Regression

Author: Wenxiao Yang

Institute: Haas School of Business, University of California Berkeley

Date: 2023

All models are wrong, but some are useful.

Contents

Chapter 1 Linear Predictors / Regression 1

1.1 Best Linear Predictor . . . 1

1.2 Convergence of OLS . . . 2

1.2.1 Approximation . . . 2

1.2.2 Testing and Confidence Interval . . . 4

1.3 Long, Short, Auxilary Regression . . . 4

1.4 Residual Regression . . . 6

1.5 Card-Krueger Model . . . 7

1.5.1 Proxy Variable Regression . . . 8

1.6 Instrumental Variables . . . 9

1.6.1 Motivation . . . 9

1.6.2 I.V. Model . . . 10

1.6.3 Weak I.V. . . . 12

1.7 Linear Generalized Method of Moments (Linear GMM) . . . 12

1.7.1 Generalized Method of Moments (GMM) . . . 12

1.7.2 Linear GMM . . . 13

1.7.3 Properties of Linear GMM Estimator . . . 14

1.7.4 Alternative: Continuous Updating Estimator . . . 15

1.7.5 Inference . . . 15

1.7.6 OVER-ID Test . . . 16

1.7.7 Bootstrap GMM . . . 18

1.8 Panel Data Models . . . 18

1.8.1 Pooled OLS . . . 19

1.8.2 Fixed Effect Model . . . 20

1.8.3 Random Effect Model . . . 21

1.8.4 Two-Way Fixed Effect Model . . . 22

1.8.5 Arellano Bond Approach . . . 23

1.9 Control Function Approach (another approach to handle endogenieity) . . . 24

1.10 LATE (Local ATE): Application of I.V. on Potential Outcomes . . . 25

1.11 Difference in Difference (DiD) 27
1.11.1 After OLS Regression 27
1.11.2 Difference in Difference 27

Chapter 1 Linear Predictors / Regression

1.1 Best Linear Predictor

Consider a prediction problem that the distribution $F_{X,Y}$ is known, we observe $X = \begin{pmatrix} 1 \\ R \end{pmatrix} \in \mathbb{R}^{K\times 1}$ and predict $Y \in \mathbb{R}$. Only linear functions of X are allowed $\mathcal{L} = \{X'b : b \in \mathbb{R}^K\}$. We use square experience loss $(Y - X'b)^2$. We want to minimze Risk (mean squared error)

$$\mathbb{E}_{X,Y}[(Y - X'b)^2] = \int_{x,y} (y - x'b)^2 f_{x,y}(x,y)dxdy$$

Assumption *Following inference is based on assumptions:*

(i). $\mathbb{E}[Y^2] < \infty$;

(ii). $\mathbb{E}[||X||^2] < \infty$ *(Frobenius norm);*

(iii). $\mathbb{E}[(\alpha' X)^2] > 0$ *for any non-zero* $\alpha \in \mathbb{R}^K$.

Let $\beta_0 = \arg\min_{b\in\mathbb{R}^k} \mathbb{E}_{X,Y}[(Y - X'b)^2]$. By the F.O.C.

$$\begin{aligned} \mathbb{E}[X(Y - X'\beta_0)] &= 0 \\ \mathbb{E}[XY] - \mathbb{E}[XX']\beta_0 &= 0 \\ \mathbb{E}[XY] &= \underbrace{\mathbb{E}[XX']}_{non-singular} \beta_0 \\ \beta_0 &= \mathbb{E}[XX']^{-1}\mathbb{E}[XY] \end{aligned}$$

Proposition 1.1 (Best Linear Predictor)

Hence, the mean-squared error minimizing linear predictor of Y given X is

$$\mathbb{E}^*[Y|X] = X'\beta_0, \text{ where } \beta_0 = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$$

$$\mathbb{E}_{X,Y}[X(\underbrace{Y - X'\beta_0}_{\triangleq u})] = \begin{pmatrix} \mathbb{E}[u] \\ \mathbb{E}[uR] \end{pmatrix} = \mathbf{0}$$

Hence, we have $\mathbb{E}[u] = 0$, then $\mathbb{E}[uR] = 0 = \text{Cov}(u, R)$.

Lemma 1.1

$\mathbb{E}[u] = \mathbb{E}[uR] = \text{Cov}(u, R) = 0$, where $u = Y - \mathbb{E}^*[Y|X]$.

If $u > 0$, it is underpredicting and if $u < 0$, it is overpredicting.

Result 1 (ure Partitioned Inverse Formula)

When we separate the constant term from other variables, we can write the Best Linear Predictor as:

Proposition 1.2 (Best Linear Predictor (ure Partitioned Inverse Formula))

$X = \begin{pmatrix} 1 \\ R \end{pmatrix}$, $\beta_0 = \begin{pmatrix} \alpha_0 \\ \beta_* \end{pmatrix}$, $\mathbb{E}[XX']^{-1} = \begin{bmatrix} 1 & \mathbb{E}[R]' \\ \mathbb{E}[R] & \mathbb{E}[RR'] \end{bmatrix}^{-1}$, $\mathbb{E}[XY] = \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}[RY] \end{pmatrix}$. Then,

$$\alpha_0 = \mathbb{E}[Y] - \mathbb{E}[R]'\beta_*$$

$$\beta_* = \underbrace{\text{Var}(R)^{-1}}_{(K-1)\times(K-1)} \times \underbrace{\text{Cov}(R, Y)}_{(K-1)\times 1}$$

1.2 Convergence of OLS

1.2.1 Approximation

OLS Fit is

$$\hat{\beta} = \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right]$$

Theorem 1.1 (Weak Law of Large Numbers (wLLN))

The weak law of large numbers (also called Khinchin's law) states that the sample average converges in probability towards the expected value.

$$\overline{X}_n \xrightarrow{P} \mu \qquad \text{when } n \to \infty.$$

That is, for any positive number ε,

$$\lim_{n\to\infty} \Pr\left(|\overline{X}_n - \mu| < \varepsilon\right) = 1.$$

1. By LLN: $\frac{1}{N}\sum_{i=1}^{N} X_i Y_i \xrightarrow{P} \mathbb{E}[XY]$
2. By LLN and $f(X) = X^{-1}$ is continuous, $\left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right] \xrightarrow{P} \mathbb{E}[XX']^{-1}$
3. Hence,

$$\hat{\beta} = \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right] \xrightarrow{P} \mathbb{E}[XX']^{-1}\mathbb{E}[XY] = \beta_0$$

Theorem 1.2 (Central Limit Theorem (CLT))

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} N(0,1) \text{ when } n \to \infty$$

Z converges in distribution to $N(0,1)$ as $n \to \infty$

(converges in distribution: $P(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq a) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx$)

Application to OLS: Let $u = Y - X'\beta_0$. Then,

$$\begin{aligned}
\hat{\beta} &= \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right] \\
&= \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} X_i (u_i + X_i'\beta_0)\right] \\
&= \beta_0 + \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i u_i\right]
\end{aligned}$$

Then,

$$\sqrt{N}(\hat{\beta} - \beta_0) = \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i u_i\right]$$

1. By LLN, $\left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \xrightarrow{P} \mathbb{E}[XX']^{-1} \triangleq \Gamma_0^{-1}$.
2. By CLT, $\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i u_i\right] \sim \mathcal{N}(0, \Omega_0)$, where

$$\Omega_0 = Var[X_i u_i] = \mathbb{E}[\|X_i u_i\|^2] = \mathbb{E}[\|x_i\|^2 u_i^2] \leq \left(\mathbb{E}[\|x_i\|^4]\right)^{\frac{1}{2}} \mathbb{E}[u_i^4]^{\frac{1}{2}}$$

Hence,

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{D} N\left(0, \Gamma_0^{-1}\Omega_0\Gamma_0^{-1}\right)$$

The estimation of Γ_0 and Ω_0:

$$\begin{aligned}
\hat{\Gamma} &= \frac{1}{N}\sum_{i=1}^{N} X_i X_i' \\
\hat{\Omega} &= \frac{1}{N}\sum_{i=1}^{N} X_i \hat{u}_i \hat{u}_i' X_i', \quad \text{where } \hat{u}_i = Y_i - X_i'\hat{\beta}
\end{aligned}$$

We have

$$\hat{\Gamma}^{-1}\hat{\Omega}\hat{\Gamma}^{-1} \xrightarrow{P} \Gamma_0^{-1}\Omega_0\Gamma_0^{-1}$$

Then,

$$\hat{\beta} \xrightarrow{approx} N\left(\beta_0, \frac{\hat{\Gamma}^{-1}\hat{\Omega}\hat{\Gamma}^{-1}}{N}\right)$$

1.2.2 Testing and Confidence Interval

Let $\hat{\Lambda} = \hat{\Gamma}^{-1}\hat{\Omega}\hat{\Gamma}^{-1}$, $\Lambda = \Gamma_0^{-1}\Omega_0\Gamma_0^{-1}$, $\sqrt{N}(\hat{\beta}_k - \beta_k) \xrightarrow{D} N\left(0, \Lambda_{kk}\right)$. Hence,

$$T_N \triangleq \sqrt{N}\Lambda_{kk}^{-\frac{1}{2}}\left(\hat{\beta}_k - \beta_k\right) \xrightarrow{D} N(0,1)$$

Consider the event $A = \mathbf{1}\left\{|T_N| \leq 1.96\right\}$. We have

$$\Pr(A = 1) = \Phi(1.96) - \Phi(-1.96) = 0.95$$

Specifically,

$$\begin{aligned} A &= \mathbf{1}\left\{|T_N| \leq 1.96\right\} \\ &= \mathbf{1}\left\{\hat{\beta}_k - 1.96\frac{\Lambda_{kk}^{\frac{1}{2}}}{\sqrt{N}} \leq \beta_k \leq \hat{\beta}_k + 1.96\frac{\Lambda_{kk}^{\frac{1}{2}}}{\sqrt{N}}\right\} \end{aligned}$$

The "Random Interval" is

$$\left[\hat{\beta}_k - 1.96\frac{\Lambda_{kk}^{\frac{1}{2}}}{\sqrt{N}}, \hat{\beta}_k + 1.96\frac{\Lambda_{kk}^{\frac{1}{2}}}{\sqrt{N}}\right]$$

Testing Linear Restrictions

Let $\theta = H\beta$, where H is $p \times k$ and β is $k \times 1$.

$$H_0 : \theta = \theta_0; \quad H_1 : \theta \neq \theta_0$$

We have

$$\sqrt{N}(\hat{\theta} - \theta_0) = H\sqrt{N}\left(\hat{\beta} - \beta_0\right) \xrightarrow[H_0]{D} N(0, H\Lambda_0 H')$$

Moreover,

$$W_0 = N\left(\hat{\theta} - \theta_0\right)(H\Lambda_0 H')^{-1}\left(\hat{\theta} - \theta_0\right) \xrightarrow[H_0]{D} \chi_p^2$$

where $\mathbb{E}[\chi_p^2] = p$.

1.3 Long, Short, Auxilary Regression

$Y \in \mathbb{R}^1$, $X \in \mathbb{R}^K$, $K \in \mathbb{R}^J$. Consider a researcher interested in the conditional distribution of the logarithm of weekly wages ($Y \in \mathbb{R}^1$) given years of competed schooling ($X \in \mathbb{R}^K$) and vector of additional worker attributes. This vector could include variables such as age, childhood test scores, and race. Let W be this $J \times 1$ vector of additional variables.

We can run regression by two ways:

1. Long regression: $\mathbb{E}^*[Y|X, W] = X'\beta_0 + W'\gamma_0$.

2. Short regression: $\mathbb{E}^*[Y|X] = X'b_0$.

Proposition 1.3 (Long Regression)

Long regression is another form of best linear predictor.

$$\begin{aligned}\mathbb{E}^*[Y|X,W] &= \mathbb{E}^*[Y|Z] \\ &= Z'\left(\mathbb{E}[ZZ']^{-1}\mathbb{E}[ZY]\right) \\ &= X'\beta_0 + W'\gamma_0\end{aligned}$$

where $\begin{pmatrix}\beta_0 \\ \gamma_0\end{pmatrix} = \mathbb{E}[ZZ']^{-1}\mathbb{E}[ZY]$, $Z = \begin{pmatrix}X \\ W\end{pmatrix}$.

Proposition 1.4 (Auxiliary Regression)

$$\mathbb{E}^*[W|X] = \Pi_0 X$$

which is multivariate regression. For each row $j = 1, ..., J$,

$$\mathbb{E}^*[W_j|X] = X'\Pi_{j0}$$

where $\Pi_{j0} = \mathbb{E}[XX']^{-1}\mathbb{E}[XW_j]$ and $\Pi_0 = \begin{pmatrix}\Pi'_{10} \\ \vdots \\ \Pi'_{J0}\end{pmatrix} = \mathbb{E}[WX']\mathbb{E}[XX']^{-1}$.

Theorem 1.3 (Law of Iterated Linear Predictors (LILP))

$$\mathbb{E}^*[Y|X] = \mathbb{E}^*[\mathbb{E}^*[Y|X,W]|X]$$

Facts: Linear predictor is linear operator, $\mathbb{E}^*[X+Y|W] = \mathbb{E}^*[X|W] + \mathbb{E}^*[Y|W]$.

Let $Y = \mathbb{E}^*[Y|X,W] + u = X'\beta_0 + W'\gamma_0 + u$. Then,

$$\begin{aligned}\mathbb{E}^*[Y|X] &= \mathbb{E}^*[X'\beta_0 + W'\gamma_0 + u|X] \\ &= \mathbb{E}^*[X'\beta_0|X] + \mathbb{E}^*[W'\gamma_0|X] + \mathbb{E}^*[u|X] \\ &= X'\beta_0 + (\Pi_0 X)'\gamma_0 + 0 \\ &= X'\underbrace{(\beta_0 + \Pi_0'\gamma_0)}_{b_0}\end{aligned}$$

Proposition 1.5 (Short Regression)

$$\mathbb{E}^*[Y|X] = X'b_0$$

where $b_0 = \beta_0 + \Pi_0'\gamma_0$.

1.4 Residual Regression

Let the variation in W unexplained by X.

$$\underbrace{V}_{J\times 1} = \underbrace{W}_{J\times 1} - \underbrace{\mathbb{E}^*[W|X]}_{J\times 1} = W - \Pi_0 X$$

Proposition 1.6 (Residual Regression)

Let $\tilde{Y} = Y - \mathbb{E}^*[Y|X]$,

$$\mathbb{E}^*[\tilde{Y}|V] = V'\gamma_0$$

Proof 1.1

$$\begin{aligned} Y &= X'\beta_0 + W'\gamma_0 + u \\ \tilde{Y} &= X'\beta_0 - \mathbb{E}^*[Y|X] + W'\gamma_0 + u \\ &= -X'(\Pi_0'\gamma_0) + W'\gamma_0 + u \\ &= V'\gamma_0 + u \\ \mathbb{E}^*[\tilde{Y}|V] &= V'\gamma_0 \end{aligned}$$

By long regression,

$$\begin{aligned} \mathbb{E}^*[Y|X,W] &= X'\beta_0 + W'\gamma_0 \\ &= X'b_0 - X'(\Pi_0'\gamma_0) + W'\gamma_0 \\ &= X'b_0 + V'\gamma_0 \\ &= \mathbb{E}^*[Y|X] + \mathbb{E}^*[\tilde{Y}|V] \end{aligned}$$

Theorem 1.4 (Frisch-Waugh Theorem)

$$\begin{aligned} \mathbb{E}^*[Y|X,V] &= \mathbb{E}^*[Y|X] + \mathbb{E}^*[Y|V] - \mathbb{E}[Y] \\ &= \mathbb{E}^*[Y|X,W] \end{aligned}$$

Lemma 1.2

If $Cov(X,W)=0$, then

$$\mathbb{E}^*[Y|X,W]=\mathbb{E}^*[Y|X]+\mathbb{E}^*[Y|W]-\mathbb{E}[Y]$$

Proof 1.2

Let $u=Y-\mathbb{E}^*[Y|X,W]$.

$$\begin{aligned}0&=\mathbb{E}[uW]\\&=\mathbb{E}[(Y-\mathbb{E}^*[Y|X]-\mathbb{E}^*[Y|W]+\mathbb{E}[Y])W]\\&=\underbrace{\mathbb{E}[(Y-\mathbb{E}^*[Y|W])W]}_{=0\text{ by F.O.C.}}-\underbrace{\mathbb{E}[\mathbb{E}^*[Y|X]]}_{=\mathbb{E}[Y]}\mathbb{E}[W]+\mathbb{E}[Y]\mathbb{E}[W]\end{aligned}$$

1.5 Card-Krueger Model

Consider a model about log-learning based on schooling, ability, luck.

$$Y(s)=\alpha_0+\beta_0\underbrace{s}_{\text{schooling } s\in\mathbb{S}}+\underbrace{A}_{\text{ability}}+\underbrace{V}_{\text{luck}}$$

Given a cost function about s:

$$C(s)=\underbrace{C}_{\text{cost heterogeneity}}s+\frac{k_0}{2}s^2$$

Assumption *We assume*

1. *Information set $I_0=(C,A)$ are known by agent when choosing schooling.*
2. *V is independent of C,A: $V|C,A\triangleq V$.*

Then, the observed schooling s should satsify

$$\begin{aligned}s&=\arg\max_s\mathbb{E}[Y(s)-C(s)\mid I_0]\\&=\arg\max_s\alpha_0+\beta_0 s+A-Cs-\frac{k_0}{2}s^2\end{aligned}$$

By F.O.C.

$$\beta_0-C-k_0s=0\Rightarrow s=\frac{\beta_0-C}{k_0}$$

1. **Long Regression**:

$$\mathbb{E}^*[Y|s,A]=\alpha_0+\beta_0s+A \tag{LR}$$

2. Short Regression:

$$\mathbb{E}^*[Y|s]=a_0+b_0s$$

3. **Auxillary Regression**: By the best linear predictor, the $\mathbb{E}^*[A|s]$ can be written as

$$\begin{aligned}\mathbb{E}^*[A|s] &= \mathbb{E}[A] - \frac{\text{Cov}(A,s)}{\text{Var}(s)}\mathbb{E}[s] + \frac{\text{Cov}(A,s)}{\text{Var}(s)}s \\ &= \mathbb{E}[A] - \eta_0\mathbb{E}[s] + \eta_0 s\end{aligned} \tag{AR}$$

where $\eta_0 = \frac{\text{Cov}(A,s)}{\text{Var}(s)}$ and $s = \frac{\beta_0 - C}{k_0}$ and $\mathbb{E}[s] = \frac{\beta_0 - \mu_C}{k_0}$,

$$\text{Cov}(A,s) = \text{Cov}\left(A, \frac{\beta_0 - C}{k_0}\right) = -\frac{\text{Cov}(A,C)}{k_0} = -\frac{\sigma_{AC}}{k_0}$$

$$\text{Var}(s) = \text{Var}\left(\frac{\beta_0 - C}{k_0}\right) = \frac{\sigma_C^2}{k_0^2}$$

$$\eta_0 = -k_0\frac{\sigma_{AC}}{\sigma_C^2} = -k_0\frac{\sigma_{AC}}{\sigma_A\sigma_C}\frac{\sigma_A}{\sigma_C} = -k_0\rho_{AC}\frac{\sigma_A}{\sigma_C}$$

The Auxillary Regression is written as

$$\begin{aligned}\mathbb{E}^*[A|s] &= \mathbb{E}[A] + k_0\rho_{AC}\frac{\sigma_A}{\sigma_C}\frac{\beta_0 - \mu_C}{k_0} - k_0\rho_{AC}\frac{\sigma_A}{\sigma_C}s \\ &= \mathbb{E}[A] + \rho_{AC}\frac{\sigma_A}{\sigma_C}(\beta_0 - \mu_C) - k_0\rho_{AC}\frac{\sigma_A}{\sigma_C}s\end{aligned} \tag{AR-1}$$

Hence, the **Short Regression**

$$\begin{aligned}\mathbb{E}^*[Y|s] &= \mathbb{E}^*\left[\mathbb{E}^*[Y|s,A]|s\right] \\ &= \mathbb{E}^*\left[\alpha_0 + \beta_0 s + A|s\right] \\ &= \alpha_0 + \beta_0 s + \mathbb{E}^*[A|s] \\ &= \underbrace{\alpha_0 + \mathbb{E}[A] + \rho_{AC}\frac{\sigma_A}{\sigma_C}(\beta_0 - \mu_C)}_{a_0} + \underbrace{\left(\beta_0 - k_0\rho_{AC}\frac{\sigma_A}{\sigma_C}\right)}_{b_0}s\end{aligned} \tag{SR}$$

1.5.1 Proxy Variable Regression

What if we don't observe A or C. We observe some observed variables W (**proxy variable**) instead.

Assumption *We assume*

1. *Redundancy:* $\mathbb{E}^*[Y|s,A,W] = \mathbb{E}^*[Y|s,A]$ *(W doesn't give extra information).*
2. *Conditional Uncorrelatedness:* $\mathbb{E}^*[A|s,W] = \mathbb{E}^*[A|W] = \Pi_0 + W'\Pi_W$ *(Auxillary Regression).*
3. *Conditional Independence:* $C \perp A|W = w$.

The **Proxy Variable Regression** is given by

$$\begin{aligned}\mathbb{E}^*[Y|s,W] &= \mathbb{E}^*\left[\mathbb{E}^*[Y|s,A,W]|s,W\right] \\ &= \mathbb{E}^*\left[\mathbb{E}^*[Y|s,A]|s,W\right] \\ &= \mathbb{E}^*[\alpha_0 + \beta_0 s + A|s,W] \\ &= \alpha_0 + \beta_0 s + (\Pi_0 + W'\Pi_W) \\ &= (\alpha_0 + \Pi_0) + \beta_0 s + W'\Pi_W\end{aligned} \tag{PVR}$$

A general form of **Proxy Variable Regression** with

1. Long Regression: $\mathbb{E}^*[Y|X,A] = X'\beta_0 + A'\gamma_0$
2. Redundancy: $\mathbb{E}^*[Y|X,A,W] = \mathbb{E}^*[Y|X,A]$
3. Conditional Uncorrelatedness: $\mathbb{E}^*[A|X,W] = \mathbb{E}^*[A|W] = \Pi_0 W$

   where Π_0 is $P \times J$, W is $J \times 1$, and A is $P \times 1$.

$$\begin{aligned}
\mathbb{E}^*[Y|X,W] &= \mathbb{E}^*\left[\mathbb{E}^*[Y|X,A,W]|X,W\right] \\
&= \mathbb{E}^*\left[\mathbb{E}^*[Y|X,A]|X,W\right] \\
&= \mathbb{E}^*\left[X'\beta_0 + A'\gamma_0|X,W\right] \\
&= X'\beta_0 + \mathbb{E}^*[A|X,W]'\gamma_0 \\
&= X'\beta_0 + W'\Pi_0'\gamma_0
\end{aligned}$$

1.6 Instrumental Variables

1.6.1 Motivation

Suppose we want to estimate an OLS model $y = \beta^T x + e$, where $x \in \mathbb{R}^k$. The OLS estimator is given by

$$\hat{\beta}_{\text{OLS}} = \left(\frac{1}{m}\sum_{i=1}^m X_i X_i^T\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^m X_i Y_i\right)$$

which converges (in probability) to

$$\mathbb{E}_{P_0}[XX^T]^{-1}\mathbb{E}_{P_0}[XY] = \beta + \mathbb{E}_{P_0}[XX^T]^{-1} \underbrace{\mathbb{E}_{P_0}[Xe]}_{\text{assumed to be 0 (Exogeneity)}}$$

What if the exogeneity doesn't hold?

Example 1.1

1. $y = \beta x^* + e$, where $\mathbb{E}[x^* e] = 0$. However, we don't have x^* and we only have a noisy variable $x = x^* + v$ (with $\mathbb{E}[v] = 0$). Then, $y = \beta(x - v) + e = \beta x + \epsilon$, where $\epsilon := e - \beta v$. The probability limits of the OLS estimator satisfies

$$\hat{\beta}_{\text{OLS}} - \beta = \frac{\mathbb{E}_{P_0}[x\epsilon]}{\mathbb{E}_{P_0}[x^2]} = \frac{\mathbb{E}_{P_0}[(x^*+v)(e-\beta v)]}{\mathbb{E}_{P_0}[(x^*+v)^2]} = -\frac{\beta\mathbb{E}_{P_0}[v^2]}{\mathbb{E}_{P_0}[(x^*+v)^2]}$$

   Hence, it is impossible to let the estimator converge to the true β.

2. Returns to Schooling: Consider a model

$$\ln \text{Wage} = \beta_0 + \beta_1 \text{EDUC} + e$$

   Suppose the e is correlated to both the wage and the education. Given e is positively correlated to the education, the OLS estimator is over-estimating.

1.6.2 I.V. Model

Consider a model $Y = X^T\beta + e$, where $X \in \mathbb{R}^k$ and $\mathbb{E}_{P_0}[xe] \neq 0$.

Definition 1.1 (Instrumental Variable)

A variable $Z \in \mathbb{R}^l$ is an **instrumental variable** if it satisfies

(1). $\mathbb{E}_{P_0}[Ze] = 0$ (exogeneity).

(2). $\mathbb{E}_{P_0}[ZZ^T]$ is non-singular (tech).

(3). $\text{Rank}(\mathbb{E}_{P_0}(ZX^T)) = k$ (relevance), which requires $l \geq k$.

Remark Exogeneity implies "exclusion restriction", which means the Z can't directly affect Y without affecting X.

Implementation:

- Outcome Equation:

$$Y = X^T\beta + e$$

- 1^{st} Stage Equation (no economic meaning, just for mathematical use):

$$X = \Gamma^T Z + u$$

  where X and u are $k \times 1$, Γ are $l \times k$, and Z is $l \times 1$. $Z \perp u$ and $\Gamma = \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZX^T]$.

- Reduced Form Equation:

$$\begin{aligned} Y &= \beta^T X + e \\ &= \beta^T(\Gamma^T Z + u) + e \\ &= \lambda^T Z + v \end{aligned}$$

  where $\lambda = \Gamma\beta$ and $v = \beta^T u + e$.

  Note that $\mathbb{E}[Zv] = 0$, which satisfies exogeneity. Hence, we can use OLS to estimate λ.

Identification: Suppose λ and Γ are known, we want to recover β.

$$\lambda = \Gamma\beta$$

1. Case 1: $l = k$,

$$\beta = \Gamma^{-1}\lambda$$

   where Γ^{-1} exists by relevance.

2. Case 2: $l > k$,

$$\Gamma^T\lambda = (\Gamma^T\Gamma)\beta \Rightarrow \beta = (\Gamma^T\Gamma)^{-1}\Gamma^T\lambda$$

Estimation of Γ and λ:

(A). "Plug In"

(a). The estimation of Γ is given by

$$\hat{\Gamma} = \left(\frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^T\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i X_i^T\right) \tag{hG}$$

The OLS estimator of regressing X on Z should converge to Γ in probability.

(b). The estimation of λ is given by

$$\hat{\lambda} = \left(\frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^T\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i Y_i\right)$$

which converges to λ in probability.

(B). "2SLS"

The reduced form can also be written as

$$\begin{aligned} Y &= \beta^T X + e \\ &= \beta^T(\Gamma^T Z + u) + e \\ &= \beta^T \underbrace{(\Gamma^T Z)}_{W} + v \end{aligned} \tag{hl}$$

Assuming Γ is known, we can regress Y on W:

$$\begin{aligned} \tilde{\beta} &= \left(\frac{1}{m}\sum_{i=1}^{m} W_i W_i^T\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} W_i Y_i\right) \\ &= \left(\Gamma^T\left(\frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^T\right)\Gamma\right)^{-1}\Gamma^T\left(\frac{1}{m}\sum_{i=1}^{m} Z_i Y_i\right) \end{aligned}$$

Hence, we can estimate β based on

$$\hat{\beta}_{\text{2SLS}} = \left(\hat{\Gamma}^T\left(\frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^T\right)\hat{\Gamma}\right)^{-1}\hat{\Gamma}^T\left(\frac{1}{m}\sum_{i=1}^{m} Z_i Y_i\right)$$

where $\hat{\Gamma}$ is given by (1.1). Specifically, in the case of $l = k$, $\hat{\beta}_{\text{2SLS}} = \left(\frac{1}{m}\sum_{i=1}^{m} Z_i X_i^T\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i Y_i\right)$.

Remark Why not use the following steps?

(a). Regress X on Z to construct $\hat{W} := \hat{\Gamma}^T Z$.

(b). Regress Y on $\hat{W}$.

(Note that the mathematical foundation of OLS doesn't hold here because $\hat{W}$ is not i.i.d.)

1.6.3 Weak I.V.

The "relevance" of the IV doesn't hold: $\mathbb{E}[ZX^T] \approx 0$. Why this is a problem?

Let's begin with a simple case that $l = k = 1$. The 2SLS estimator is given by

$$\hat{\beta}_{\text{2SLS}} = \frac{\frac{1}{m}\sum_{i=1}^m Z_i Y_i}{\frac{1}{m}\sum_{i=1}^m Z_i X_i} = \beta + \frac{\frac{1}{m}\sum_{i=1}^m Z_i e_i}{\frac{1}{m}\sum_{i=1}^m Z_i X_i}$$

where the small $Z_i X_i$ may lead to a large bias.

Consider the $\mathbb{E}[ZX] = \frac{c}{\sqrt{m}}, c \neq 0$. Then, the 2SLS estimator can be written as

$$\hat{\beta}_{\text{2SLS}} = \beta + \frac{\frac{1}{m}\sum_{i=1}^m Z_i e_i}{\frac{c}{\sqrt{m}}\frac{1}{m}\sum_{i=1}^m Z_i^2 + \frac{1}{m}\sum_{i=1}^m Z_i v_i} = \beta + \frac{\frac{1}{\sqrt{m}}\sum_{i=1}^m Z_i e_i}{c\frac{1}{m}\sum_{i=1}^m Z_i^2 + \frac{1}{\sqrt{m}}\sum_{i=1}^m Z_i u_i}$$

where the $\lim_{m\to\infty} \frac{1}{\sqrt{m}}\sum_{i=1}^m Z_i e_i \sim \mathcal{N}(0, \sigma^2)$ and $\lim_{m\to\infty} \frac{1}{\sqrt{m}}\sum_{i=1}^m Z_i u_i \sim \mathcal{N}(0, r^2)$ by LLN, and $\frac{1}{m}\sum_{i=1}^m Z_i^2 \to 1 + 0_P(1)$ with normalized Z. Hence, As $m \to \infty$,

$$\hat{\beta}_{\text{2SLS}} \approx \beta + \frac{\mathcal{N}(0, \sigma^S)}{\mathcal{N}(c, r^2)}$$

which gives that $\hat{\beta}_{\text{2SLS}}$ is not good for nonzero $\mathbb{E}[ZX]$.

1.7 Linear Generalized Method of Moments (Linear GMM)

1.7.1 Generalized Method of Moments (GMM)

Assumption *GMM model assumes that, given the true probability of data P_0, there exists a unique parameter β such that*

$$\mathbb{E}_{P_0}[g(\text{Data}, \beta_0)] = 0$$

where $g(\cdot)$ is a residual function.

β_0 is given by

$$\beta_0 = \underset{\beta}{\text{argmin}}\, J(\beta, P_0)$$

where

$$J(\beta, P_0) := \left(\mathbb{E}_{P_0}[g(Y, X, Z, \beta)]\right)^T W \left(\mathbb{E}_{P_0}[g(Y, X, Z, \beta)]\right)$$

and the weight matrix $W \succ 0$ (is positive definite and symmetric).

The GMM estimator is given by

$$\hat{\beta}_{\text{GMM}} = \underset{\beta}{\text{argmin}}\, J(\beta, P_m)$$

Using this for

1. Linear Regression: $g(Y, X, \beta) := (Y - X^T\beta)X$;

2. IV Model: $g(Y, X, Z, \beta) = Z(Y - X^T\beta)$, which is called Linear GMM.

1.7.2 Linear GMM

Definition 1.2 (Linear GMM)

A **Linear GMM** is defined as

$$\mathbb{E}_{P_0}[\underbrace{Z}_{l\times 1}(\underbrace{Y}_{1\times 1} - \beta_0^T \underbrace{X}_{k\times 1})] = 0$$

If Rank $\left(\mathbb{E}_{P_0}[ZX^T]\right) = k$, there is a unique $\beta_0 =$ minimizes $J(\beta, P_0)$ with

$$J(\beta, P_0) := \left(\mathbb{E}_{P_0}[Z(Y - X^T\beta)]\right)^T W \left(\mathbb{E}_{P_0}[Z(Y - X^T\beta)]\right)$$

$$J(\hat{\beta}, P_0) := \left(\frac{1}{m}\sum_{i=1}^m Z_i(Y_i - X_i^T\beta)\right)^T W \left(\frac{1}{m}\sum_{i=1}^m Z_i(Y_i - X_i^T\beta)\right)$$

The GMM estimator is given by

$$\hat{\beta}_{\text{GMM}} = \underset{\beta}{\text{argmin}} \left(\frac{1}{m}\sum_{i=1}^m Z_i(Y_i - X_i^T\beta)\right)^T W \left(\frac{1}{m}\sum_{i=1}^m Z_i(Y_i - X_i^T\beta)\right) \tag{1.1}$$

Remark W matters for $\hat{\beta}_{\text{GMM}}$.

The FOC of (1.1) is given by

$$\left(\frac{1}{m}\sum_{i=1}^m Z_i X_i^T\right)^T W \left(\frac{1}{m}\sum_{i=1}^m Z_i Y_i - (\frac{1}{m}\sum_{i=1}^m Z_i X_i^T)\hat{\beta}_{\text{GMM}}\right) = 0$$

Let $\hat{Q} := \frac{1}{m}\sum_{i=1}^m Z_i X_i^T \in \mathbb{R}^{l\times k}$. Then,

$$\hat{\beta}_{\text{GMM}} = \left(\hat{Q}^T W \hat{Q}\right)^{-1} \hat{Q}^T W \frac{1}{m}\sum_{i=1}^m Z_i Y_i$$

Lemma 1.3

If $W = (\frac{1}{m}\sum_{i=1}^m Z_i Z_i^T)^{-1}$, then $\hat{\beta}_{\text{GMM}} = \hat{\beta}_{\text{2SLS}}$

Proof 1.3

With $W^T = W$,

$$\hat{\beta}_{\text{GMM}} = \left(\hat{Q}^T W \hat{Q}\right)^{-1} \hat{Q}^T W \frac{1}{m}\sum_{i=1}^m Z_i Y_i$$

$$\left(\hat{Q}^T W W^{-1} W \hat{Q}\right)^{-1} \hat{Q}^T W \frac{1}{m}\sum_{i=1}^m Z_i Y_i$$

$$= \left((W\hat{Q})^T W^{-1} (W\hat{Q})\right)^{-1} (W\hat{Q})^T \frac{1}{m}\sum_{i=1}^m Z_i Y_i$$

Substitute W by $W = (\frac{1}{m}\sum_{i=1}^m Z_i Z_i^T)^{-1}$. We have $W\hat{Q} = \hat{\Gamma}$. The lemma is proved.

1.7.3 Properties of Linear GMM Estimator

Theorem 1.5 (Asymptotic)

$\sqrt{m}\left(\hat{\beta}_{\text{GMM}} - \beta_0\right) \to \mathcal{N}(0, V_{P_0})$.

Proof 1.4

$$
\begin{aligned}
\hat{\beta}_{\text{GMM}} &= \left(\hat{Q}^T W \hat{Q}\right)^{-1} \hat{Q}^T W \frac{1}{m}\sum_{i=1}^m Z_i \underbrace{Y_i}_{X_i^T\beta_0 + e_i} \\
&= \left(\hat{Q}^T W \hat{Q}\right)^{-1} \hat{Q}^T W \left(\underbrace{(\frac{1}{m}\sum_{i=1}^m Z_i X_i^T)}_{\hat{Q}} \beta_0 + \frac{1}{m}\sum_{i=1}^m Z_i e_i \right) \\
&= \beta_0 + \left(\hat{Q}^T W \hat{Q}\right)^{-1} \hat{Q}^T W \frac{1}{m}\sum_{i=1}^m Z_i e_i
\end{aligned}
$$

By LLN, $\hat{Q} \xrightarrow{P} Q := \mathbb{E}[ZX^T]$. Then we have, $\hat{Q}^T W \hat{Q} \xrightarrow{P} Q^T W Q$. Because $Q^T W Q$ is invertible, $(\hat{Q}^T W \hat{Q})^{-1} \xrightarrow{P} (Q^T W Q)^{-1}$. So, $(\hat{Q}^T W \hat{Q})^{-1} = (Q^T W Q)^{-1} + o_{P_0}(1)$. Hence,

$$
\begin{aligned}
\hat{\beta}_{\text{GMM}} &= \beta_0 + \left((Q^T W Q)^{-1} + o_{P_0}(1)\right)\left(Q^T W + o_{P_0}(1)\right)\frac{1}{m}\sum_{i=1}^m Z_i e_i \\
&= \beta_0 + \left((Q^T W Q)^{-1} Q^T W + o_{P_0}(1)\right)\frac{1}{m}\sum_{i=1}^m Z_i e_i \\
&= \beta_0 + (Q^T W Q)^{-1} Q^T W \frac{1}{m}\sum_{i=1}^m Z_i e_i + o_{P_0}(1)\frac{1}{m}\sum_{i=1}^m Z_i e_i
\end{aligned}
$$

By orthogonality condition, $\mathbb{E}_{P_0}[Ze] = 0$. And by central limit theorem, we have $\sqrt{m}\frac{1}{m}\sum_{i=1}^m Z_i e_i \to \mathcal{N}(0, \Omega_{P_0})$. Then, we represent $\hat{\beta}_{\text{GMM}}$ as

$$
\hat{\beta}_{\text{GMM}} = \beta_0 + (Q^T W Q)^{-1} Q^T W \frac{1}{m}\sum_{i=1}^m Z_i e_i + o_{P_0}(\frac{1}{\sqrt{m}}) \tag{1.2}
$$

which is called **asymptotic linear representation**.

Multiplying $\sqrt{m}$,

$$
\begin{aligned}
\sqrt{m}(\hat{\beta}_{\text{GMM}} - \beta_0) &= (Q^T W Q)^{-1} Q^T W \underbrace{\frac{1}{\sqrt{m}}\sum_{i=1}^m Z_i e_i}_{\to \mathcal{N}(0, \Omega_{P_0})} + o_{P_0}(1) \\
&\to \mathcal{N}\left(0, \underbrace{(Q^T W Q)^{-1} Q^T W \Omega_{P_0} W Q (Q^T W Q)^{-1}}_{\triangleq V_{P_0}}\right)
\end{aligned}
$$

Corollary 1.1

$\hat{\beta}_{\text{GMM}} \xrightarrow{P} \beta_0$.

Proof 1.5

$\hat{\beta}_{\text{GMM}} - \beta_0 = O_{P_0}(\frac{1}{\sqrt{m}}) \to o_{P_0}(1)$.

Efficiency Consideration We want to choose the weight matrix to minimize the asymptotic variance within GMM estimator, $W^* = \operatorname{argmin}_W V_{P_0}$.

Theorem 1.6

$W^* = \Omega_{P_0}^{-1}$. That is, $V_{P_0}^* := \left(Q^T \Omega_{P_0}^{-1} Q\right)^{-1} \leq V_{P_0}, \forall W$.

Then, we want to compute the efficient GMM by $\Omega_{P_0} := \mathbb{E}[e^2 Z Z^T]$.

$$\hat{W}^* = \left(\hat{\Omega}\right)^{-1}$$

where $\hat{\Omega} = \frac{1}{m}\sum_{i=1}^m \hat{e}_i^2 Z Z^T$ and $\hat{e}_i$ is given by

$$\hat{e}_i := Y_i - X_i^T \hat{\beta}$$

where $\hat{\beta}$ can be any GMM estimator, e.g., $W = I$ or a 2SLS estimator. As long as we can make sure $\hat{\Omega} \xrightarrow{P} \Omega_{P_0}$. Finally, we have $\hat{\beta}_{\text{EFFI}} := \hat{W}^* = W^* + o_{P_0}(1)$,

$$\sqrt{m}\left(\hat{\beta}_{\text{EFFI}} - \beta_0\right) \to \mathcal{N}(0, \left(Q^T \Omega_{P_0}^{-1} Q\right)^{-1})$$

Remark If $\mathbb{E}_{P_0}[e^2|Z] = \sigma_e^2$, then 2SLS is efficient.

$$\Omega^{-1} = \left(\mathbb{E}_{P_0}[e^2 Z Z^T]\right)^{-1} = \frac{1}{\sigma_e^2} \underbrace{\left(\mathbb{E}_{P_0}[Z Z^T]\right)^{-1}}_{W \text{ used in 2SLS}}$$

1.7.4 Alternative: Continuous Updating Estimator

Based on the idea of efficiency, we may use

$$\hat{\beta}_{\text{CUE}} = \operatorname*{argmin}_{\beta} \left(\frac{1}{m}\sum_{i=1}^m g(\text{Data}_i, \beta)\right)^T \left(\frac{1}{m}\sum_{i=1}^m \hat{e}_i^2 Z Z^T\right) \left(\frac{1}{m}\sum_{i=1}^m g(\text{Data}_i, \beta)\right)$$

However, it may not be convex.

1.7.5 Inference

Suppose we want test $H_0 : \Gamma(\beta_0) = \theta_0 = 0$ or $H_0 : \theta_0 = \Gamma(\beta_0) \neq \hat{\theta} = \Gamma(\hat{\beta})$.

Theorem 1.7 (Construct Chi-square)

By using the asymptotic variance of GMM, V_{P_0},

$$m(\hat{\theta}-\theta)^T \left(R(\beta_0)^T V_{P_0} R(\beta_0)\right)^{-1} (\hat{\theta}-\theta) \Rightarrow \chi_l^2$$

where $R(\beta_0) := \frac{d\Gamma(\beta_0)}{d\beta} \in \mathbb{R}^{k\times l}$.

Proof 1.6

Let

$$\overbrace{m(\hat{\theta}-\theta)^T \underbrace{\left(R(\beta_0)^T V_{P_0} R(\beta_0)\right)^{-1}}_{\triangleq\Omega} (\hat{\theta}-\theta)}^{\mathcal{W}} \Rightarrow \chi_l^2$$

We have

$$\hat{\theta}-\theta_0 = \Gamma(\hat{\beta}) - \Gamma(\beta_0) = \underbrace{\frac{d\Gamma(\beta_0)}{d\beta}}_{R(\beta_0)}(\hat{\beta}-\beta_0) + o_{P_0}(m^{-\frac{1}{2}})$$

$$\mathcal{W} = \left(\sqrt{m}R(\beta_0)(\hat{\beta}-\beta_0) + o_{P_0}(1)\right)^T \Omega \left(\sqrt{m}R(\beta_0)(\hat{\beta}-\beta_0) + o_{P_0}(1)\right)$$

As $\sqrt{m}\left(\hat{\beta}-\beta_0\right) \Rightarrow \mathcal{N}(0, V_{P_0})$, by continuous mapping theorem, we have

$$\mathcal{W} \Rightarrow \left(\mathcal{N}(0, R(\beta_0)V_{P_0}R(\beta_0)^T)\right)^T \Omega \left(\mathcal{N}(0, R(\beta_0)V_{P_0}R(\beta_0)^T)\right)$$

Let $M := R(\beta_0)V_{P_0}R(\beta_0)^T$. Since M is symmetric, it can be decomposed by $M = LL^T$. Then, $M^{-1} = (L^T)^{-1}L^{-1}$. We have $L^{-1}M(L^T)^{-1} = I$.

Since $\Omega = M^{-1} = (L^{-1})^T L^{-1}$,

$$\mathcal{W} \Rightarrow \left(\mathcal{N}(0, I)\right)^T \left(\mathcal{N}(0, I)\right) = \chi_l^2$$

Based on this theorem, we have the "real" Wald test for $H_0 : \Gamma(\beta_0) = \theta_0 = 0$.

$$\mathcal{W} = m(\hat{\theta}-\theta)^T \left(R(\hat{\beta})^T \hat{V}_{P_0} R(\hat{\beta})\right)^{-1} (\hat{\theta}-\theta) \Rightarrow \chi_l^2$$

1.7.6 OVER-ID Test

Remind that

$$J(\beta, P_0) := \left(\mathbb{E}_{P_0}[Z(Y - X^T\beta)]\right)^T W \left(\mathbb{E}_{P_0}[Z(Y - X^T\beta)]\right)$$

We want to test

$$H_0 : J(\beta, P_0) = 0$$

which is equivalent to $\mathbb{E}[Ze] = 0$. $H_1 : J(\beta, P_0) > 0$, which is equivalent to $\mathbb{E}[Ze] \neq 0$.

Theorem 1.8

If W is efficient weighting matrix ($W = \hat{\Omega}^{-1}$), then $mJ(\hat{\beta}, P_m) \Rightarrow \chi^2_{l-k}$

Proof 1.7

Remind (1.2) that $\hat{\beta} = \beta_0 + (Q^T W Q)^{-1} Q^T W \frac{1}{m} \sum_{i=1}^m Z_i e_i + o_{P_0}(\frac{1}{\sqrt{m}})$ and $Q := \mathbb{E}[ZX^T]$. Then,

$$\begin{aligned} Z_i(Y_i - X_i^T \hat{\beta}) &= Z_i(X_i^T \beta_0 + e_i - X_i^T \hat{\beta}) \\ &= -Q(\hat{\beta} - \beta_0) + \frac{1}{m} \sum_{i=1}^m Z_i e_i + o_{P_0}(\frac{1}{\sqrt{m}}) \end{aligned}$$

which gives

$$\frac{1}{m} \sum_{i=1}^m Z_i(Y_i - X_i^T \hat{\beta}) = \left(I - Q(Q^T W Q)^{-1} Q^T W\right) \frac{1}{m} \sum_{i=1}^m Z_i e_i + o_{P_0}(\frac{1}{\sqrt{m}})$$

By decomposing W by $W := LL^T$,

$$mJ(\hat{\beta}, P_m) = \left(L^T \frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i(Y_i - X_i^T \hat{\beta})\right)^T \left(L^T \frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i(Y_i - X_i^T \hat{\beta})\right)$$

where

$$\begin{aligned} L^T \frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i(Y_i - X_i^T \hat{\beta}) &= \left(L^T - \underbrace{L^T Q}_{:=M}((L^T Q)^T (L^T Q))^{-1} (L^T Q)^T L^T\right) \frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i e_i + o_{P_0}(1) \\ &= \underbrace{(I - M(M^T M)^{-1} M^T)}_{:=R_M} \left(L^T \left(\frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i e_i\right)\right) + o_{P_0}(1) \end{aligned}$$

where R_M satisfies $R_M = R_M^T R_M$, which shows R_M has eigenvalues $\in \{0, 1\}$ and its number of eigenvalues equal to 1 is $l - k$.

Hence,

$$mJ(\hat{\beta}, P_m) = \left(L^T \left(\frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i e_i\right)\right)^T R_M \left(L^T \left(\frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i e_i\right)\right) + o_{P_0}(1)$$

As $\left(L^T \left(\frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i e_i\right)\right) \Rightarrow \xi \sim \mathcal{N}(0, L^T \Omega L)$. So,

$$mJ(\hat{\beta}, P_m) \Rightarrow \xi^T R_m \xi$$

If $W = \Omega^{-1}$, then $L^T \Omega L = I$, which gives

$$\begin{aligned} mJ(\hat{\beta}, P_m) &\Rightarrow \xi_*^T R_m \xi_*, \ \xi_* \sim \mathcal{N}(0, I) \\ &= \sum_{j=1}^{l-k} \omega_j^2, \omega_j \sim \mathcal{N}(0, 1) \\ &\sim \chi^2_{l-k} \end{aligned}$$

Remark

1. Test by c_α, which gives $\Pr(\chi^2_{l-k} \geq c_\alpha) = \alpha \in (0, 1)$.

2. Only make sense for $l > k$.
   (a). You "spent" k degrees of freedom estimating β_0.
   (b). The rest $(l - k)$ is "spent" on testing.

1.7.7 Bootstrap GMM

Now, we gives estimator by using bootstrap data,

$$\hat{\beta}^* = \underset{\beta}{\operatorname{argmin}}\, J(\beta, P_m^*)$$

where

$$J(\beta, P_m^*) := \left(\frac{1}{m}\sum_{i=1}^{m} Z_i^*(Y_i^* - X_i^{*T}\beta) - \mathbb{E}_{P_m}[Z(Y - X^T\hat{\beta})]\right)^T W \left(\frac{1}{m}\sum_{i=1}^{m} Z_i^*(Y_i^* - X_i^{*T}\beta) - \mathbb{E}_{P_m}[Z(Y - X^T\hat{\beta})]\right)$$

where $\mathbb{E}_{P_m}[Z(Y - X^T\hat{\beta})] = \frac{1}{m}\sum_{i=1}^{m} Z_i\hat{e}_i$, which is used to debias. Then,

$$\hat{\beta}_{\text{GMM}} = \left(\hat{Q}^{*T} W \hat{Q}^*\right)^{-1} \hat{Q}^{*T} W \left(\frac{1}{m}\sum_{i=1}^{m}(Z_i^* Y_i^* - Z_i\hat{e}_i)\right)$$

Bootstrap OVER-ID Test The distribution $mJ(\hat{\beta}^*, P_m^*)$ is the <u>same</u> as $mJ(\hat{\beta}, P_m)$ regardless of W.

1.8 Panel Data Models

Definition 1.3 (Panel Data)

For each unit i, it has time $\{1, ..., T\}$.

	$t = 1$
$i = 1$	$\vdots$
	$t = T$
	$t = 1$
$i = 2$	$\vdots$
	$t = T$
$\vdots$	$\vdots$

The typical model is given by

$$Y_{i_t} = \underbrace{\alpha_i}_{\text{Fixed Effect}} + X_{i_t}^T\beta + \epsilon_{i_t}$$

α_i is a fixed effect, which is unobserved, random, and time invariant.

Assumption

1. *$\{\alpha_i, (X_{i_t})_{t=1}^T, (Y_{i_t})_{t=1}^T, (\epsilon_{i_t})_{t=1}^T\}$ is i.i.d. for all $i \in \{1, ..., N\}$. (Within a unit, data at different time can be dependent, which means there are no estimators within units.)*
2. *$N \to \infty$, T is fixed.*

1.8.1 Pooled OLS

$$Y_{i_t} = X_{i_t}^T \beta_0 + \underbrace{e_{i_t}}_{:=\alpha_i + \epsilon_{i_t}}$$

Use the notations of vectors $\vec{Y}_i := \begin{bmatrix} Y_{i_1} \\ \vdots \\ Y_{i_T} \end{bmatrix}$, $\vec{X}_i := \begin{bmatrix} X_{i_1} \\ \vdots \\ X_{i_T} \end{bmatrix}$, $\vec{e}_i := \mathbf{1}\alpha_i + \vec{\epsilon}_i$, where $\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$. Then, the equation can be written as

$$\vec{Y}_i = \vec{X}_i \beta_0 + \vec{e}_i$$

The pooled OLS estimator is

$$\hat{\beta}_{\text{pool}} := \left(\frac{1}{N}\sum_{i=1}^N \vec{X}_i^T \vec{X}_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^N \vec{X}_i^T \vec{Y}_i\right)$$

Properties

$$\hat{\beta}_{\text{pool}} = \beta_0 + \left(\frac{1}{N}\sum_{i=1}^N \vec{X}_i^T \vec{X}_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^N \vec{X}_i^T \vec{e}_i\right)$$

For consistency:

1. $\frac{1}{N}\sum_{i=1}^N \vec{X}_i^T \vec{X}_i \xrightarrow{P} \mathbb{E}[\vec{X}^T \vec{X}]$, which is required to be non singular.
2. $\frac{1}{N}\sum_{i=1}^N \vec{X}_i^T \vec{e}_i \xrightarrow{P} \mathbb{E}[\vec{X}^T \vec{e}]$, where

$$\mathbb{E}[\vec{X}^T \vec{e}] = \underbrace{\mathbb{E}[\vec{X}^T \mathbf{1}\alpha]}_{\text{need assumed to be 0}} + \underbrace{\mathbb{E}[\vec{X}^T \vec{\epsilon}]}_{:=0\text{, by assumption}}$$

The pooled OLS estimator is inconsistent if X_{it} is correlated with α_i.

Assumption *X_{it} is uncorrelated with α_i, $\mathbb{E}[X_{it}\alpha_i] = 0$.*

Asymptotic Normality:

$$\sqrt{N}\left(\hat{\beta}_{\text{pool}} - \beta_0\right) = \underbrace{\left(\frac{1}{N}\sum_{i=1}^N \vec{X}_i^T \vec{X}_i\right)^{-1}}_{\mathbb{E}[\vec{X}^T\vec{X}]+o_{P_0}(1)} \underbrace{\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \vec{X}_i^T \vec{e}_i\right)}_{\text{by CLT:}\Rightarrow N(0,\mathbb{E}[\vec{X}^T\vec{e}\vec{e}^T\vec{X}])}$$

$$\Rightarrow N\left(0, \mathbb{E}[\vec{X}^T\vec{X}]^{-1}\mathbb{E}[\vec{X}^T\vec{e}\vec{e}^T\vec{X}]\mathbb{E}[\vec{X}^T\vec{X}]^{-1}\right)$$

where $\mathbb{E}[\vec{X}^T\vec{e}\vec{e}^T\vec{X}] = \vec{X}^T\mathbb{E}[\vec{e}\vec{e}^T \mid \vec{X}]\vec{X}$. Specifically, $\mathbb{E}[e_s e_t \mid \vec{X}] = \mathbb{E}[\alpha^2 + \epsilon_s\epsilon_t \mid \vec{X}] \neq 0, \forall s \neq t$. Hence, the variance of the normal distribution is not identical matrix. We need to compute the variance:

$$[\frac{1}{N}\sum_{i=1}^{N}\vec{X}_i^T\vec{X}_i]^{-1}[\frac{1}{N}\sum_{i=1}^{N}\vec{X}_i^T\hat{\vec{e}}_i\hat{\vec{e}}_i^T\vec{X}_i][\vec{X}_i^T\vec{X}_i]^{-1}$$

where $\hat{\vec{e}}_i = \vec{Y}_i - \vec{X}_i\hat{\beta}_{\text{pool}}$.

1.8.2 Fixed Effect Model

$$Y_{i_t} = \underbrace{\alpha_i}_{\text{Fixed Effect}} + X_{i_t}^T\beta + \epsilon_{i_t}$$

where is no assumption over α and $\vec{X}_i$.

"Naive" Time Difference (losing many data, inefficient):

$$\Delta Y_i = Y_{i_t} - Y_{i_{t-1}}, \text{ for some } t$$

$$\Delta Y_i = \Delta X_i\beta_0 + \Delta\epsilon_i$$

We get OLS estimator

$$\hat{\beta}_{\text{Diff}} = \frac{\sum_{i=1}^{n}\Delta X_i\Delta Y_i}{\sum_{i=1}^{n}\Delta X_i^2}$$

With assumptions $\mathbb{E}[X_t\epsilon_t] = \mathbb{E}[X_t\epsilon_{t-1}] = \mathbb{E}[X_{t-1}\epsilon_t] = \mathbb{E}[X_{t-1}\epsilon_{t-1}] = 0$, we have $\mathbb{E}[\Delta X\Delta\epsilon] = 0$, which gives the consistency.

Fixed Effect Estimator (most used): Let

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T}Y_{i_t} = \alpha_i + \bar{X}_i\beta + \bar{\epsilon}_i$$

"Dot" Model:

$$\dot{Y}_{i_t} = Y_{i_t} - \bar{Y}_i = \dot{X}_{i_t}\beta_0 + \dot{\epsilon}_{i_t}$$

Use the notations of vectors $\dot{\vec{Y}}_i := \begin{bmatrix}\dot{Y}_{i_1}\\ \vdots \\ \dot{Y}_{i_T}\end{bmatrix} = \vec{Y}_i - \mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T\vec{Y}_i =: Q\vec{Y}_i$, where $Q := I - \mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T$ (notice that $QQ = Q$).

Then, the equation $\dot{\vec{Y}}_i = \dot{\vec{X}}_i\beta_0 + \dot{\vec{\epsilon}}_i$ can be written as

$$Q\vec{Y}_i = Q\vec{X}_i\beta_0 + Q\vec{\epsilon}_i$$

Run OLS

$$\hat{\beta}_{FE} = \left(\frac{1}{N}\sum_{i=1}^{N}\vec{X}_i^TQ\vec{X}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\vec{X}_i^TQ\vec{Y}_i\right)$$

Assumption *We assume* $\mathbb{E}[\vec{X}^T Q\vec{\epsilon}] = 0$*, which is equivalent to* $\mathbb{E}[\vec{X}_i^T \vec{\epsilon}_i] = 0$*.*

Note *"Strict exogeneity" is sufficient for above assumption:* $\mathbb{E}[X_s \epsilon_t] = 0, \forall s, t$ *(*ϵ *is uncorrelated with past, present, and future* X*'s).*

Consistency:

$$\hat{\beta}_{FE} = \beta_0 + \left(\frac{1}{N}\sum_{i=1}^{N} \vec{X}_i^T Q \vec{X}_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \vec{X}_i^T Q \vec{\epsilon}_i\right)$$

The sufficient condition is $\mathbb{E}[\vec{X}^T Q\vec{\epsilon}] = 0$, that is the motivation of giving the above assumption.

Theorem 1.9

$$\sqrt{N}(\hat{\beta}_{FE} - \beta_0) \Rightarrow N\left(0, (\mathbb{E}[\vec{X}^T Q\vec{X}])^{-1}\mathbb{E}[\vec{X}^T Q\vec{\epsilon}\vec{\epsilon}^T Q\vec{X}](\mathbb{E}[\vec{X}^T Q\vec{X}])^{-1}\right)$$

Remark

1. Actually, all we want to do is constructing a matrix Q such that $Q\alpha_i = 0$, so that we can get rid of fixed effect. Another example of this kind of matrix is $D = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ & & & \cdots & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$.
2. Time invariant covariant? No.
3. Dummy interpretation:

$$Y_{i_t} = \gamma_1 D1_{i_t} + \gamma_2 D2_{i_t} + \vdots + \gamma_N DN_{i_t} + X_{i_t}\beta + \epsilon_{i_t}$$

   where $Dj_{i_t} = 1$ if $i = j$ and $Dj_{i_t} = 0$ if $i \neq j$.
4. Fixed effect can't be estimated.

1.8.3 Random Effect Model

(Based on many assumptions, but more efficient than fixed effect. However, still not suggested.)

Assumption α_i *is orthogonal to* X_{it}*,* $\mathrm{Cov}(\alpha_i X_{i_t}) = 0$.

$$Y_{i_t} = X_{i_t}\beta_0 + e_{i_t},\ e_{i_t} = \alpha_i + \epsilon_{i_t}$$

which can be written as the form of vector

$$\vec{Y}_i = \vec{X}_i\beta_0 + \vec{e}_i, \vec{e}_i = \alpha_i \mathbf{1} + \vec{\epsilon}_i \tag{1.3}$$

The R.E. estimator is the OLS estimator for (1.3). The pooled OLS estimator:

$$\sqrt{N}\left(\hat{\beta}_{\text{pool}} - \beta_0\right) \Rightarrow N\left(0, \mathbb{E}[\vec{X}^T\vec{X}]^{-1}\mathbb{E}[\vec{X}^T\vec{e}\vec{e}^T\vec{X}]\mathbb{E}[\vec{X}^T\vec{X}]^{-1}\right)$$

where $\mathbb{E}[\vec{X}^T\vec{e}\vec{e}^T\vec{X}] = \vec{X}^T\mathbb{E}[\vec{e}\vec{e}^T \mid \vec{X}]\vec{X}$. Specifically, $\mathbb{E}[e_s e_t \mid \vec{X}] = \mathbb{E}[\alpha^2 + \epsilon_s\epsilon_t \mid \vec{X}] \neq 0, \forall s \neq t$.

$$\begin{aligned}
\mathbb{E}[\vec{e}\vec{e}^T \mid \vec{X}] &= \mathbb{E}[(\alpha\mathbf{1} + \vec{\epsilon})(\alpha\mathbf{1} + \vec{\epsilon})^T \mid \vec{X}] \\
\text{(assuming } \alpha \perp \vec{\epsilon}) \quad &= \mathbb{E}[\alpha^2\mathbf{1}\mathbf{1}^T \mid \vec{X}] + \mathbb{E}[\vec{\epsilon}\vec{\epsilon}^T \mid \vec{X}] \\
\text{(assuming homoscedasticity and } \mathrm{Cov}(\epsilon_s, \epsilon_t) = 0) \quad &= \sigma_\alpha^2\mathbf{1}\mathbf{1}^T + \sigma_\epsilon^2 I \\
&:= \Omega
\end{aligned}$$

Given Ω (or $\hat{\Omega}$),

$$\hat{\beta}_{RE} = \left(\frac{1}{N}\sum_{i=1}^N \vec{X}_i^T\Omega^{-1}\vec{X}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N \vec{X}_i^T\Omega^{-1}\vec{Y}_i\right)$$

So,

$$\sqrt{N}\left(\hat{\beta}_{RE} - \beta_0\right) \Rightarrow N\left(0, \underbrace{(\mathbb{E}[\vec{X}^T\Omega^{-1}\vec{X}])^{-1}}_{V_{RE}}\right)$$

Hausmon Test We want to test $H_0 : \mathrm{Cov}(\alpha_i, X_{i_t}) = 0$. Under H_0:

$$\begin{aligned}
\sqrt{N}\left(\hat{\beta}_{RE} - \beta_0\right) &\Rightarrow N(0, V_{RE}) \\
\sqrt{N}\left(\hat{\beta}_{FE} - \beta_0\right) &\Rightarrow N(0, V_{FE}) \\
\text{where } V_{FE} &\geq V_{RE}
\end{aligned}$$

Theorem 1.10

Under H_0, $\hat{H} := N\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)^T(V_{FE} - V_{RE})^{-1}\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \Rightarrow \chi^2_{\dim(\beta_0)}$.

1.8.4 Two-Way Fixed Effect Model

In this model, we consider an extra "time fixed effect" V_t.

$$Y_{i_t} = \alpha_i + V_t + X_{i_t}\beta_0 + \epsilon_{i_t}$$

1. Principle of deleting fixed effect:

$$\dot{Y}_{i_t} = Y_{i_t} - \bar{Y}_i - \bar{Y}_t + \bar{Y}$$

   where $\bar{Y}_t := \frac{1}{N}\sum_{i=1}^N Y_{i_t}$ and $\bar{Y} := \frac{1}{NT}\sum_{t,i} Y_{it}$. Then,

$$\dot{Y}_{i_t} = \dot{X}_{i_t}\beta_0 + \dot{\epsilon}_{i_t}$$

   where $\dot{X}_{i_t}$ and $\dot{\epsilon}_{i_t}$ are given in the same way.

2. Hybrid Model (better?):

$$Y_{i_t} = \alpha_i + \gamma_2\delta 2_t + \gamma_3\delta 3_t + \cdots + \gamma_T\delta T_t + X_{i_t}\beta_0 + \epsilon_{i_t}$$

where $\delta s_t = \begin{cases} 1, & s = t \\ 0, & s \neq t \end{cases}$. Then, in the matrix form,

$$Y_{i_t} = \alpha_i + Z_{i_t}^T \Theta + \epsilon_{i_t}, \text{ where } Z_{i_t}^T = \begin{bmatrix} X \\ \delta 2 \\ \vdots \\ \delta T \end{bmatrix}$$

1.8.5 Arellano Bond Approach

1. "Strict exogeneity": $\mathbb{E}[X_s \epsilon_t] = 0, \forall s, t$ (ϵ is uncorrelated with past, present, and future X's).
2. "Sequential exogeneity": $\mathbb{E}[X_s \epsilon_t] = 0, \forall t \geq s$ (ϵ is uncorrelated with past X's).

Reminds that Fixed Effect model has assumption $\mathbb{E}[\vec{X}_i \vec{\epsilon}_i] = 0$, which can be given by "strict exogeneity". However, the assumption of "strict exogeneity" is too strong.

Example 1.2

$Y_{i_t} = \alpha_i + \rho \underbrace{Y_{i_{t-1}}}_{X_{i_t}} + \epsilon_{i_t}$, which doesn't satisfy the "strict exogeneity": $\mathbb{E}[X_{i_{t+1}} \epsilon_{i_t}] = \mathbb{E}[Y_{i_t} \epsilon_{i_t}] \neq 0$.

Instead of using the "strict exogeneity" assumption, <u>we can use "sequential exogeneity" assumption</u>.

Consider model

$$\Delta Y_{i_t} = \Delta X_{i_t} \beta_0 + \Delta \epsilon_{i_t}$$

we have

$$\mathbb{E}[X_s(\Delta \epsilon_t)] = \underbrace{\mathbb{E}[X_s \epsilon_t]}_{=0, \forall s \leq t} - \underbrace{\mathbb{E}[X_s \epsilon_{t-1}]}_{=0, \forall s \leq t-1}$$

Moreover, we suppose $\mathbb{E}[X_s \Delta X_t] \neq 0$, then $\{X_s, s \leq t-1\}$ are I.V. for the model above!

$\mathbb{E}[X_s (\Delta Y_t - \Delta X_t \beta_0)] = 0, \forall t, s : s \leq t-1$.

$t = 2$	$\mathbb{E}[X_1 (\Delta Y_2 - \Delta X_2 \beta_0)]$
$t = 3$	$\mathbb{E}[X_1 (\Delta Y_3 - \Delta X_3 \beta_0)]$
	$\mathbb{E}[X_2 (\Delta Y_3 - \Delta X_3 \beta_0)]$
$\vdots$	$\vdots$

All in all, we have

$$\mathbb{E}[g(\Delta\vec{Y}, \Delta\vec{X}, \vec{X}, \beta_0)] = \begin{bmatrix} \mathbb{E}[X_1 (\Delta Y_2 - \Delta X_2 \beta_0)] \\ \mathbb{E}[X_1 (\Delta Y_3 - \Delta X_3 \beta_0)] \\ \mathbb{E}[X_2 (\Delta Y_3 - \Delta X_3 \beta_0)] \\ \vdots \end{bmatrix} = 0$$

We can use GMM to estimate the parameters:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} g(\Delta\vec{Y}_i, \Delta\vec{X}_i, \vec{X}_i, \beta_0) \right)^T W \left(\frac{1}{N} \sum_{i=1}^{N} g(\Delta\vec{Y}_i, \Delta\vec{X}_i, \vec{X}_i, \beta_0) \right)$$

Arellano Bond estimator is GMM estimator over I.D.

1.9 Control Function Approach (another approach to handle endogenieity)

Another approach to handle endogenieity.

Suppose we are facing the problem of endogenieity that

$$Y_i = X_i \beta_i + U_i,\ \mathbb{E}[U|X] \neq 0$$

Suppose W is a variable that

$$\mathbb{E}[U|X, W] = \varphi(W)$$

which is only a function of W. That is, the relationship between X and U can only be determined by W: $X \to W \to U$.

Definition 1.4 (Control Variable)

W is a **Control Variable**.

A control variable doesn't have to be an I.V.

Example 1.3

$X = Z\gamma + V$, where Z is I.V. that $\mathbb{E}[ZU] = 0$. $\mathbb{E}[U|X, V] = \varphi(V)$.

Based on the control variable, we can write the regression as

$$Y_i = X_i \beta_0 + \gamma W_i + U_i$$

$$Y_i = X_i \beta_0 + \gamma W_i + \varphi(W_i) + \underbrace{U_i - \varphi(W_i)}_{\xi_i}$$

where $\mathbb{E}[\xi_i | X_i, W_i] = 0$.

To implement this, we can decompose $\varphi(W_i) := \sum_{l=1}^{L} \pi_l \phi_l(W_i)$ (e.g. polynomial).

Note *We may get inconsistent* γ.

Example 1.4

Suppose $\varphi(W) = \Pi W$, then $Y_i = X_i\beta_0 + \underbrace{(\gamma+\Pi)}_{\beta_1} W_i + \xi_i$. Hence, in OLS, $\hat{\beta}_0 \xrightarrow{P} \beta_0$ and $\hat{\beta}_1 \xrightarrow{P} \beta_1 = \gamma + \Pi$.

1.10 LATE (Local ATE): Application of I.V. on Potential Outcomes

(Application of I.V.)

Consider the potential outcome framework: $X \in \{0,1\}, Y(0), Y(1) : Y := XY(1) + (1-X)Y(0)$.

The Average treatment effect (ATE) is

$$ATE = \mathbb{E}[Y(1) - Y(0)]$$

Consider another variable $Z \in \{0,1\}$.

1. X: the assigned treatment of an agent.
2. Z: the intended treatment of an agent. (instrument)

Suppose $X(Z)$ be the potential treatment status $X(0), X(1)$. $X = ZX(1) + (1-Z)X(0)$.

Example 1.5

Some people are suggested to stay at home, but they don't.

We have $Z \to X \to Y$ and Z doesn't have a direct effect on Y.

There are four possible cases:

1. Never Treated (NT): $X(0) = X(1) = 0$.
2. Always Treated (AT): $X(0) = X(1) = 1$.
3. Complies (C): $X(0) = 0, X(1) = 1$.
4. Defiers (D): $X(0) = 1, X(1) = 0$.

Usually, we assume the instruments are relevant and rule out the defiers.

Assumption $X_i(0) \le X_i(1), \forall i$ *and* $X_j(0) < X_j(1)$ *for some* j.

$\hat{\beta}_{2SLS} = \frac{\hat{\mathrm{Cov}}(Y,Z)}{\hat{\mathrm{Cov}}(X,Z)} \xrightarrow{P} \frac{\mathrm{Cov}(Y,Z)}{\mathrm{Cov}(X,Z)}$

Theorem 1.11

$\frac{\mathrm{Cov}(Y,Z)}{\mathrm{Cov}(X,Z)} = \frac{\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1]-\mathbb{E}[X|Z=0]}$

Proof 1.8

$$
\begin{aligned}
\mathrm{Cov}(Y, Z) &= \mathbb{E}[YZ] - \mathbb{E}[Y]P(Z=1) \\
&= \mathbb{E}[Y|Z=1]P(Z=1) - (\mathbb{E}[Y|Z=1]P(Z=1) + \mathbb{E}[Y|Z=0]P(Z=0))P(Z=1) \\
&= P(Z=1)\left(\mathbb{E}[Y|Z=1](1-P(Z=1)) - \mathbb{E}[Y|Z=0]P(Z=0)\right) \\
&= P(Z=1)P(Z=0)\left(\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]\right)
\end{aligned}
$$

Similarly,

$$
\mathrm{Cov}(X, Z) = P(Z=1)P(Z=0)\left(\mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0]\right)
$$

Since we rule out the possible of (D), we can write

$$
\begin{aligned}
&\mathbb{E}[Y|Z=1] \\
=&\mathbb{E}[Y|AT, Z=1]\mathrm{Pr}(AT|Z=1) + \mathbb{E}[Y|NT, Z=1]\mathrm{Pr}(NT|Z=1) + \mathbb{E}[Y|C, Z=1]\mathrm{Pr}(C|Z=1) \\
=&\mathbb{E}[Y(1)|AT]\mathrm{Pr}(AT) + \mathbb{E}[Y(0)|NT]\mathrm{Pr}(NT) + \mathbb{E}[Y(1)|C]\mathrm{Pr}(C)
\end{aligned}
$$

We can also decompose the $\mathbb{E}[Y|Z=1]$.

$$
\begin{cases}
\mathbb{E}[Y|Z=1] & = \mathbb{E}[Y(1)|AT]\mathrm{Pr}(AT) + \mathbb{E}[Y(0)|NT]\mathrm{Pr}(NT) + \mathbb{E}[Y(1)|C]\mathrm{Pr}(C) \\
\mathbb{E}[Y|Z=0] & = \mathbb{E}[Y(1)|AT]\mathrm{Pr}(AT) + \mathbb{E}[Y(0)|NT]\mathrm{Pr}(NT) + \mathbb{E}[Y(0)|C]\mathrm{Pr}(C)
\end{cases}
$$

Then, we have

$$
\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] = \mathrm{Pr}(C)\left(\mathbb{E}[Y(1)|C] - \mathbb{E}[Y(0)|C]\right)
$$

We also have $\mathbb{E}[X|Z=1] = \mathrm{Pr}(AT) + \mathrm{Pr}(C)$ and $\mathbb{E}[X|Z=0] = \mathrm{Pr}(AT)$. Hence,

$$
\begin{aligned}
\frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0]} &= \frac{\mathrm{Pr}(C)\left(\mathbb{E}[Y(1)|C] - \mathbb{E}[Y(0)|C]\right)}{\mathrm{Pr}(C)} \\
&= \mathbb{E}[Y(1)|C] - \mathbb{E}[Y(0)|C] \\
&= \mathbb{E}[Y(1) - Y(0)|C]
\end{aligned}
$$

which is called **LATE**.

Proposition 1.7

With Assumption 1.10, the **LATE** is given by

$$
\mathbb{E}[Y(1) - Y(0)|C] = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0]} = \frac{\mathrm{Cov}(Y, Z)}{\mathrm{Cov}(X, Z)}
$$

Remark

1. In RCT, $\mathrm{Pr}(C) = 1$, in which case ATE=LATE.

1.11 Difference in Difference (DiD)

The setup is the potential outcomes in Panel data.

Consider a two-way fixed effect model on the potential outcomes. For $D_{i_t} \in \{0, 1\}$, Y_{i_t} is given by

$$Y_{i_t}(0) = \alpha_i + \delta_t + \gamma X_{i_t} + \epsilon_{i_t}(0)$$

$$Y_{i_t}(1) = \alpha_i + \delta_t + \gamma X_{i_t} + \epsilon_{i_t}(1) + \theta$$

Assumption *We use following assumptions:*

1. $\epsilon_{i_t}(0) = \epsilon_{i_t}(1) := \epsilon_{i_t}$
2. $\mathbb{E}[\epsilon_{i_t} | X_{i_t}] = 0$

The ATE is given by

$$ATE := \mathbb{E}[Y_t(1) - Y_t(0)] = \theta + \underbrace{\mathbb{E}[\epsilon_{i_t}(1) - \epsilon_{i_t}(0)]}_{\text{by assumption } = 0}$$

Lemma 1.4

With Assumption 1.11, $ATE = \theta$.

$$Y_{i_t} = D_{i_t} Y_{i_t}(1) + (1 - D_{i_t}) Y_{i_t}(0) = \alpha_i + \delta_t + \theta D_{i_t} + \gamma X_{i_t} + \epsilon_{i_t}$$

1.11.1 After OLS Regression

Let $T = 2$, we have

$$Y_{i_2} = \delta_2 + \theta D_{i_2} + \gamma X_{i_2} + e_{i_2}, \text{ where } e_{i_2} = \alpha_i + \epsilon_{i_2}$$

Theorem 1.12

If $\mathbb{E}[e_{i_2} | X_{i_2}, D_{i_2}] = \Pi_0 + \Pi_1 X_{i_2}$, then the control function estimator (OLS) is consistent:

$$\hat{\theta}_{\text{CF}} \xrightarrow{P} ATE = \theta$$

However, what if $\alpha_i < \alpha_j$, the assumption $\mathbb{E}[e_{i_2} | X_{i_2}, D_{i_2}] = \Pi_0 + \Pi_1 X_{i_2}$ doesn't hold.

1.11.2 Difference in Difference

$$\Delta Y_i := Y_{i_2} - Y_{i_1} = \underbrace{\delta_2 - \delta_1}_{\delta} + \theta \Delta D_i + \gamma \Delta X_i + \Delta \epsilon_i$$

Case without covariate ($\gamma = 0$)

$$\Delta Y_i = \delta + \theta D_{i_2} + \Delta \epsilon_i$$

Assumption *[Parallel Trends Assumption]* $\mathbb{E}[\Delta\epsilon | D_2 = 1] = \mathbb{E}[\Delta\epsilon | D_2 = 0]$.

Theorem 1.13

Parallel Trends Assumption is equivalent to each of following conditions.

$$\begin{aligned} PT &\Leftrightarrow \mathbb{E}[\Delta Y(1) | D_2 = 1] = \mathbb{E}[\Delta Y(1) | D_2 = 0] \\ &\Leftrightarrow \mathbb{E}[\Delta Y(0) | D_2 = 1] = \mathbb{E}[\Delta Y(0) | D_2 = 0] \\ &\Leftrightarrow \text{Cov}(D_2, \Delta\epsilon) = 0 \end{aligned}$$

The DiD estimator is numerically same with OLS:

$$\hat{\theta}_{\text{DiD}} = \frac{\frac{1}{N}\sum_{i=1}^N \Delta Y_i D_{i_2}}{\frac{1}{N}\sum_{i=1}^N D_{i_2}} - \frac{\frac{1}{N}\sum_{i=1}^N \Delta Y_i (1 - D_{i_2})}{1 - \frac{1}{N}\sum_{i=1}^N D_{i_2}} \tag{DiD}$$

Case with covariates

$$\Delta Y_i = \delta + \theta D_{i_2} + \gamma \Delta X_i + \Delta\epsilon_i$$

Assumption $\mathbb{E}[\Delta\epsilon | D_2 = 1, \Delta X] = \mathbb{E}[\Delta\epsilon | D_2 = 0, \Delta X]$, *which is equivalent to* $\mathbb{E}[\Delta Y(d) | D_2 = 1, \Delta X] = \mathbb{E}[\Delta Y(d) | D_2 = 0, \Delta X], \forall d \in \{0, 1\}$.

Remark The DiD estimator (DiD) is no longer consistent:

$$\hat{\theta}_{\text{DiD}} \xrightarrow{P} \theta + \underbrace{\gamma\left(\mathbb{E}[\Delta X | D_2 = 1] - \mathbb{E}[\Delta X | D_2 = 0]\right)}_{\text{“selection on observables”}}$$