdots, \Delta y_{i,t-s})$. The estimator requires $T \geq s+2$.

$$\mathbb{E}[y_{is}\Delta u_{it}] = 0, \forall s \leq t-2.$$

The Anderson-Hsiao estimator is IV using $Y_{i,t-2}$ as an instrument for $\Delta Y_{i,t-1}$. Equivalently, this is IV using the instruments $(Y_{i,t-2},...,Y_{i,t-p-1})$ for $(\Delta Y_{i,t-1},...,\Delta Y_{i,t-p})$. The estimator requires $T \geq p+2$.

To show that this estimator is consistent, for simplicity assume we have a balanced panel with $T=3$, $p=1$, and no regressors. In this case the Anderson-Hsiao IV estimator is

$$\widehat{\alpha}_{\text{iv}} = \left(\sum_{i=1}^{N} Y_{i1}\Delta Y_{i2}\right)^{-1}\left(\sum_{i=1}^{N} Y_{i1}\Delta Y_{i3}\right) = \alpha + \left(\sum_{i=1}^{N} Y_{i1}\Delta Y_{i2}\right)^{-1}\left(\sum_{i=1}^{N} Y_{i1}\Delta \varepsilon_{i3}\right).$$

Under the assumption that ε_{it} is serially uncorrelated, (17.88) shows that $\mathbb{E}[Y_{i1}\Delta\varepsilon_{i3}]=0$. In general, $\mathbb{E}[Y_{i1}\Delta Y_{i2}] \neq 0$. As $N \to \infty$

$$\widehat{\alpha}_{\text{iv}} \underset{p}{\longrightarrow} \alpha - \frac{\mathbb{E}[Y_{i1}\Delta\varepsilon_{i3}]}{\mathbb{E}[Y_{i1}\Delta Y_{i2}]} = \alpha.$$

Thus the IV estimator is consistent for α.

The Anderson-Hsiao IV estimator relies on two critical assumptions. First, the validity of the instrument (uncorrelatedness with the equation error) relies on the assumption that the dynamics are correctly specified so that ε_{it} is serially uncorrelated. For example, many applications use an AR(1). If instead the true model is an AR(2) then Y_{it-2} is not a valid instrument and the IV estimates will be biased. Second, the relevance of the instrument (correlatedness with the endogenous regressor) requires $\mathbb{E}[Y_{i1}\Delta Y_{i2}] \neq 0$. This turns out to be problematic and is explored further in Section 17.40. These considerations suggest that the validity and accuracy of the estimator are likely to be sensitive to these unknown features.

Figure 3.1: Anderson and Hsiao(1981)

Using similar reasoning, other approaches use sequential exogeneity to circumvent FE methods altogether rather than to save their consistency. For example, Blundell and Bond (1998) start from the original specification:

$$y_{it} = x_{it}'\beta + \alpha_i + u_{it},$$

where correlation between α_i an x_{it} is suspected to be due to $y_{i,t-1}$, contained in x_{it}.

Definition 3.3.4 (Blundell and Bond(1998)).

$$\begin{aligned} y_{it} &= \alpha_i + \beta_1 y_{i,t-1} + u_{it} \\ &= \beta_1 y_{i,t-1} + (u_{it} + \alpha_i) \end{aligned}$$

Use $\Delta y_{i,t-1}$ as the IV for $y_{i,t-1}$

Lecture 4.

Time Series

4.1 Univariate Time Series

We have a sample $\{w_i\}_{i=1}^n$, with $w_i = (y_i, x_i')'$,

$\{w_{it}\}_{i=1:n,t=1:T}$.

Now, we look at $\{w_t\}_{i=1}^T$, usually written as y_t, is univariate time series data.

In the cross-sectional context, we average over i to get

$$\mathbb{E}[u_i] = \int u_i f_u(u_i) du_i.$$

Under tiem series data, we also think y_t as a RV. without i.i.d. assumption, we generallt have T realizations of different and mutually dependent variables.

$$\mathbb{E}[y_t] = \int y_t f_{y_t}(y_t) dy_t = \mu_t,$$
$$\mathbb{V}[y_t] = \mathbb{E}\Big[(y_t - \mu_t)^2\Big] = \gamma_{0,t},$$
$$\text{Cov}(y_t, y_{t-h}) = \mathbb{E}\Big[(y_t - \mu_t)(y_{t-h} - \mu_{t-h})\Big] = \gamma_{h,t}.$$

Definition 4.1.1 (Weak Stationarity).

y_t is a weakly stationary process if

1. $\mu_t = \mu$ for all t,
2. $\gamma_{h,t} = \gamma_h$ for all t.

autocovariance function (ACF): $\{\gamma_0, \gamma_1, \cdots\}$ autocorrelation function: $\{\rho_0, \rho_1, \cdots\}$, where $\rho_h = \frac{\gamma_h}{\gamma_0}$.

Appendix

Recommended Resources

Books

[1] Peng Ding. *A First Course in Causal Inference*. 2023. arXiv: 2305.18793 [stat.ME]. URL: https://arxiv.org/abs/2305.18793 (p. 8)

[2] Bruce E. Hansen. *Econometrics*. Princeton, New Jersey: Princeton University Press, 2022 (pp. 11, 19, 23, 25)

[3] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, Massachusetts: The MIT Press, 2010 (p. 14)

[4] James H. Stock and Mark W. Watson. *Introduction to Econometrics*. 4th ed. New York: Pearson, 2003

[5] Jeffrey M. Wooldridge. *Introductory Econometrics: A Modern Approach*. 7th ed. Cengage Learning, 2020

[6] Fumio Hayashi. *Econometrics*. Princeton, New Jersey: Princeton University Press, 2000

[7] Joshua Chan et al. *Bayesian Econometric Methods*. 2nd ed. Cambridge, United Kingdom: Cambridge University Press, 2019

[8] Badi H. Baltagi. *Econometric Analysis of Panel Data*. 6th ed. Cham, Switzerland: Springer, 2021

[9] James D. Hamilton. *Time Series Analysis*. Princeton, New Jersey: Princeton University Press, 1994. ISBN: 9780691042893

[10] Takeshi Amemiya. *Advanced Econometrics*. Cambridge, MA: Harvard University Press, 1985

Others

[11] Donald B. Rubin. "Bayesian Inference for Causality: The Importance of Randomization". In: *The Annals of Statistics* 3.1 (1975), pp. 121–131. DOI: 10.1214/aos/1176343238 (p. 8)

[12] Paul W. Holland. "Statistics and Causal Inference(with discussion)". In: *Journal of the American Statistical Association* 81.396 (1986), pp. 945–960. DOI: 10.1080/01621459.1986.10478373 (p. 8)

[13] Jerzy Neyman. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9". In: *Statistical Science* 5.4 (1923), pp. 465–472 (p. 8)

[14] Donald B. Rubin. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". In: *Journal of Educational Psychology* 66.5 (1974), pp. 688–701. DOI: 10.1037/h0037350 (p. 8)

[15] Donald B. Rubin. "Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test" by D. Basu". In: *Journal of the American Statistical Association* 75.371 (1980), pp. 591–593. DOI: 10.1080/01621459.1980.10477410 (pp. 8, 9)

[16] Donald B. Rubin. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions". In: *Journal of the American Statistical Association* 100.469 (2005), pp. 322–331. DOI: 10.1198/016214504000001880 (p. 9)

[17] Roger Bowden. "The Theory of Parametric Identification". In: *Econometrica* 41.6 (1973), pp. 1069–1074. DOI: 10.2307/1914036

[18] Robert I. Jennrich. "Asymptotic Properties of Non-linear Least Squares Estimators". In: *The Annals of Mathematical Statistics* 40.2 (1969), pp. 633–643. DOI: 10.1214/aoms/1177697731

[19] Michael P. Keane. "A Note on Identification in the Multinomial Probit Model". In: *Journal of Business & Economic Statistics* 10.2 (1992), pp. 193–200. DOI: 10.1080/07350015.1992.10509906

[20] Thomas J. Rothenberg. "Identification in Parametric Models". In: *Econometrica* 39.3 (1971), pp. 577–591. DOI: 10.2307/1913267

[21] George Tauchen. "Diagnostic Testing and Evaluation of Maximum Likelihood Models". In: *Journal of Econometrics* 30 (1985), pp. 415–443. DOI: 10.1016/0304-4076(85)90149-6

[22] Abraham Wald. "Note on the Consistency of the Maximum Likelihood Estimate". In: *The Annals of Mathematical Statistics* 20.4 (1949), pp. 595–601. DOI: 10.1214/aoms/1177729952

[23] Halbert White. "Maximum Likelihood Estimation of Misspecified Models". In: *Econometrica* 50.1 (1982), pp. 1–25. DOI: 10.2307/1912526