Introduction to Econometrics II: The Book[1]

Seung-hun Lee[2]

Spring 2020 Semester

[1] This book is meant to leave what was learned in the spring semester of the year 2020 as a record, hoping to help future first-year students navigate through the Econometrics sequence. I do not intend to publish this in any form. Thus, while sharing this file with others not necessarily affiliated to Columbia is allowed, I forbid reproducing this book for commercial purposes.

[2] sl4436@columbia.edu. Contact me if you find any errors.

Preface

This ‘Book’ is a compilation of all the recitation notes that I have used in the Spring 2020 semester when I worked as a teaching fellow for the Introduction to Econometrics II class at Columbia University. The intention behind The Book is to create a one-stop reference for all things that are related to graduate-level econometrics. In particular, this book focuses on the following topics

1. Instrumental Variables

2. Generalized Method of Moments

3. Panel Data (both static and dynamic)

4. Bootstrapping and Multiple Tests

5. Quantile Regression

6. Nonparametric and Semiparametric Estimation

7. Program Evaluation and Treatment Effects: Random Assignment, Selection on Observables, Selection on Unobservables, and Regression Discontinuity

8. Introduction to Machine Learning

In creating this book, I have been indebted to many individuals. First, the contents are mainly based on the lecture notes of Professors Jushan Bai and Bernard Salanie. I was given a rare opportunity to go over the first-year material once more. With this opportunity, I was able to gain mastery of the relevant methods that I can take anywhere in my research career. I am grateful to both professors for giving me this chance.

I was also greatly helped by two individuals - Paul Sungwook Koh and Dong Woo Hahm. They provided me an access to their prevous recitation notes. I was able to refine my explanation based on their notes, which benefitted those who learned from me and myself. Without them, this teaching fellowship would have been monumenally difficult.

Last but not least, I thank the first year cohort of 2019-2020 academic year. Not only have they corrected mistakes in my notes as classes went on, but they also gave me constructive and positive feedbacks for the recitation notes that became this book. The motivation from these reviews helped me complete this work.

Any remaining errors - typos, concepts that could be explained more efficiently than stated, etc. - are mine. I would appreciate any comments regarding the errors you may find here.

Contents

Chapter 1

Review on Asymptotic Theories

1.1 Motivation

Econometrics is about drawing an empirical conclusion from the given data. In particular, we are interested in the features of the estimators that we obtain from the data. While we cannot be sure of the distribution of such estimators in a finite sample context, we can still use asymptotic theories to derive the approximate distribution. In particular, we are interested in the consistency and the asymptotic variance of an estimator. This is where various modes of convergence and related theorems come in handy.

Before moving on, here are the Markov Inequality and the Chevyshev Inequality that would serve useful in proving some key theorems on this note and throughout the semester.

Theorem 1.1.1 (Markov Inequality)**.** *For any non-negative random variable X and $\epsilon > 0$,*

$$\Pr(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}$$

Theorem 1.1.2 (Chevyshev Inequality)**.** *For any random variable X and $\epsilon > 0$,*

$$\Pr(|X - E(X)| \geq \epsilon) \leq \frac{var(X)}{\epsilon^2}$$

Another useful tool is the asymptotic notations, or what we know as "Big O and small o". The definitions and properties of each notations are as follows

Definition 1.1.1 (Asymptotic Notations)**.** A sequence of non-random vectors X_n is said to be $O(1)$ if it is bounded and $o(1)$ if it converges to 0. In addition, if a_n is a sequence of non-random positive scalars, we say X_n is $O(a_n)$ if $\frac{X_n}{a_n}$ is bounded and $o(a_n)$ if $\frac{X_n}{a_n} \to 0$

Definition 1.1.2 (Asymptotic Notations (Stochastic))**.** A sequence of random vectors X_n is $O_p(1)$ if it is bounded in probability[a] and $o_p(1)$ if it converges to zero in probability[b]. If a_n is a random variable, then X_n is $O_p(a_n)$ if $\frac{X_n}{a_n}$ is $O_p(1)$. X_n is $o_p(a_n)$ if $\frac{X_n}{a_n} \xrightarrow{p} 0$.

Property 1.1.1. *The following holds true for O_p and o_p*

1. $o_p(1) \implies O_p(1)$
2. *If $X_n = O_p(a_n)$, then $X_n = o_p(b_n)$ for any b_n s.t. $a_n/b_n \to 0$*
3. *If random vector X_n converges in distrubtion to X, Then $X_n = O_p(1)$*
4. $o_p(1) + o_p(1) = o_p(1), O_p(1) + O_p(1) = O_p(1), o_p(1) + O_p(1) = O_p(1)$
5. $o_p(1)o_p(1) = o_p(1), O_p(1)O_p(1) = O_p(1), o_p(1)O_p(1) = o_p(1)$

[a] It means that for every $\epsilon > 0$, there is a constant $M_\epsilon > 0$ s.t. $\Pr(|X_n| > M_\epsilon) < \epsilon$ for all $n \in \mathbb{N}$

[b] Or for every $\epsilon > 0$, $\Pr(|X_n| > \epsilon) \to 0$

1.2 Various Modes of Convergence

A minimal criterion of a good estimator is consistency, which tells us that some estimator $\hat{\theta}_n$ converges to a true value θ for sufficiently large n. Convergence in probability is used to analyze whether the estimator is consistent or not.

Definition 1.2.1 (Convergence in Probability)**.** A sequence random variable $X_n \in \mathbb{R}$ **converges in probability** to a random variable X if for any $\epsilon > 0$, we have either one of

$$\lim_{n\to\infty} \Pr(|X_n - X| \leq \epsilon) = 1 \text{ or } \lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0$$

We denote this as $X_n \xrightarrow{p} X$.

The estimator $\hat{\theta}_n$ is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$. To show, consistency, we rely on the Weak Law of Large Numbers (hereafter WLLN).

Theorem 1.2.1 (WLLN)**.** *Let $\{X_n | n \geq 1\}$ be a sequence of IID random variables with $E|X| < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\mu = E[X_1]$. Then, $\bar{X}_n \xrightarrow{p} \mu$.*

Proof. For simplicity, I work with the case where the variance of the random variables are finite, namely $var(X_1) = \sigma^2 < \infty$. For any $\epsilon > 0$,

$$\begin{aligned}\Pr(|\bar{X}_n - \mu| > \epsilon) &\leq \frac{E(|\bar{X}_n - \mu|^2)}{\epsilon^2} \text{ } (\because \text{Chevyshev}) \\ &= \frac{\sigma^2}{n\epsilon^2} \text{ } (\because \text{Sample variance of } X_n = \sigma^2/n)\end{aligned}$$

By taking $n \to \infty$, $\frac{\sigma^2}{n\epsilon^2} \to 0$. Therefore, $\lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| > \epsilon) = 0$ □

We are also interested in the asymptotic distribution of the estimator. In particular, we are interested in checking whether the estimator is asymptotically normal.

Definition 1.2.2 (Convergence in Distribution)**.** Let X_n be a sequence of random variables with distribution $F_n(x) = \Pr(X_n \leq x)$. Let X be a random variable whose distribution is $F(x) = \Pr(X \leq x)$. A sequence random variable $X_n \in \mathbb{R}$ **converges in distribution** to a random variable X if at all x at which $F(x)$ is continuous,

$$\Pr(X_n \leq x) \to \Pr(X \leq x)$$

We denote this as $X_n \xrightarrow{d} X$.

In words, $X_n \xrightarrow{d} X$ does not necessarily mean that X_n and X are close - but their CDFs are close. If $X_n \xrightarrow{p} X$, X_n and X are close. Therefore, convergence in distribution has a weaker requirement than convergence in probability.

Convergence in distribution is applied when we use a central limit theorem.

Theorem 1.2.2 (Lindberg-Levy CLT)**.** *If $\{X_n | n \geq 1\}$ is a sequence of IID random variables with $E(X_1) = \mu < \infty, var(X_1) = \sigma^2 < \infty$, then*

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

or in standardized notation,

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right) \xrightarrow{d} N(0,1)$$

To answer whether the approximation provided by the CLT is accurate, we refer to the Berry-Esseen Theorem.

Theorem 1.2.3 (Berry-Esseen Theorem)**.** *For all sample sizes n and for all $t \in \mathbb{R}$,*

$$\left|\Pr\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \le t\right) - \Phi(t)\right| < \frac{C \times E(|X_i - \mu|^3)}{\sigma^3\sqrt{n}}$$

where the estimate of C has changed over time: $C \in (0.4097, 0, 4748)$.

The key advancement of this theorem is that it contains some implication about the rate of a convergence of a probability distribution of a random sample to a normal distribution. It also implies that the accuracy depends on the standardized absolute third moment of the distribution of X_i, $\frac{E(|X_i-\mu|^3)}{\sigma^3}$, and sample size.

We can extend the CLT to a multi-variate setting using Cramer-Wold Device.

Theorem 1.2.4 (Cramer-Wold Device)**.** *Let X_n and X be random vector in $\mathbb{R}^k$. Then,*

$$X_n \xrightarrow{d} X \iff \lambda' X_n \xrightarrow{d} \lambda' X, \ \ \forall \lambda \in \mathbb{R}^k$$

Theorem 1.2.5 (Multivariate Lindberg-Levy CLT)**.** *Let $X_1, .., X_n \in \mathbb{R}^k$ be IID and $E||X_1||^2 < \infty$. Then, as $n \to \infty$,*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, E[(X-\mu)(X-\mu)'])$$

where $X \in \mathbb{R}^k$.

Proof. Since X_i is IID, and $\lambda \in \mathbb{R}^k$, $\lambda' X_i$ is also IID, where $E[\lambda' X_i] = \lambda'\mu$ and $var(\lambda' X_i) = \lambda' var(X_i)\lambda$. By Lindberg-Levy CLT (scalar version), we can get

$$\sqrt{n}(\lambda'\bar{X}_n - \lambda'\mu) \xrightarrow{d} N(0, \lambda' var(X_i)\lambda) \iff \lambda'(\sqrt{n}(\bar{X}_n - \mu)) \xrightarrow{d} \lambda'(N(0, var(X_i)))$$

Then, apply Cramer-Wold device to get $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, var(X_i))$ □

1.3 Some Useful Theorems

To get the distribution and check for consistency of a seemingly complicated estimators, we can use the following theorems.

Theorem 1.3.1 (Continuous Mapping Theorem). *If* $X_n \xrightarrow{p,d} X$ *as* $n \to \infty$ *and* $g : \mathbb{R}^k \to \mathbb{R}^m$ *is continuous at every point of a set* C *s.t.* $P(X \in C) = 1$ *(or* g *has the set of discontinuity points* D *s.t.* $\Pr(X \in D) = 0$*), then*

$$g(X_n) \xrightarrow{p,d} g(X)$$

Theorem 1.3.2 (Slutsky's Theorem). *If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$ *(*$c \in \mathbb{R}$*) as* $n \to \infty$

1. $X_n \pm Y_n \xrightarrow{d} X \pm c$
2. $X_n Y_n \xrightarrow{d} Xc$
3. $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$, *provided that* $c \neq 0$

If $X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y$, *then the following holds*

1. $X_n \pm Y_n \xrightarrow{p} X \pm Y$
2. $X_n Y_n \xrightarrow{p} XY$
3. $\frac{X_n}{Y_n} \xrightarrow{p} \frac{X}{Y}$, *provided that* $\Pr(Y = 0) = 0$

However, the Slutsky's Theorem does not work well when both random variables converge in distribution, but not in probability. For such cases, we need to consider whether the two different random variables JOINTLY converge to another vector of random variable.

Example 1.3.1 (Counter-example). *Let* $Z_n \xrightarrow{d} Z$, *where* $Z \sim N(0,1)$. *Also define* $Y_n = (-1)^n Z_n$. *Then,* $Y_n \xrightarrow{d} N(0,1)$ *as well. However,* $Z_n + Y_n$ *displays following patterns*

$$Z_n + Y_n = \begin{cases} 2Z_n & n \text{ being even} \\ 0 & n \text{ being odd} \end{cases}$$

Thus, for even numbered n*, the distribution converges to* $N(0,4)$*. For odd numbered* n*, it con-*

verges to a distribution centered at 0. Therefore, $X_n + Y_n$ fails to converge in distribution to $2N(0,1) = N(0,4)$

Example 1.3.2 (Relationship between convergence of marginal and joint distribution)**.** *Suppose $\mathbf{X}_n = (A_n\ B_n)'$ and $\mathbf{X} = (A\ B)'$, where A_n, B_n, A, B are random variables.*

- *Suppose that $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$. From Cramer-Wold Device, let $\lambda = (1\ 0)'$. By applying Cramer-Wold Device, we can obtain that $A_n \xrightarrow{d} A$. We can obtain similar results for $B_n \xrightarrow{d} B$.*
- *Convergence in distribution for marginal distributions does not imply the same for joint distribution. Let A_n and B_n be independent standard normal random variables for all n. Thus, $A_n, B_n \xrightarrow{d} Z \sim N(0,1)$ individually. Jointly, however, $(A_n\ B_n)'$ is bivariate normal with correlation 0 while $(Z\ Z)'$ is bivariate normal with correlation 1. Therefore, we cannot say $(A_n\ B_n)'$ converges in distribution to $(Z\ Z)'$*
- *However, if it is the case that $A_n \xrightarrow{d} A$ and $B_n \xrightarrow{d} B$, where A_n is independent of B_n for all n and A is independent of B, it can be said that $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$. This is because*

$$F_{A_n,B_n}(a,b) = F_{A_n}(a)F_{B_n}(b) \to F_A(a)F_B(b) = F_{A,B}(a,b)$$

for all a, b that are the continuity points of F_A and F_B

The delta method provides us with another way to approximate the distribution of a smooth transformations of simpler objects.

Theorem 1.3.3 (Delta Method (Univariate))**.** *Suppose that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim N(0, \sigma^2)$ for some $\sigma^2 > 0$. If $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable at θ_0 with a nonzero derivative, then*

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} g'(\theta_0)Z \sim N(0, (g'(\theta_0))^2\sigma^2)$$

Proof. Take first order Taylor approximation of $g(\cdot)$ around θ_0, so that $g(x) = g(\theta_0) + g'(\bar{x})(x - \theta_0)$ for some $\bar{x}$ in between θ_0 and x. Then, let $x = \hat{\theta}_n$ and $\bar{x} = \bar{\theta}_n$ that is in between θ_0 and $\hat{\theta}_n$. We can then obtain

$$g(\hat{\theta}_n) - g(\theta_0) = g'(\bar{\theta}_n)(\hat{\theta}_n - \theta_0) \iff \sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) = g'(\bar{\theta}_n)\sqrt{n}(\hat{\theta}_n - \theta_0)$$

Since $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z$, we have $\hat{\theta}_n \xrightarrow{p} \theta_0$. In addition, $|\bar{\theta}_n - \theta_0| \leq |\hat{\theta}_n - \theta_0|$. In this context, $\bar{\theta}_n \xrightarrow{p} \theta_0$. By Continuous mapping theorem (and the fact that g' is continuous at θ_0), $g'(\bar{\theta}_n) \xrightarrow{p} g'(\theta_0)$. By Slutsky theorem (part (2)),

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} g'(\theta_0)Z = N(0, (g'(\theta_0)^2\sigma^2)$$

□

Theorem 1.3.4 (Delta Method (Multidimensional)). *Suppose that θ_0 is an m-dimensional vector and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim N(0, \Sigma)$ as $n \to \infty$. Let $g : \mathbb{R}^m \to \mathbb{R}^n$ be a function that is differentiable at θ_0 with a continuous partial derivatives at θ_0. Define $G(\theta_0)$ as*

$$G(\theta_0) = \frac{\partial g(\theta)}{\partial \theta'} |_{\theta=\theta_0} = \left[\frac{\partial g(\theta_0)}{\partial \theta_1}, ..., \frac{\partial g(\theta_0)}{\partial \theta_m}\right] \in \mathbb{R}^{n \times m}$$

Then, we can get

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} G(\theta_0)Z \sim N(0, G(\theta_0)\Sigma G(\theta_0)') \text{ as } n \to \infty$$

1.4 Problem with Standard Asymptotics

While the asymptotic methods introduced here are the most commonly used approach, it is not without problems. There could be more problems here, but I will introduce one of them. When the functions transforming the random variable to a non-linear form, the accuracy of the approximation may fall.

Example 1.4.1 (Non-linear Transformation). *If $X_i \sim N(\mu, \sigma^2)$, so that $\bar{X} \sim N(\mu, \sigma^2/n)$, and $g(\cdot) : x \to x^2$, delta method suggests that*

$$g(\bar{X}) \sim N\left(g(\mu), g'(\mu)^2 \frac{\sigma^2}{n}\right)$$

However, we have

$$E[g(\bar{X})] = E[\bar{X}^2] = var(\bar{X}) + \{E[\bar{X}]\}^2 = \mu^2 + \sigma^2/n$$

whereas $g(\mu) = \mu^2$. This implies that $g(\bar{X})$ has a bias of σ^2/n.

This happens since delta method is based on first-order Taylor approximation. Therefore, the accuracy of the approximation hinges on the performance of the first-order approximation of a non-linear function. One of the possible solutions is to use high-order approximation to obtain a smaller bias.

Chapter 2

Instrumental Variables

2.1 Classical Linear Models: Ordinary Least Squares

We will assume a data generating process that looks like

$$y_i = x_i'\beta + e_i,\ x_i = \begin{pmatrix} x_{i1} \\ ... \\ x_{ik} \end{pmatrix},\ i = 1, ..., n$$

where x_i and β are both in $\mathbb{R}^k$ and y_i and e_i are scalars. In a matrix notation, this can be written as

$$y = X\beta + e,\ y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix} \in \mathbb{R}^n, X = \begin{pmatrix} x_1' \\ ... \\ x_n' \end{pmatrix} \in \mathbb{R}^{n \times k}, e = \begin{pmatrix} e_1 \\ ... \\ e_n \end{pmatrix} \in \mathbb{R}^n$$

To demonstrate the consistency and the limiting distribution of the OLS estimators, I will use some of these assumptions

Assumption 2.1.1 (Assumptions for Classical Linear Models)**.** *The following assumptions are used in showing consistency and the limiting distribution of OLS estimators*

A1 (y_i, x_i) *are IID across i's*

A2 $E(x_i e_i) = 0$

A2′ $E(e_i|x_i) = 0$ *(Problem set 1 includes a question that asks you to derive A2 from A2′)*

A3 $E(x_i x_i') = Q$ *is a positive definite matrix (hereafter PD matrix)*

A4 $E||x_i^4|| < \infty, E||y_i^4|| < \infty$

The OLS estimator can be found by minimizing the sum of squared errors. In other words

$$\hat{\beta} = \min_b \sum_{i=1}^{n} (y_i - x_i' b)^2 = \left(\sum_{i=1}^{n} x_i x_i' \right)^{-1} \left(\sum_{i=1}^{n} x_i y_i \right)$$

or in matrix notation, $(X'X)^{-1}(X'y)$. The consistency and the limiting distribution of OLS estimators can be demonstrated as follows

Theorem 2.1.1 (Consistency of $\hat{\beta}$)**.** *Under assumptions **A1-A3**, $\hat{\beta} \xrightarrow{p} \beta$*

Proof. Rewrite $\hat{\beta}$ as $\beta + \left(\sum_{i=1}^n x_i x_i'\right)^{-1} \left(\sum_{i=1}^n x_i e_i\right)$. To carry out the asymptotic analysis on the summation terms, multiply $\frac{1}{n}$ to both. By **A1**, we can deduce that x_i and e_i are IID. Then, we can apply weak law of large numbers and continuous mapping theorem to show that

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

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

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

By assumptions **A2,A3**,$\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \xrightarrow{p} Q^{-1}$ and $\frac{1}{n}\sum_{i=1}^n x_i e_i' \xrightarrow{p} 0$. By Slutsky's theorem, $\left(\sum_{i=1}^n x_i x_i'\right)^{-1} \left(\sum_{i=1}^n x_i e_i\right) \xrightarrow{p} 0$. Thus, $\hat{\beta} \xrightarrow{p} \beta$. □

Theorem 2.1.2 (Limiting distribution of $\hat{\beta}$)**.** *Under assumptions **A1-A4**, the limiting distribution of $\hat{\beta}$ is characterized by $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, Q^{-1}\Omega Q^{-1})$, where $\Omega = E(x_i x_i' e_i^2)$*

Proof. We can write $\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i e_i\right)$. We know from the previous theorem that $\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \xrightarrow{p} Q^{-1}$. So we need to work on $\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i e_i\right)$.

From the central limit theorem, we can obtain the limiting distribution of $\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i\right)$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i \xrightarrow{d} N(0,\Omega)$$

since $E(x_i e_i) = 0$ by **A2** and $var(x_i e_i) = E(x_i e_i e_i x_i') - (E(x_i e_i))^2 = E(x_i x_i' e_i^2) = \Omega$ (In using CLT, we need assumption **A4** so that the variance-covariance matrix obtained from here is finite.) Then, by Slutsky's theorem,

$$\sqrt{n}(\hat{\beta}-\beta) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i\right) \xrightarrow{d} Q^{-1}N(0,\Omega) = N(0,\underbrace{Q^{-1}\Omega Q^{-1}}_{V})$$

□

If we are interested in a particular element of β, namely β_j, we will need to work on the following limiting distribution

$$\sqrt{n}(\hat{\beta}_j-\beta_j) \xrightarrow{d} N(0,V_{jj}) \text{ (V_{jj} is the (j,j)th element of V)}$$

2.2 Hypothesis Tests on OLS

- **Single element**: Consider the following setting

  $$H_0 : \beta_j = \beta_j^0, \;\; H_1 : \beta_j \neq \beta_j^0$$

  From the limiting distribution of β_j, we can show that the test statistic is distributed as a standard normal under H_0. It is characterized as

  $$\frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\widehat{V}_{jj}/n}} \xrightarrow{d} N(0,1)$$

  where $\widehat{V}$ is estimated version of the variance (more on that later).

- **Multiple element**: Sometimes, we may be interested in the features of the linear combinations of the elements of β, or even multiple restrictions, Let $R \in \mathbb{R}^{k\times r}$ characterize such restrictions. Then we can write

$$H_0 : R'\beta = c, \;\; H_1 : \neg H_0$$

Then from the limiting distribution of $\hat{\beta}$, we can apply Slutsky's theorem to get the necessary limiting distribution

$$\sqrt{n}(R'\hat{\beta} - R'\beta) = \sqrt{n}R'(\hat{\beta} - \beta) \xrightarrow{d} N(0, R'VR)$$

Since $R'\beta = c$ under H_0, we can obtain the following Wald test statistic. (For convenience, we also assume that V is known)

$$n(R'\hat{\beta} - c)'(R'VR)^{-1}(R'\hat{\beta} - c) \xrightarrow{d} \chi^2_r$$

To see why we ended up with a χ^2_r distribution, we need to understand the following.

Theorem 2.2.1 (Chi-squared distribution). *If $\eta \sim N(0, A)$, where A is PD, then*

$$\eta' A^{-1} \eta \sim \chi^2_{rank(A)}$$

- **Notes on $\widehat{V}$**: The definition of $\widehat{V}$ is $\widehat{V} \equiv \widehat{Q}^{-1}\widehat{\Omega}\widehat{Q}^{-1}$, where
  - $\widehat{Q}$ is a sample analogue of Q, written as $\frac{1}{n}\sum_{i=1}^{n} x_i x_i'$
  - $\widehat{\Omega}$ is a sample analogue of $E(x_i x_i' e_i^2)$, also with the consideration that the true value of e_i is replaced with the residual $\hat{e}_i$. Therefore, we can write $\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \hat{e}_i^2$

2.3 Source of Endogeneity

We still work with the data generating process $y_i = x_i'\beta + e_i$, but now with the possibility that $E(x_i e_i) \neq 0$. In other words, the error term and the regressor can now be correlated (or the regressors are endogenous). Since we required **A2** assumption in showing that OLS estimators are consistent, the fact that $E(x_i e_i)$ is not necessarily zero implies that OLS may no longer be consistent. In this section, we study cases in which regressors can become endogenous and how instrumental variables allow us to address this problem.

- **Measurement Error in Regressors:** Suppose that the linear model we want to estimate is as follows
$$y_i = {x_i^*}' \beta + e_i \text{ (We assume } E(x_i{}^* e_i) = 0)$$
However, we cannot observe x_i^*. Instead, we can observe $x_i = x_i^* + v_i$, where v_i has mean zero and independent of both x_i^* and e_i. So we have a classical measurement error in which x_i is unbiased, but noisy measure of x_i^*. If we use x_i instead,
$$y_i = (x_i - v_i)'\beta + e_i \quad = x_i'\beta \underbrace{-v_i'\beta + e_i}_{=u_i}$$
Then $E(x_i u_i)$ is as follows
$$E(x_i u_i) = E[x_i(-v_i'\beta + e_i)] = E[(x_i^* + v_i)(-v_i'\beta + e_i)] = -E(v_i v_i')\beta$$
So unless $\beta = 0$, or $E(v_i v_i') = 0$, $E(x_i u_i) \neq 0$. When we use OLS on this context, the probability limit of the OLS estimator would be
$$\begin{aligned}\hat{\beta}_{OLS} &= \beta + E(x_i x_i')^{-1} E(x_i u_i) \\ &= \beta - E(x_i x_i')^{-1} E(v_i v_i)\beta \\ &= \beta - E[(x_i^* + v_i)(x_i^* + v_i)']^{-1} E(v_i v_i)\beta \\ &= \beta - [E(x_i^* x_i^{*'}) + E(v_i v_i')]^{-1} E(v_i v_i)\beta \\ &= \frac{E(x_i^* x_i^{*'})}{E(x_i^* x_i^{*'}) + E(v_i v_i')}\beta \ (\leq \beta)\end{aligned}$$
The only time that $\frac{E(x_i^* x_i^{*'})}{E(x_i^* x_i^{*'}) + E(v_i v_i')}\beta$ would equal β is when β itself is zero or when $E(v_i v_i') = 0$. The latter, however, implies that $var(v_i) = 0$ and the noise v_i has mean 0 and has a point mass at 0 - so no measurement error exists. In usual cases, the OLS estimator has a probability limit of something less than β. This is what is also known as **attenuation bias**.

Comment 2.3.1 (Comment on Measurement Errors). *So how do we address the endogeneity problem?*

- *If there exists another noisy, but unbiased measure of x_i^*, namely $w_i = x_i^* + \delta_i$, we can use w_i to instrument for x_i. The condition is that η_i has mean zero and uncorrelated* *with $(x_i^*, e_i.v_i)$. Try verifying that this satisfies all IV conditions.*

- *If there is a measurement error in y_i, the only this it does is to change the component of e_i. Assuming all the old assumptions hold, this does not pose as much problem as having a measurement error in the regressor.*

- **Simultaneity Bias:** A classic example of this would be a supply and demand system type of setting:

$$q_i = \beta_1 p_i + u_i \tag{Supply}$$
$$q_i = -\beta_2 p_i + v_i \tag{Demand}$$

I will assume $e_i = (u_i\ v_i)'$ is IID, $E(e_i) = 0, E(e_i e_i') = I_2$ When you do some algebra, the equilibrium of this system is

$$p_i = \frac{v_i - u_i}{\beta_1 + \beta_2}, q_i = \frac{\beta_1 v_i + \beta_2 u_i}{\beta_1 + \beta_2}$$

So for both supply and demand equations, we have $E(p_i u_i) \neq 0$ and $E(p_i v_i) \neq 0$. When naively applying OLS to this equation, the result is as follows.

$$q_i = \beta^* p_i + \eta_i,\ E(p_i \eta_i) = 0 \implies \hat{\beta}^* = \frac{E(p_i q_i)}{E(p_i^2)} = \frac{\beta_1 - \beta_2}{2}$$

Thus, OLS estimators does not converge to either one of β_1 or β_2, resulting in a **simultaneity bias**.

- **Omitted Variable Bias (OVB):** Suppose that we are interested in the determinant of wages (y_i). Also assume that education, x_i, and innate ability, a_i, determine wages in the following manner

$$y_i = x_i \beta_1 + a_i \beta_2 + e_i,\ \ E(x_i e_i) = 0, E(a_i e_i) = 0$$

However, instead of observing (y_i, x_i, a_i), we can only observe (y_i, x_i). the best we can do at the moment is to estimate the following equation

$$y_i = x_i \beta_1 + u_i, \text{ where } u_i = a_i \beta_2 + e_i$$

Then $E(x_iu_i)$ becomes

$$E(x_iu_i) = E(x_i(a_i\beta_2 + e_i)) = E(x_ia_i)\beta_2 + 0 = E(x_ia_i)\beta_2$$

Therefore, when 1)x_i and a_i are correlated and 2)$\beta_2 \neq 0$, x_i is endogenous with respect to u_i. On the flip side, if either one of the condition is not met, $E(x_iu_i) = 0$ again. Moreover, the OLS estimator acquired here has a probability limit of

$$\hat{\beta}_{OLS} = \beta_1 + E(x_i^2)^{-1}E(x_iu_i) = \beta_1 + E(x_i^2)^{-1}E(x_ia_i)\beta_2$$

So if 1) and 2) occurs, the above does not converge in probability to β_1. Also note that we can determine the direction of the bias by the sign of $E(x_ia_i)$ and β_2.

2.4 Instrumental Variable Estimators

Assume that the data generating process is as follows

$$y_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + e_i$$

where $E(x_{1i}e_i) = 0, E(x_{2i}e_i) \neq 0$, and $\dim(x_{1i}) = k_1, \dim(x_{2i}) = k_2,\ k_1 + k_2 = k$. In our case, x_{2i} is the collection of endogenous variables. It can be shown that consistency of the OLS estimators of β_2 and β_1 will not be guaranteed under this situation.

Example 2.4.1 (When $k_1 = k_2 = 1$)**.** *In this case, we can write $\hat{\beta}_1$ as*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_{2i}^2 \sum_{i=1}^n x_{1i}y_i - \sum_{i=1}^n x_{1i}x_{2i} \sum_{i=1}^n x_{2i}y_i}{\sum_{i=1}^n x_{1i}^2 \sum_{i=1}^n x_{2i}^2 - (\sum_{i=1}^n x_{1i}x_{2i})^2}$$

When we replace y_i with $x_{1i}\beta_1 + x_{2i}\beta_2 + e_i$, we end up with

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n x_{2i}^2 \sum_{i=1}^n x_{1i}e_i - \sum_{i=1}^n x_{1i}x_{2i} \sum_{i=1}^n x_{2i}e_i}{\sum_{i=1}^n x_{1i}^2 \sum_{i=1}^n x_{2i}^2 - (\sum_{i=1}^n x_{1i}x_{2i})^2}$$

Since $E(x_{2i}e_i) \neq 0$, then $\frac{1}{n}\sum_{i=1}^n x_{2i}e_i$ converges to something that is not zero (whereas $\frac{1}{n}\sum_{i=1}^n x_{1i}e_i$ does converge in probability to 0). So the whole fraction term does not converge in probability to 0 and even $\hat{\beta}_1$ is not consistent.

Let $z_i \in \mathbb{R}^l = \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} = \begin{pmatrix} x_{1i} \\ z_{2i} \end{pmatrix}$, where $\dim(z_{2i}) = l - k_1$ and $l - k_1 \geq k_2$. For z_i to be a valid IV, the following conditions must be satisfied

Definition 2.4.1 (IV conditions)**.** z_i is a valid IV if

1. **Exogeneity**: $E(z_i e_i) = 0$
   (a) **Exclusion**: $E(z_i y_i) = \beta_1 E(z_i x_{1i}) + \beta_2 E(z_i x_{2i})$, in other words, z_i should impact y_i through x_{1i} and x_{2i}
2. **Relevancy**: $rank[E(z_i x_i')] = \dim(x_i) = k$
3. **PD**: $E(z_i z_i') > 0$

We will derive IV estimators in two ways: Reduced form approach and 2SLS approach

- **Reduced form**: In this approach, we assume that z_i is a least squares projection. So we can write

$$\begin{aligned} x_i &= \Gamma' z_i + u_i \ (\Gamma \in \mathbb{R}^{l \times k}) \\ \implies z_i x_i' &= z_i z_i' \Gamma + z_i u_i' \\ \implies E(z_i x_i') &= E(z_i z_i')\Gamma + E(z_i u_i') \end{aligned}$$

Since z_i is a least squares projection, $E(z_i u_i) = 0$. So

$$\Gamma = E(z_i z_i')^{-1} E(z_i x_i')$$

and the estimator for Γ would be its sample analogue, $\widehat{\Gamma} = \left(\frac{1}{n}\sum_{i=1}^n z_i z_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n z_i x_i'\right) = (Z'Z)^{-1}(Z'X)$.

Then we get to the structural equation

$$y_i = x_i'\beta + e_i \iff y_i = (z_i'\Gamma + u_i')\beta + e_i \iff y_i = z_i' \underbrace{\Gamma\beta}_{=\lambda} + \underbrace{u_i'\beta + e_i}_{=v_i}$$

From the assumptions, we can show that $E(z_i v_i) = 0$. As such,

$$\lambda = E(z_i z_i')^{-1} E(z_i y_i)$$

with its sample analogue being $\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i z_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i\right) = (Z'Z)^{-1}Z'y$

In case where $k = l$, Z itself becomes invertible. Then we can show that $\beta = \Gamma^{-1}\lambda$ and thus

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

If otherwise, We note that $\widehat{\Gamma}\beta + \text{error} = \hat{\lambda}$ and show that

$$\hat{\beta}_{IV} = (\widehat{\Gamma}'\widehat{\Gamma})^{-1}\widehat{\Gamma}'\hat{\lambda}$$

- **2SLS**: Suppose the structural equation and the first-stage regression is as follows.

$$\begin{aligned} y &= X\beta + e && \text{(Structural)} \\ X &= Z\Gamma + u && \text{(First Stage)} \end{aligned}$$

where $Z \in \mathbb{R}^{n\times l}, \Gamma \in \mathbb{R}^{l\times k}$. Z is our IV and we still maintain the least square projection assumption. We proceed as follows

1. Regress the first stage and obtain $\widehat{\Gamma} = (Z'Z)^{-1}Z'X$. Then the predicted value of X, denoted as $\widehat{X} = Z(Z'Z)^{-1}Z'X = P_Z X$.

Property 2.4.1 (Properties of Projection Matrix P_Z). *Note the following*

– *Symmetric:* $P_Z' = (Z(Z'Z)^{-1}Z')' = Z(Z'Z)^{-1}Z' = P_Z$

– *Idempotent:* $P_Z^2 = Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z' = Z(Z'Z)^{-1}Z' = P_Z$

2. In the structural equation, replace X with $\widehat{X}$ and obtain

$$\begin{aligned} \hat{\beta}_{2SLS} &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'y = (X'P_Z'P_Z X)^{-1}(X'P_Z'y) \\ &= (X'P_Z X)^{-1}X'P_Z y = (\widehat{X}'X)^{-1}\widehat{X}'y \end{aligned}$$

which is effectively replacing Z in the previous approach with $\widehat{X}$.

2.5 Consistency and Limiting Distribution of 2SLS

To show the asymptotic properties of the 2SLS estimators, we need the following set of assumptions

Assumption 2.5.1 (2SLS Assumptions)**.** *This is a list of required assumptions.*

T1 (y_i, x_i, z_i) *are IID*

T2 *Finite second moments:* $E||y_i^2|| < \infty, E||x_i^2|| < \infty, E||z_i^2|| < \infty$

T3 $E(z_i z_i') > 0$

T4 $rank[E(z_i x_i')] = k$

T5 $E(z_i e_i) = 0$

T6 *Finite fourth moments:* $E||y_i^4|| < \infty, E||x_i^4|| < \infty, E||z_i^4|| < \infty$

T7 $E(z_i z_i' e_i^2) = \Omega > 0$

Then, we can show the consistency and asymptotic normality of 2SLS estimators.

Theorem 2.5.1 (Consistency of 2SLS)**.** *Under assumptions* ***T1-T5****,* $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$

Proof. We can rewrite 2SLS estimators as $\hat{\beta}_{2SLS} = \beta + (X'P_Z X)^{-1} X'P_Z e$, which becomes

$$\beta + \left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'X}{n}\right]^{-1}\left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'e}{n}\right]$$

Define $Q_{ZX} = E(z_i x_i'), Q_{XX} = E(x_i x_i')$. Then by weak law of large numbers,

$$\frac{Z'X}{n} \xrightarrow{p} Q_{ZX}, \frac{X'X}{n} \xrightarrow{p} Q_{XX}, \frac{Z'e}{n} \xrightarrow{p} 0$$

By applying Slutsky's theorem and continuous mapping theorem, all the terms right of β converge in probability to 0. Thus, $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$ □

Theorem 2.5.2 (Limiting Distribution of 2SLS)**.** *Under assumptions* ***T1-T7****, the limiting distribution of the 2SLS estimator is characterized by* $\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, V_\beta)$

Proof. Note that

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

By CLT, $\frac{Z'e}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i e_i \xrightarrow{d} N(0, \Omega)$. Then apply Slutsky theorem and continuous mapping theorem to obtain

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, \underbrace{A_n^{-1}Q'_{ZX}Q_{ZZ}^{-1}\Omega Q_{ZZ}^{-1}Q_{ZX}A_n^{-1}}_{=V_\beta})$$

where $A_n = \left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'X}{n}\right]$ □

2.6 Some words of Caution

- **Correct Standard Errors** When estimating $\Omega = E(z_i z_i' e_i^2)$, we are not aware of the true error. So we need to estimate this as well. Note that we need to use a proper residual. Namely, we must use

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

  This is correct, as $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$. However, we frequently make a mistake of using

  $$\tilde{e}_i = y_i - \hat{x}_i'\hat{\beta}_{2SLS}$$

  Since $\tilde{x}_i$ is not exactly x_i, this converges in probability to something else.

  Comment 2.6.1 (on STATA). *To see this, compare the standard errors from doing* `ivregress 2sls y (x=z)` *and 'hard-coding' 2SLS by using* `reg x z`*, then* `predict xhat`*, and then coding* `reg y xhat`*.*

- **Nonlinear Extension**: If we are instead interested in the properties of $g(\hat{\beta}_{2SLS})$, where $g(\cdot)$ is not necessarily nonlinear, we can apply delta method here.

- **Too Many IV?**: In practice, when we have too many IV's (large dimension for z_i), it may be possible for the instruments to 'perfectly' predict x_i. In other words, $\hat{x}_i$ becomes nearly identical to x_i. This can become a problem because the endogeneity bias that plagued original structural equation may not be mitigated.
  One way of getting around this is to use a regression method that involves penalty mechanism - like LASSO, for instance. We will also formally treat this issue when we learn over-identification tests in the near future.

2.7 Limited Information Maximum Likelihood: Setup and Motivation

Assume a data generating process

$$y_i = \beta_1' x_{1i} + \beta_2' x_{2i} + e_i, \ (x_{1i} \in \mathbb{R}^{k_1}, x_{1i} \in \mathbb{R}^{k_2})$$

where $E(x_{1i}e_i) = 0$, but $E(x_{2i}e_i) \neq 0$. A **limited information maximum likelihood (LIML) estimator** derives the maximum likelihood estimator for the joint distribution of (y_i, x_{2i}) using structural equation of y_i and the reduced form equation for x_{2i}. This differs from the full information maximum likelihood (FIML) in the sense that the FIML requires structural equation for x_{2i} as well.

Briefly speaking, the reason we prefer LIML is as follows. When the number of the instruments are fixed, 2SLS and LIML have the same asymptotic distribution, with the former not requiring the assumption of normality (Greene, 2012). However, when there is a problem of weak instrument variable or too many instrumental variables, it can be shown that 2SLS becomes biased towards OLS. Others have shown that LIML estimators perform better in the presence of weak IVs and/or too many IVs.[1].

2.8 Limited Information Maximum Likelihood: Derivation

The structural equation for y_i is introduced in the previous page. Assume that there is a $z_i = \begin{pmatrix} x_{1i} \\ z_{2i} \end{pmatrix} \in \mathbb{R}^l$ where $E(z_i e_i) = 0$. I now define a reduced form equation for x_{2i}, written as

$$x_{2i} = \Gamma_{12}' x_{1i} + \Gamma_{22}' z_{2i} + u_{2i}, \ (\Gamma_{12}' \in \mathbb{R}^{k_2 \times k_1}, \Gamma_{22}' \in \mathbb{R}^{k_2 \times (l-k_1)}, z_{2i} \in \mathbb{R}^{l-k_1})$$

By putting the structural and reduced form equation together in a matrix form, we get

$$\underbrace{\begin{pmatrix} 1 & -\beta_2' \\ 0 & I_{k_2} \end{pmatrix}}_{=A} \underbrace{\begin{pmatrix} y_i \\ x_{2i} \end{pmatrix}}_{=w_i} = \underbrace{\begin{pmatrix} \beta_1' & 0 \\ \Gamma_{12}' & \Gamma_{22}' \end{pmatrix}}_{=B} \underbrace{\begin{pmatrix} x_{1i} \\ z_{2i} \end{pmatrix}}_{z_i} + \underbrace{\begin{pmatrix} e_i \\ u_{2i} \end{pmatrix}}_{=\eta_i}$$

[1] I refer further explanation to Jorn-Steffen Pischke's notes, found at http://econ.lse.ac.uk/staff/spischke/ec533/Weak%20IV.pdf

This now puts all endogenous variables on the left hand side and the exogenous variables on the right. To use the maximum likelihood approach, we need some assumptions on η_i. Specifically, we assume that η_i is conditionally IID normal and

$$\eta_i | z_i \sim N(0, \Sigma_\eta)$$

If we apply these facts, then we can derive the log-likelihood function for η_i

Lemma 2.8.1 (Multivariate Normal Density)**.** *Let $X \in \mathbb{R}^k$ be a multivariate normal distribution s.t. $X \sim N(\mu, \Sigma)$. Then the density function can be written as*

$$f_X(x) = (2\pi)^{-\frac{k}{2}} \det(\Sigma)^{-\frac{1}{2}} \times e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

Lemma 2.8.2 (Jacobian Transformation)**.** *If $X \in \mathbb{R}^k$ is a continuous random vector and $g : \mathbb{R}^k \to \mathbb{R}^k$ is one-to-one and onto function, the density of $Y = g(X)$ is given by*

$$f_Y(y) = f_X(g^{-1}(y))\,|J|$$

where $J = \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \dots & \frac{\partial x_1}{\partial y_k} \\ \dots & \dots & \dots \\ \frac{\partial x_k}{\partial y_1} & \dots & \frac{\partial x_k}{\partial y_k} \end{pmatrix}$. *Alternatively, we can write*

$$f_X(x) = f_Y(g(x))|H|, H = \det \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_k} \\ \dots & \dots & \dots \\ \frac{\partial y_k}{\partial x_1} & \dots & \frac{\partial y_k}{\partial x_k} \end{pmatrix}$$

So to work through the likelihood function, we take these steps

1. **Re-express the PDF**: The PDF pf η_i is

$$(2\pi)^{-\frac{k_2+1}{2}} \det(\Sigma)^{-\frac{1}{2}} \times e^{-\frac{1}{2}(\eta_i)'\Sigma_\eta^{-1}(\eta_i)}$$

   Note that $\eta_i = Aw_i - Bz_i$, Then we can write the pdf of w_i in terms of A and B as [2]

$$(2\pi)^{-\frac{k_2+1}{2}} \det(\Sigma)^{-\frac{1}{2}} \times e^{-\frac{1}{2}(Aw_i - Bz_i)'\Sigma_\eta^{-1}(Aw_i - Bz_i)}|A|$$

[2] Note that $\chi_i = w_i - A^{-1}Bz_i$ is also another normal distribution. You can notice that $\eta_j = A\chi_i$. Try matching this with the second version of the Jacobian transformation.

by Applying a Jacobian Transformation. Note that the determinant of the A matrix is 1. So this term would be dropped. By IID assumption, the log likelihood for $i = 1, ..., n$ becomes

$$C - \frac{n}{2}\log(\det(\Sigma_\eta)) - \frac{1}{2}\sum_{i=1}^{n}(Aw_i - Bz_i)'\Sigma_\eta^{-1}(Aw_i - Bz_i)$$

2. **Concentrating out**: To find A, B that maximizes the log - likelihood, we concentrate out the Σ_η term by using its estimator,

$$\widehat{\Sigma}_\eta = \frac{1}{n}\sum_{i=1}^{n}(Aw_i - Bz_i)(Aw_i - Bz_i)'$$

Then plug this estimator into the log-likelihood function to get

$$C - \frac{n}{2}\log(\det(\widehat{\Sigma}_\eta)) - \frac{1}{2}\sum_{i=1}^{n}(Aw_i - Bz_i)'\widehat{\Sigma}_\eta^{-1}(Aw_i - Bz_i)$$

Since the last term is a scalar, we can take

$$(Aw_i - Bz_i)'\widehat{\Sigma}_\eta^{-1}(Aw_i - Bz_i) = tr\left((Aw_i - Bz_i)'\widehat{\Sigma}_\eta^{-1}(Aw_i - Bz_i)\right)$$

Also, using the fact that $tr(AB) = tr(BA)$,

$$tr\left((Aw_i - Bz_i)'\widehat{\Sigma}_\eta^{-1}(Aw_i - Bz_i)\right) = tr\left(\widehat{\Sigma}_\eta^{-1}(Aw_i - Bz_i)(Aw_i - Bz_i)'\right)$$

This reduced the likelihood function to

$$C - \frac{n}{2}\log(\det(\widehat{\Sigma}_\eta)) - \frac{1}{2}tr\left(\widehat{\Sigma}_\eta^{-1}\sum_{i=1}^{n}(Aw_i - Bz_i)(Aw_i - Bz_i)'\right) = C - \frac{n}{2}\log(\det(\widehat{\Sigma}_\eta)) - \frac{n(k_2+1)}{2}$$

3. **Maximize**: The only thing left to maximize over now is $-\frac{n}{2}\log(\det(\widehat{\Sigma}_\eta))$. This is equivalent to maximizing

$$-\det\left(\frac{1}{n}\sum_{i=1}^{n}(Aw_i - Bz_i)'(Aw_i - Bz_i)\right) = -\det\left(\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} y_i - \beta_1'x_{1i} - \beta_2 x_{2i} \\ x_{2i} - \Gamma_{12}'x_{1i} - \Gamma_{22}'z_{2i}\end{pmatrix}\begin{pmatrix} y_i - \beta_1'x_{1i} - \beta_2 x_{2i} \\ x_{2i} - \Gamma_{12}'x_{1i} - \Gamma_{22}'z_{2i}\end{pmatrix}'\right)$$

The combination of β and Γ that maximizes this is the LIML estimator. [3]

[3] Further steps involve heavy amounts of algebra involving multi-dimensional matrices. For details, I will have you refer to the recitation notes from Dong Woo last year, which I will upload.

2.9 k-class Estimators

Another way to compute the LIML estimator is to use a k-class estimator with a particular choice for k. In general, k-class estimator is defined as

$$\hat{\beta}_k = \arg\min_{\beta}(y - X\beta)'(I_n - kM_Z)(y - X\beta)$$

where $M_Z = I - Z(Z'Z)^{-1}Z' = I - P_Z$. The other way to express this, after some matrix differentiation, is

$$\begin{gathered} -2X'y + 2kX'M_Z y + 2(X'X)\beta - 2k(X'M_Z X)\beta = 0 \\ \iff (X'(I_n - kM_Z)X)\beta = X'(I_n - kM_Z)y \\ \implies \hat{\beta}_k = (X'(I_n - kM_Z)X)^{-1}(X'(I_n - kM_Z)y) \end{gathered}$$

In fact, we can show that OLS ($k = 0$) and 2SLS ($k = 1$) are k-class estimators. LIML is a k-class estimator with a parameter choice of some $k > 1$. What we need to do find the associated value of k. To do so, define

$$W = (y \;\; X_2) \in \mathbb{R}^{n\times(k_2+1)},\; M_1 = I - X_1(X_1'X_1)^{-1}X_1'$$

Then we compute the minimum eigenvalue of [4]

$$(W'M_1W)(W'M_ZW)^{-1}$$

This minimum eigenvalue will be our choice of k, which will be denoted as $\hat{k}$, and the LIML estimator would be

$$\hat{\beta}_{LIML} = (X'(I_n - \hat{k}M_Z)X)^{-1}(X'(I_n - \hat{k}M_Z)y)$$

From here, we can also find that LIML is also a type of an IV estimator. We can see that by rewriting above equation as

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

where $\tilde{X} = (I_n - \hat{k}M_Z)X$. This also hints that the asymptotic properties of LIML and some other types of IV estimators are similar.

[4] This is shown in Anderson and Rubin (1949) and Amemiya (1985)

2.10 Asymptotics of LIML

Given the above definition of $\hat{\beta}_{LIML} = (X'(I_n - \hat{k}M_Z)X)^{-1}(X'(I_n - \hat{k}M_Z)y)$, we can use this to study the asymptotic properties of LIML estimators. Note that

$$\begin{aligned}
\hat{\beta}_{LIML} - \beta &= (X'(I_n - \hat{k}M_Z)X)^{-1}(X'(I_n - \hat{k}M_Z)e) \\
&= (X'(P_Z - (\hat{k}-1))M_Z)X)^{-1}(X'(P_Z - (\hat{k}-1))e) \\
&(\because I - M_Z = P_Z, \implies I - \hat{k}M_Z = P_Z - (\hat{k}-1)M_Z) \\
\implies \sqrt{n}(\hat{\beta}_{LIML} - \beta) &= \left(\frac{X'P_ZX}{n} - (\hat{k}-1)\frac{X'M_ZX}{n}\right)^{-1}\left(\frac{X'P_ZX}{\sqrt{n}} - (\hat{k}-1)\frac{X'M_Ze}{\sqrt{n}}\right)
\end{aligned}$$

Note that

- Anderson and Rubin (1949) shows that $\hat{k} - 1 \xrightarrow{p} 0$, which we accept as given.
- $\frac{X'M_ZX}{n} = \frac{X'X}{n} - \frac{X'P_ZX}{n} = \frac{X'X}{n} - \frac{X'P_Z'P_ZX}{n} = \frac{X'X}{n} - \frac{(P_ZX)'(P_ZX)}{n} \leq \frac{X'X}{n} \xrightarrow{p} E(x_ix_i')$. Therefore, $\frac{X'M_ZX}{n}$ is bounded (and thus $O_p(1)$)
- $\frac{X'M_Ze}{\sqrt{n}}$ converges in distribution to a normal distribution (CLT), so it is $O_p(1)$
- From the third and first points, we can infer that $(\hat{k}-1)\frac{X'M_Ze}{\sqrt{n}} = o_p(1)$

Combining these leads to the result that

$$\sqrt{n}(\hat{\beta}_{LIML} - \beta) = \left(\frac{X'P_ZX}{n}\right)^{-1}\frac{X'P_ZX}{\sqrt{n}} + o_p(1)$$

which is equivalent to $\sqrt{n}(\hat{\beta}_{2SLS} - \beta) + o_p(1)$. So for a fixed n, the 2SLS and LIML converge to a similar limiting distribution. However, when either instrument is weakly related to X or large in numbers, 2SLS bias becomes large, giving advantage to the LIML[5].

[5] This is the key takeaway from Dr. Pischke's slides.

2.11 Control Function Approach

This is another way to derive a 2SLS estimator. As before, I will assume $E(x_{1i}e_i) = 0, E(x_{2i}e_i) \neq 0$ and write the structural and reduced form regression as

$$y_i = x'_{1i}\beta_1 + x'_{2i}\beta_2 + e_i \tag{Structural}$$
$$x_{2i} = \Gamma'_{12}x_{1i} + \Gamma'_{22}z_{2i} + u_{2i} \tag{Reduced Form}$$

Also, we have a $z_i \in (x_{1i}\ z_{2i})' \in \mathbb{R}^l$ that satisfies $E(z_ie_i) = 0$. The key driving idea for the control function method is that we can write $E(x_{2i}e_i)$ differently. Specifically,

$$\begin{aligned}
E[x_{2i}e_i] &= E[(\Gamma'_{12}x_{1i} + \Gamma'_{22}\beta_2 + u_{2i})e_i] \\
&= \Gamma'_{12}E(x_{1i}e_i) + \Gamma'_{22}E(z_{2i}e_i) + E(u_{2i}e_i) \\
&= E(u_{2i}e_i) \ (\because \text{The first by exogeneity, the second by IV conditions}) \\
&\therefore E[x_{2i}e_i] \neq 0 \iff E[u_{2i}e_i] \neq 0
\end{aligned}$$

From here, we consider a linear projection of e_i onto u_{2i}, which we write as

$$e_i = u'_{2i}\alpha + \epsilon_i \tag{LP}$$

where $E(u_{2i}\epsilon_i) = 0$ and the population analogue of $\alpha = E(u_{2i}u'_{2i})^{-1}E(u_{2i}e_i)$. What we now do is to substitute the e_i term in the (Structural) equation with the (LP) equation to get

$$y_i = x'_{1i}\beta_1 + x'_{2i}\beta_2 + u'_{2i}\alpha + \epsilon_i \tag{CFA}$$

where the following are satisfied

$$\begin{aligned}
E[x_{1i}\epsilon_i] &= E[x_{1i}(e_i - u'_{2i}\alpha)] = 0 - 0 = 0 \\
E[u_{2i}\epsilon_i] &= E[u_{2i}(e_i - u'_{2i}\alpha)] \\
&= E[u_{2i}e_i] - E[u_{2i}u'_{2i}]E[u_{2i}u'_{2i}]^{-1}E[u_{2i}e_i] = 0 \\
E[x_{2i}\epsilon_i] &= E[x_{2i}(e_i - u'_{2i}\alpha)] = E[x_{2i}e_i] - E[x_{2i}u'_{2i}]\alpha \\
&= E[u_{2i}e_i] - \Gamma'_{12}E[x_{1i}u'_{2i}]\alpha - \Gamma'_{22}E[z_{2i}u'_{2i}]\alpha - E[u_{2i}u'_{2i}]\alpha \\
&= E[u_{2i}e_i] - 0 - 0 - E[u_{2i}e_i] = 0 \\
&(\because \text{LS assumption from Reduced Form, Recitation 2})
\end{aligned}$$

So the key takeaway is that now x_{2i} is exogenous with ϵ_i. In words, x_{2i} is correlated with e_i through u_{2i} and ϵ_i is the error after e_i has been projected onto u_{2i}.

If we know the true u_{2i}, the (CFA) equation can be regressed with an OLS straight away. This is not usually the case, so we need to use the residual $\hat{u}_{2i}$ which we can get from the (Reduced Form) equation. We work in these steps

1. **Obtain** $\hat{u}_{2i}$: This is done by regressing (Reduced Form) equation.

2. **Work with (CFA)**: However, instead of u_{2i}, use $\hat{u}_{2i}$. Then we can run an OLS on the REWRITTEN (CFA) equation.

Once this is done, we can show that the estimates from control function approach is numerically identical to 2SLS.[6]. One thing to note is that with this setup, we can conduct a test of endogeneity. Formally we want to test

$$H_0 : E(x_{2i}e_i) = 0,\ H_1 : E(x_{2i}e_i) \neq 0$$

If it is the case that $E(x_{2i}e_i) = 0$, then $E(u_{2i}e_i) = 0$. Then, by how we constructed α, this implies that $\alpha = 0$. We can then show that the Wald statistics, under H_0, is distributed as

$$\hat{\alpha}'(var(\hat{\alpha}))^{-1}\hat{\alpha} \xrightarrow{d} \chi^2_{k_2}$$

To test the null against the alternative hypothesis, define $C_{1-\alpha}$ as $\Pr(\chi^2_{k_2} \leq C_{1-\alpha}) = 1 - \alpha$. Then, we can reject the H_0 hypothesis if $\hat{\alpha}'(var(\hat{\alpha}))^{-1}\hat{\alpha} > C_{1-\alpha}$.

2.12 Hausmann Tests for Specification

Hausmann Test can be utilized as a general test of specification. In this context, we are interested in using Hausmann Tests for exogeneity / endogeneity of the regressors. Assume the following data generating process

$$y_i = x'_{1i}\beta_1 + x'_{2i}\beta_2 + e_i$$

and we are interested in checking the exogeneity / endogeneity of x_{2i}. So we test $H_0 : E(x_{2i}e_i) = 0$ against $H_1 : E(x_{2i}e_i) \neq 0$. Consider these properties of 2SLS and OLS estimators

[6] A more detailed explanation can be found in Hansen (2019) pages 426-427.

- $\hat{\beta}_{OLS}$: Consistent and minimal variance under H_0, inconsistent under H_1
- $\hat{\beta}_{2SLS}$: Consistent in either H_0 or H_1. Inefficient under H_0.

Under H_0, $\sqrt{n}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})$ converges in distribution to $N(0, var(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}))$. Hausmann also shows that in H_0, $var(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}) = var(\hat{\beta}_{2SLS}) - var(\hat{\beta}_{OLS})$ holds in general (you are asked to show this for a scalar case). Given this, we can write the Hausmann test statistic as

$$(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})'(var(\hat{\beta}_{2SLS}) - var(\hat{\beta}_{OLS}))^{-}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}) \xrightarrow{d} \chi^2_{k_2}$$

where $(var(\hat{\beta}_{2SLS}) - var(\hat{\beta}_{OLS}))^{-}$ indicates a generalized inverse. We can use the test statistic, construct a critical value and reject/not reject H_0 based on comparison.

Note that generalized inverse is required since $(var(\hat{\beta}_{2SLS}) - var(\hat{\beta}_{OLS}))$ is usually not a full (column) rank matrix. The following trick is useful in showing this.

Lemma 2.12.1 (Breakdown of OLS coefficients, from Hansen (2019): pp. 80-82)**.** *The OLS estimator for β_1, β_2 in $y = X_1\beta_1 + X_2\beta_2 + e$ can be written as*

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}(X_1'M_2y),\ \hat{\beta}_2 = (X_2'M_1X_2)^{-1}(X_2'M_1y)$$

where $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$ and M_2 is defined similarly.

Proof. I will provide some sketches of this proof. Suppose that we want to find β_1's OLS estimate. Then the problem is identical to

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

In words, we concentrate out β_2 and express it as a function of β_1. We put this expression back into the original equation and find β_1 from there. As a result, we can get that

$$\beta_2 = (X_2'X_2)^{-1}(X_2'(y - X_1\beta_1))$$

Put this back in to obtain the OLS estimation for β_1, which is

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}(X_1'M_2y)$$

It is also useful to note that M_1 is symmetric and idempotent. □

In our context, $\hat{\beta}_{1,OLS}$ would be $(X_1'X_1)^{-1}(X_1'(y - X_2\hat{\beta}_{2,OLS}))$ and its 2SLS estimator is $(\widehat{X}_1'\widehat{X}_1)^{-1}(\widehat{X}_1'(y - \widehat{X}_2\hat{\beta}_{2,2SLS}))$. Since $Z = [X_1 \ Z_2]' \in \mathbb{R}^{n\times l}$, we can write that $X_1 = Z\Gamma$ with $\Gamma = \begin{pmatrix} I_{k_1} \\ 0 \end{pmatrix} \in \mathbb{R}^{l\times k_1}$. So we can write $\widehat{X}_1 = Z(Z'Z)^{-1}Z'X_1 = Z\Gamma = X_1$. Therefore,

$$\begin{aligned}
\hat{\beta}_{1,OLS} - \hat{\beta}_{1,2SLS} &= (X_1'X_1)^{-1}X_1'(\widehat{X}_2\hat{\beta}_{2,2SLS} - X_2\hat{\beta}_{2,OLS}) \\
&= (X_1'X_1)^{-1}X_1'(P_Z X_2\hat{\beta}_{2,2SLS} - X_2\hat{\beta}_{2,OLS}) \\
&= (X_1'X_1)^{-1}X_1'P_Z X_2\hat{\beta}_{2,2SLS} - (X_1'X_1)^{-1}X_1'X_2\hat{\beta}_{2,OLS}) \\
&= (X_1'X_1)^{-1}X_1'X_2\hat{\beta}_{2,2SLS} - (X_1'X_1)^{-1}X_1'X_2\hat{\beta}_{2,OLS} \ (\because X_1'P_Z = (P_Z X_1)' = X_1') \\
&= (X_1'X_1)^{-1}X_1'X_2(\hat{\beta}_{2,2SLS} - \hat{\beta}_{2,OLS})
\end{aligned}$$

So $\hat{\beta}_{1,OLS} - \hat{\beta}_{1,2SLS}$ is merely obtained by linearly transforming $\hat{\beta}_{2,OLS} - \hat{\beta}_{2,2SLS}$. Therefore, $var(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})$ is not full rank.

2.13 Subset Endogeneity Test

This section can be seen as an extension to the endogeneity test we showed earlier. Instead, we break down x_{2i} into two parts - one that is 'potentially' endogenous (x_{2i}) and one that is endogenous for sure (x_{3i}). Then, we again test for $H_0 : E(x_{2i}e_i) = 0$ against $H_1 : E(x_{2i}e_i) \neq 0$. We now go over how control function approach can be utilized in this context. We have the set of required equations as follows

$$y_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + x_{3i}'\beta_3 + e_i \tag{Structural}$$
$$x_{2i} = \Gamma_2' z_i + u_{2i} \tag{Reduced Form 2}$$
$$x_{3i} = \Gamma_3' z_i + u_{3i} \tag{Reduced Form 3}$$

We then project e_i onto u_{2i}, u_{3i} to obtain

$$e_i = u_{2i}'\alpha_2 + u_{3i}'\alpha_3 + \epsilon_i \tag{LP2}$$

where the following is satisfied

$$\begin{pmatrix} E(u_{2i}u'_{2i}) & E(u_{2i}u'_{3i}) \\ E(u_{3i}u'_{2i}) & E(u_{3i}u'_{3i}) \end{pmatrix} \begin{pmatrix} \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} E(u_{2i}e_i) \\ E(u_{3i}e_i) \end{pmatrix}$$

then the control function approach would require us to apply an OLS to the following

$$y_i = x'_{1i}\beta_1 + x'_{2i}\beta_2 + x'_{3i}\beta_3 + u'_{2i}\alpha_2 + u'_{3i}\alpha_3 + \epsilon_i \tag{CFA2}$$

As before, $E(x_{2i}e_i) = 0$ implies $E(u_{2i}e_i) = 0$. This allows us to replace the H_0 with

$$H_0 : E[u_{2i}u'_{2i}]\alpha_2 + E[u_{2i}u'_{3i}]\alpha_3 = 0$$

This is a special case of $R\beta = c$ type of hypothesis test studied in Recitation 1. However, since we cannot usually know what the true value of u_{2i}, u_{3i} are, we use the residuals, and replace R with $\widehat{R}$ constructed from the sample analogue using residuals.

2.14 Overidentification Test

Consider the following setup

$$y_i = x'_i\beta + e_i, \ (\dim(x_i) = k < \dim(z_i) = l)$$

where we have an overidentified model, with number of instruments being larger than endogenous variables. We can test for the validity of the instruments by testing $H_0 : E(z_ie_i) = 0$ vs. $H_1 : E(z_ie_i) \neq 0$. To do this, we construct a linear projection equation by projecting e_i onto z_i, obtaining

$$e_i = z'_i\alpha + \epsilon_i$$

Thus, $\alpha = E(z_iz'_i)^{-1}E(z_ie_i)$. Because of this, the null and alternative hypotheses can be expressed by $H_0 : \alpha = 0$ vs. $H_1 : \alpha \neq 0$. This leaves us with the challenge of estimating α. Since we do not know what the true value of the error terms are, we take the following steps.

1. **Obtain** $\hat{e}_i$: This can be done using 2SLS estimates of β. As a result,

$$\hat{e}_i = y_i - x'_i\hat{\beta}_{2SLS} \implies \hat{e} = y - X\hat{\beta}_{2SLS}$$

2. **Obtain** $\hat{\alpha}$: Replace e with $\hat{e}$ to get

$$\hat{\alpha} = (Z'Z)^{-1}Z'\hat{e}$$

3. **Sargan Test**: We will assume homoskedasticity (otherwise, the test statistic does not converge to χ^2 distribution). Then we make use of the following test statistic

$$S = \hat{\alpha}'(var(\hat{\alpha}))^{-}\hat{\alpha} = \frac{\hat{e}'Z(Z'Z)^{-1}Z'\hat{e}}{\hat{\sigma}^2}$$

where $\hat{\sigma}^2$ can be obtained from two paths - one using $\frac{1}{n}\hat{e}'\hat{e}$ and the other using $\frac{1}{n}\hat{\epsilon}'\hat{\epsilon} = \frac{1}{n}(\hat{e} - Z\hat{\alpha})'(\hat{e} - Z\hat{\alpha})$. While they are slightly different in that the first estimate has larger variances[7], they are asymptotically equal. Under the null, $S \xrightarrow{d} \chi^2_{l-k}$

Some note of caution include:

- Even if the null hypothesis is rejected, the test alone is not enough to pick up which variable is violating the exclusion restriction.

- Since $E(z_i e_i) = 0 \iff E(z_i y_i) - E(z_i x_i')\beta = 0$, we are effectively solving a system of $l \times 1$ restrictions, despite β being a k-vector. Depending on the subset of z_i vectors used, the β may differ in finite samples. If the overidentification hypothesis is correct, the estimates that are different in finite samples are consistent for true β. You can think of the overidentification test as an attempt to see whether different β estimates are converging to a same parameter as $n \to \infty$

2.15 Weak Instrument Test

Assume that the structural and reduced form equations are (we are working with a scalar regressors)

$$\begin{aligned} y_i &= x_i\beta + e_i \\ x_i &= z_i\gamma + u_i \end{aligned}$$

[7] It can be shown that $\hat{\epsilon}'\hat{\epsilon} = \hat{e}'\hat{e} - \hat{e}'Z(Z'Z)^{-1}Z'\hat{e}$ Since $\hat{e} \xrightarrow{p} e$ and $E(z_i e_i) = 0$, it can be shown that $\hat{e}'Z(Z'Z)^{-1}Z'\hat{e}$ goes to 0.

We say that there is a problem of weak instrument if $\gamma \simeq 0$. This can cause the 2SLS estimates to be biased and the distribution to be affected. Specifically, let $\gamma = \frac{\mu}{\sqrt{n}}$. For simplicity, I will assume

- $var\left(\begin{pmatrix} e_i \\ u_i \end{pmatrix} |z_i\right) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \Sigma$
- $E(z_i^2) = 1$
- $\frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} z_i e_i \\ z_i u_i \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = N(0, \Sigma)$

Now I will show that neither OLS nor IV estimator is not consistent.

- OLS: Note that $\hat{\beta}_{OLS} - \beta$ can be written as $\frac{n^{-1}\sum_{i=1}^n x_i e_i}{n^{-1}\sum_{i=1}^n x_i^2}$, equivalent to

$$\begin{aligned}
\frac{n^{-1}\sum_{i=1}^n x_i e_i}{n^{-1}\sum_{i=1}^n x_i^2} &= \frac{n^{-1}\sum_{i=1}^n (z_i\gamma + u_i)e_i}{n^{-1}\sum_{i=1}^n (z_i\gamma + u_i)^2} \\
&= \frac{n^{-1}\sum_{i=1}^n (z_i e_i \gamma + u_i e_i)}{n^{-1}\sum_{i=1}^n (z_i^2\gamma^2 + u_i^2 + 2z_i u_i \gamma)} \\
&= \frac{n^{-1}\sum_{i=1}^n u_i e_i}{n^{-1}\sum_{i=1}^n u_i^2} + o_p(1) \\
&\xrightarrow{p} E(u_i e_i)/E(u_i^2) = \rho
\end{aligned}$$

- IV: For $\hat{\beta}_{IV} - \beta$, we can write $\frac{n^{-1/2}\sum_{i=1}^n z_i e_i}{n^{-1/2}\sum_{i=1}^n z_i x_i}$ or equivalently

$$\begin{aligned}
\hat{\beta}_{IV} - \beta &= \frac{n^{-1/2}\sum_{i=1}^n z_i e_i}{n^{-1/2}\sum_{i=1}^n z_i(z_i\gamma + u_i)} \\
&= \frac{n^{-1/2}\sum_{i=1}^n z_i e_i}{n^{-1/2}\sum_{i=1}^n z_i u_i + n^{-1}\sum_{i=1}^n z_i^2 \mu} \xrightarrow{d} \frac{\xi_1}{\xi_2 + \mu}
\end{aligned}$$

  which is not centered at 0, making it inconsistent.

In STATA, such test can be implemented by typing in `estat firststage` after 2SLS or LIML regression.

2.16 Problem with Many IVs

In some instances, we may work with "many" IVs. This indicates a situation where the number of IV is large relative to the sample size. To formalize the idea, Bekker (1994)[8] intro-

[8] Bekker, Paul A. (1994), "Alternative Approximations to the Distributions of Instrumental Variable Estimators", Econometrica 62(3), 657-681

duces an asymptotic approximation which treats the number of instruments l as proportional to the sample size, so that $l = \alpha n$. This is equivalent to

$$l/n \to \alpha$$

When α is not zero, then it can be said that this setup has many IVs. This could cause the 2SLS estimators to be inconsistent as well.

Consider the setup where x_i is endogenous and is a scalar.

$$y_i = x_i'\beta + e_i \iff Y = X\beta + e$$
$$x_i = z_i'\beta + u_i \iff X = Z\Gamma + u \ \ (z_i \in \mathbb{R}^l)$$

To make the analysis more simpler, I assume that z_i is still a valid IV (relevant, exogenous) and that $var\begin{pmatrix} e_i \\ u_i \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \Sigma$. In addition, note that $var(x_i) = var(z_i'\gamma) + var(u_i)$ and assume that

$$\frac{1}{n}\sum_{i=1}^{n} \gamma' z_i z_i' \gamma \xrightarrow{p} c > 0$$

and that variance of x_i and u_i are unchanging with respect to l. This implies that the variance of $var(z_i'\gamma)$ is not changing as well an that that R^2 of the reduced form[9] converges to a constant.

To understand the behavior of some of the estimators in many IV situations, I will first start with OLS as a benchmark.

- OLS: We know that $\hat{\beta}_{OLS}$ can be written as

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

[9] In this context, it would be $\frac{var(z_i'\gamma)}{var(x_i)}$

Applying our setup, we can re-write this as

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' = \frac{1}{n}\sum_{i=1}^{n}\gamma' z_i z_i'\gamma + \frac{1}{n}\sum_{i=1}^{n} u_i u_i' + \frac{2}{n}\sum_{i=1}^{n}\gamma' z_i u_i'$$
$$\xrightarrow{p} c+1$$
$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \xrightarrow{p} (c+1)^{-1}$$
$$\frac{1}{n}\sum_{i=1}^{n} x_i e_i \xrightarrow{p} \rho$$

Therefore,

$$\hat{\beta}_{OLS} - \beta \xrightarrow{p} \frac{\rho}{c+1}$$

This also captures the idea that if x_i is endogenous (regardless of size), OLS estimators fail to be consistent

- 2SLS: The 2SLS estimator can be characterized by

$$\begin{aligned}\hat{\beta}_{2SLS} - \beta &= (X'P_Z X)^{-1}(X'P_Z e)\\ &= [(\Gamma'Z' + u')Z(Z'Z)^{-1}Z'(Z\Gamma + u)]^{-1}[(\Gamma'Z' + u')Z(Z'Z)^{-1}Z'e]\\ &= [\frac{\Gamma'Z'Z\Gamma}{n} + \frac{\Gamma'Z'u}{n} + \frac{u'Z\Gamma}{n} + \frac{u'P_Z u}{n}]^{-1}[\frac{\Gamma'Z'e}{n} + \frac{u'P_Z e}{n}]\end{aligned}$$

We know from the above discussion about $\frac{1}{n}\sum_{i=1}^{n}\gamma' z_i z_i'\gamma$ and the properties of Z that $\frac{\Gamma'Z'Z\Gamma}{n} \xrightarrow{p} c$, $\frac{\Gamma'Z'u}{n} \xrightarrow{p} 0$, $\frac{\Gamma'Z'e}{n} \xrightarrow{p} 0$. We need to know what happens to the rest of them. Using the fact that $u'P_Z e$ and $u'P_Z u$ are scalars and the properties of the trace operators,

$$\begin{aligned}E\left[\frac{1}{n}u'P_Z e\right] &= \frac{1}{n}E[tr(u'P_Z e)] = \frac{1}{n}E[tr(P_Z e u')] = \frac{1}{n}tr[E(P_Z e u')]\\ &= \frac{1}{n}tr[E(P_Z)\rho] = \frac{1}{n}E[tr(P_Z)]\rho = \frac{l}{n}\rho\end{aligned}$$

where I use the fact that $tr(E(X)) = E(tr(X))$ and $tr(AB) = tr(BA)$ multiple times. In a similar fashion,

$$E\left[\frac{1}{n}u'P_Z u\right] = \frac{l}{n}$$

Based on the two facts and $l/n \to \alpha$, I can make use of Markov inequality to show that $\frac{1}{n}u'P_Z u \xrightarrow{p} \frac{l}{n}$ and $\frac{1}{n}u'P_Z e \xrightarrow{p} \frac{l}{n}\rho$. Since $l/n \to \alpha$, I can write

$$\frac{u'P_Z e}{n} \xrightarrow{p} \alpha\rho, \frac{u'P_Z u}{n} \xrightarrow{p} \alpha$$

Therefore,

$$\hat{\beta}_{2SLS} - \beta \xrightarrow{p} \frac{\alpha\rho}{c+\alpha}$$

If we do not have many IVs, $\alpha = 0$ and 2SLS estimator is consistent. Otherwise, inconsistency of the above form occurs.

Theorem 2.16.1 (Inconsistency in Large Scale IVs)**.** *If above assumptions hold, together with $E(e_i^2|z_i) < \infty, E(u_i^4|z_i) < \infty$. Then*

$$\hat{\beta}_{OLS} - \beta \xrightarrow{p} \frac{\rho}{c+1}, \hat{\beta}_{2SLS} - \beta \xrightarrow{p} \frac{\alpha\rho}{c+\alpha}$$

Hansen (2019, pp. 447-448) shows why LIML is immune to this problem.

Chapter 3

Generalized Method of Moments

GMM methods utilize the method of moments estimators to identify the values of the parameters of interest. It can be generalized in the sense that the number of moment conditions can be greater than the number of unknown parameters. In fact, we can show that OLS, IV, MLE estimators are part of GMM.

3.1 Framework

The moment equation takes the following form: Let w_i be IID across $i = 1, .., n$, $g_i(w_i, \theta)$ be a $l \times 1$ function of the ith observation, and $\theta \in \mathbb{R}^{k \times 1}$ be the parameter of interest. Note that $l \geq k$. Then, the **moment equation model** is characterized by

$$E[g(w_i, \theta)] = 0$$

We say θ is identified if there is a unique mapping from the data distribution to θ, or that there is a unique θ satisfying $E[g(w_i, \theta)] = 0$. When $l = k$, then we are in a just-identified case. If $l > k$, then we are in the over-identified case in the sense that there is excess information. We will not bother with the case of $l < k$, as they are under-identified and the moment equation model is unsolvable.

- **Just-identified**: In this case, we can work with the sample analogue of $g(w_i.\theta)$ straight away. Define $\bar{g}_n(\theta)$ as

  $$\bar{g}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\theta)$$

  The **method of moments estimators** $\hat{\theta}_{MM}$ is defined as the parameter value which sets

$\bar{g}_n(\theta) = 0$. In other expression:

$$\bar{g}_n(\hat{\theta}_{MM}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\theta}_{MM}) = 0$$

- **General Case**: If we can generalize to include the case where $l > k$, we are likely to run into a situation where $\hat{\theta}_{MM}$ cannot be found. This is because there may be no choice of θ that sets the moment equations to 0, implying the nonexistence of the method of moments estimator. In such case, we work with a different approach.

  Define $J(\theta)$ as

  $$J(\theta) = n\bar{g}_n(\theta)'W\bar{g}_n(\theta)$$

  where $W \in \mathbb{R}^{l\times l}$ is a positive definite weight matrix that is given. n does not really affect our estimation, but it makes the analysis of the asymptotic features much easier. The **generalized method of moments estimator** is defined as the minimizer of the GMM criterion above, or

  $$\hat{\theta}_{GMM} = \arg\min_{\theta} J_n(\theta)$$

Note that when $l = k$, then method of moments estimator solve $\bar{g}_n(\hat{\theta}_{MM}) = 0$. Given that $J(\theta)$ is a positive definite matrix, the method of moments estimator in this case also minimizes $J(\theta)$. Thus, method of moments estimator is a special case of GMM estimator.

If $l > k$, then the problem we should solve is

$$\min_{\theta} J_n(\theta) = n\bar{g}(\theta)'W\bar{g}(\theta)$$
$$\implies \frac{\partial J_n(\theta)}{\partial\theta} = 2n\frac{\partial\bar{g}(\theta)}{\partial\theta}' W\bar{g}(\theta) = 0$$

To obtain this, I have used the following properties of matrix differentiation.

Property 3.1.1 (Matrix Differentiation Tricks)**.** *Let Let $x \in \mathbb{R}^{k\times 1}$ and $A \in \mathbb{R}^{k\times l}, y \in \mathbb{R}^{l\times 1}$. Then we have*

$$\frac{\partial x'Ay}{\partial x} = Ay, \frac{\partial x'Ay}{\partial y} = A'x$$

Let $x \in \mathbb{R}^{k\times 1}$ and $A \in \mathbb{R}^{k\times k}$. Then the following holds

$$\frac{\partial x'Ax}{\partial x} = (A + A')x$$

Example 3.1.1 (Various Examples of GMM). *These are some types of a GMM estimator.*

- *OLS: In a data generating process $y_i = x_i'\beta + e_i$, $x_i \in \mathbb{R}^k$ and the moment condition $E(x_i e_i) = 0$, we have k parameters β_k and k equations for each of the k variables. We can rewrite the moment condition as*

  $$E(x_i(y_i - x_i'\beta)) = 0$$

  And the method of moments estimators imply that we should solve

  $$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) = 0 \iff \hat{\beta}_{OLS} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

- *MLE: If w_i is IID across $i = 1, ..., n$, then we can write the joint likelihood function as*

  $$\prod_{i=1}^{n} f(w_i|\theta)$$

  and thus, the log-likelihood function

  $$\sum_{i=1}^{n} \log f(w_i|\theta)$$

  When we take partial differentiation w.r.t θ,

  $$\sum_{i=1}^{n} \frac{\partial \log f(w_i|\theta)}{\partial\theta} = 0 \implies \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(w_i|\theta)}{\partial\theta} = 0$$

  which is equivalent to $E\left(\frac{\partial \log f(w_i|\theta)}{\partial\theta}\right) = 0$. Practically, $E\left(\frac{\partial \log f(w_i|\theta)}{\partial\theta}\right) = 0$ becomes the moment condition applicable to MLE.

- *IV: Suppose we have a data generating process $y_i = x_i'\beta + e_i$ with x_i being k dimensions.*

Suppose we have an $l > k$ dimensional IV with $E(z_i e_i) = 0$ $\bar{g}(\beta)$ in our context would be

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

Then, we can write our $J_n(\beta)$ as

$$n\left(\frac{Z'y}{n} - \frac{Z'X\beta}{n}\right)' W \left(\frac{Z'y}{n} - \frac{Z'X\beta}{n}\right)$$

By solving the minimization problem we can obtain

$$\begin{aligned}\frac{\partial J_n(\beta)}{\partial \beta} &= -\frac{2}{n}(X'ZWZ'y) + \frac{2}{n}(X'ZWZ'X)\beta = 0\\ &\implies \hat{\beta} = (X'ZWZ'X)^{-1}(X'ZWZ'y)\end{aligned}$$

In fact, we can show that by setting $W = \left(\frac{Z'Z}{n}\right)^{-1}$, the above estimator is in fact, a 2SLS estimator.

3.2 Limiting Distribution of GMM

From above, we know that $\hat{\beta}_{GMM} = (X'ZWZ'X)^{-1}(X'ZWZ'y)$ for the case of an overidentified IV model. We can rewrite this by replacing y with $X\beta + e$, As a result, the limiting distribution of $\hat{\beta}_{GMM}$ is characterized by

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) = \left(\frac{X'Z}{n}W\frac{Z'X}{n}\right)^{-1}\left(\frac{X'Z}{n}W\frac{Z'e}{\sqrt{n}}\right)$$

For convenience, I assume the following

Assumption 3.2.1. *Assume that*

1. $E(z_i x_i') = Q$*, and that* $\frac{Z'X}{n} \xrightarrow{p} Q$
2. $\frac{Z'e}{\sqrt{n}} \xrightarrow{d} N(0, \Omega)$*, where* $\Omega = E(z_i z_i' e_i^2)$
3. *(If we are willing to assume W depends on n, thus W_n): $W_n \xrightarrow{p} W$, where W is a positive definite weight matrix*

Then we can say the following about the distribution of the GMM estimator

Theorem 3.2.1 (Limiting Distribution of the GMM estimator)**.** *If the above assumptions are satisfied, the limiting distribution of the GMM estimator can be characterized by*

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) \xrightarrow{d} N(0, (Q'WQ)^{-1}(Q'W\Omega W'Q)(Q'WQ)^{-1})$$

Where $(Q'WQ)^{-1}$ can be replaced with A, if we stick to the notation in class.

So far, we haven't said too much about W. We just assumed that they are fixed. Even if we suppose that W depends on n somehow, the above theorem still holds, provided that W_n converges in probability to W. Then, the natural question should be "What is the optimal selection of W?". This naturally moves us to...

3.3 Efficient GMM

To select an optimal W matrix, it must be that the resulting variance should be the smallest. If we let $W = \Omega^{-1}$ and work with $(Q'WQ)^{-1}(Q'W\Omega W'Q)(Q'WQ)^{-1} - (Q'\Omega Q)^{-1}$, we can see that it is positive semidefinite[1]. If we put $W = \Omega^{-1}$ and recalculate the variance, we get that the efficient GMM has a limiting distribution characterized by

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) \xrightarrow{d} N(0, (Q'WQ)^{-1})$$

Note that this weighting matrix, which can be rewritten as

$$W = \Omega^{-1} = E(z_i z_i e_i^2)^{-1}$$

is not exactly same as the weighting matrix we used for deriving the 2SLS estimator from GMM, which is $\left(\frac{Z'Z}{n}\right)^{-1}$. In the $W = E(z_i z_i e_i^2)^{-1}$ setup, we allowed for heteroskedasticity. If we impose conditional homoskedasticity in the sense that $E(e_i^2|z_i) = \sigma^2$, we can rewrite $E(z_i z_i' e_i^2)$ as

$$\begin{aligned} E(z_i z_i' e_i^2) &= E(E(z_i z_i' e_i^2|z_i)) = E(z_i z_i' E(e_i^2|z_i)) \\ &= E(z_i z_i' \sigma^2) = E(z_i z_i')\sigma^2 \end{aligned}$$

[1] This is your homework, so I will not proceed further with any more explanations.

Thus, $E(z_iz_i'e_i^2)^{-1} = (Z'Z)^{-1}(\sigma^2)^{-1}$. However, scaling does not affect how the estimator is defined (by calculating $\hat{\beta}_{GMM}$ the scalar scales disappear). So $E(z_iz_i'e_i^2)^{-1}$ is effectively $(Z'Z)^{-1}$ and we get the conclusion that in a conditional homoskedastic setting, 2SLS estimator is the efficient GMM.

3.4 Two-step Optimal GMM

Since we have no idea what the true value of e_i is, the true value of Ω is naturally unknown. Therefore, we require a consistent estimator, denoted as $\widehat{W}$, for $W = \Omega^{-1}$. To do so, we need a preliminary consistent estimator for the true θ, which I will denote with $\tilde{\theta}$. You can use any W matrix for this step (I or $(Z'Z)^{-1}$). Then, define

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(w_i.\tilde{\theta})g(w_i.\tilde{\theta})'$$

since $\Omega = E[g(w_i,\theta)g(w_i,\theta)']$. Then we can identify what $\widehat{\Omega}^{-1}$ is. This will allow us to compute the efficient GMM estimator $\hat{\theta}_{GMM}$

Definition 3.4.1 (Optimal Two-step GMM Estimator)**.** We can compute Optimal Two-step GMM in these steps

1. Compute a preliminary, but consistent estimator for the true θ. Denote this as $\tilde{\theta}$.
2. Using $\Omega = E[g(w_i,\theta)g(w_i,\theta)']$, create a sample analogue of this, defined as $\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(w_i.\tilde{\theta})g(w_i.\tilde{\theta})'$. We can find our $\widehat{\Omega}^{-1}$ here.
3. Using this $\widehat{\Omega}^{-1}$, construct an efficient GMM estimator $\hat{\theta}_{GMM}$

There is another way to come up with a $\widehat{\Omega}^{-1}$ in this context. Define $\bar{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i,\tilde{\theta})$. Then, an alternative definition of $\widehat{\Omega}$ can be written as

$$\widehat{\Omega}^{+} = \frac{1}{n}\sum_{i=1}^{n}(g(w_i,\tilde{\theta}) - \bar{g}(\theta))(g(w_i,\tilde{\theta}) - \bar{g}(\theta))'$$

So why use this? For one thing, both $\widehat{\Omega}$ and $\widehat{\Omega}^{+}$ converge in probability to $E[g(w_i,\theta)g(w_i,\theta)']$.

This is because

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(g(w_i,\tilde{\theta})-\bar{g}(\theta))(g(w_i,\tilde{\theta})-\bar{g}(\theta))' \\
=&\frac{1}{n}\sum_{i=1}^{n}g(w_i,\tilde{\theta})g(w_i,\tilde{\theta})'-\frac{1}{n}\sum_{i=1}^{n}g(w_i,\tilde{\theta})\bar{g}(\theta)'-\frac{1}{n}\sum_{i=1}^{n}\bar{g}(\theta)g(w_i,\tilde{\theta})'+\frac{1}{n}\sum_{i=1}^{n}\bar{g}(\theta)\bar{g}(\theta)' \\
=&\frac{1}{n}\sum_{i=1}^{n}g(w_i,\tilde{\theta})g(w_i,\tilde{\theta})'-\bar{g}(\theta)\bar{g}(\theta)' \quad \left(\because \bar{g}(\theta)=\frac{1}{n}\sum_{i=1}^{n}g(w_i,\tilde{\theta})\right)
\end{aligned}
$$

Under the assumption that $E[g(w_i,\theta)]=0$, (or $E(z_ie_i)=0$) is a correct specification, we have $\bar{g}(\theta)\xrightarrow{p}E(g(w_i,\theta))=0$. So $\widehat{\Omega}^+$ and $\widehat{\Omega}$ both converge to $E[g(w_i,\theta)g(w_i,\theta)']$.

The difference? If it is the case that $E[g(w_i,\theta)]\neq 0$ we view $\widehat{\Omega}^+$ as a robust estimator. Note that

- $\widehat{\Omega}=\frac{1}{n}\sum_{i=1}^{n}g(w_i.\tilde{\theta})g(w_i.\tilde{\theta})'\xrightarrow{p}E[g(w_i,\tilde{\theta})g(w_i,\tilde{\theta})']$
- $\widehat{\Omega}^+=\frac{1}{n}\sum_{i=1}^{n}g(w_i,\tilde{\theta})g(w_i,\tilde{\theta})'-\bar{g}(\theta)\bar{g}(\theta)'\xrightarrow{p}E[g(w_i,\tilde{\theta})g(w_i,\tilde{\theta})']-E[g(w_i,\tilde{\theta})]E[g(w_i,\tilde{\theta})]'=var[g(w_i,\tilde{\theta})]$

In other words, $\widehat{\Omega}$ is inconsistent in case where $E[g(w_i,\tilde{\theta})]=0$ is not guaranteed. Therefore, for tests, such as overidentification tests, it is much more desirable to use $\widehat{\Omega}^+$.

Since we know how to find the optimal $\widehat{W}$, we can estimate the asymptotic variance of the GMM estimators. This can be done by replacing matrices in the original variance with their sample counterparts. In general, we can estimate by

$$
\widehat{V}_{GMM}=\left(\widehat{Q}'\widehat{W}\widehat{Q}\right)^{-1}\left(\widehat{Q}'\widehat{W}\widehat{\Omega}\widehat{W}\widehat{Q}\right)\left(\widehat{Q}'\widehat{W}\widehat{Q}\right)^{-1}
$$

where $\widehat{Q}=\frac{1}{n}\sum_{i=1}^{n}z_ix_i'=\frac{Z'X}{n}$, $\widehat{W}$ is expressed by either $\widehat{\Omega}$ or $\widehat{\Omega}^+$. The residuals used in this estimation is defined as $\hat{e}_i=y_i-x_i'\hat{\beta}_{GMM}$

One alternative to the two-step GMM estimator is contructed by letting weight matrix be an explicit function θ. The criterion function is now

$$
J(\theta)=n\bar{g}_n(\theta)'\left(\frac{1}{n}\sum_{i=1}^{n}g(w_i,\theta)g(w_i,\theta)'\right)^{-1}\bar{g}_n(\theta)
$$

The $\hat{\theta}$ that minimizes this function is the continuously-updated GMM estimator. However, this setup is nonlinear, implying that acquiring this estimator requires numerical methods.

3.5 Wald Test Statistics

Assume that you have found any kind of a GMM estimator. So we are not limiting our discussion to the efficient GMM. The GMM estimator $\hat{\beta}_{GMM}$ that you found has a limiting distribution that can be characterized as

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

Now, we define a fuction $r(\beta) : \mathbb{R}^k \xrightarrow{p} \Theta \in \mathbb{R}^q$ that characterizes the type of limitations we put on our parameter of interest β. I will write $\theta = r(\beta)$. The GMM estimator of θ would naturally be $\hat{\theta}_{GMM} = r(\hat{\beta}_{GMM})$. Then, by delta method, we can characterize the limiting distribution of $\hat{\theta}_{GMM}$ as

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta) \xrightarrow{d} R'N(0, V_\beta) = N(0, \underbrace{R'V_\beta R}_{=V_\theta})$$

where $R = \frac{\partial r(\beta)'}{\partial \beta} \in \mathbb{R}^{k\times q}$.

The hypothesis test can be set up in the following manner

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

Here, we make use of the Wald test statistic, defined as

$$W \equiv n(\hat{\theta} - \theta)'\widehat{V}_\theta^{-1}(\hat{\theta} - \theta)$$

where $W \xrightarrow{d} \chi^2_q$ under H_0. If we conduct a test with a significance level α, we need to find a critical value C such that $\alpha = 1 - F(C)$ where F is the CDF for χ^2_q. In such test, we reject H_0 if $W > C$ and do not reject if otherwise.

3.6 Restricted GMM

There are many cases where we may want to impose restrictions on the coefficients (e.g. hypothesis testing). In this section, we will consider $r(\beta) = 0$ as a constraint on β. Finding a restricted GMM (or constrained GMM) is not too different from the unconstrained GMM we have been solving so far. The only difference is that we are now solving a constrained minimization problem. The constrained GMM estimator, in math, is the solution to the following problem

$$\hat{\beta}_{CGMM} = \arg\min_{\beta} J_n(\beta) \text{ s.t. } r(\beta) = 0$$

For instance, we may consider a linear constraint on β coefficients in the form of $R'\beta = c$, where $R \in \mathbb{R}^{k\times q}$ with q being number of constraints. We can write this out in a Lagrangian form.

$$n\bar{g}(\beta)'W\bar{g}(\beta) + \lambda'[R'\beta - c]$$

where $\lambda \in \mathbb{R}^q$ is the vector of Lagrange multipiers.

Suppose that the moment condition given is $E(z_i e_i) = 0$. Then

$$\bar{g}(\beta) = \left(\frac{Z'y}{n} - \frac{Z'X\beta}{n}\right)$$

To find the optimal β, we differentiate the objective function with respect to β. This will become

$$\frac{y'ZWz'Y}{n} - \frac{y'ZWZX\beta}{n} - \frac{\beta'X'ZWZ'y}{n} + \frac{\beta'X'ZWZ'X\beta}{n} + \lambda'R'\beta - \lambda'C$$
$$\xrightarrow{\partial\beta} -2\frac{X'ZWZ'y}{n} + 2\frac{X'ZWZ'X\beta}{n} + R\lambda = 0$$

We can use the knowledge that the unconstrained GMM can be written as $\left(\frac{X'ZWZ'X}{n}\right)^{-1}\frac{X'ZWZ'y}{n}$. So the trick is to pre-multiply $-R'\left(\frac{X'ZWZ'X}{n}\right)^{-1}$. This would change the first order condition to

$$2R'\hat{\beta}_{GMM} - 2R'\beta - R'\left(\frac{X'ZWZ'X}{n}\right)^{-1}R\lambda = 0$$

Since $R'\beta = c$ under H_0, we write

$$\lambda = 2\left(R'\left(\frac{X'ZWZ'X}{n}\right)^{-1}R\right)^{-1}R'(\hat{\beta}_{GMM} - \beta)$$

We can then put this back into the λ in the first order condition to get

$$-2\frac{X'ZWZ'y}{n}+2\frac{X'ZWZ'X\beta}{n}+2R\left(R'\left(\frac{X'ZWZ'X}{n}\right)^{-1}R\right)^{-1}R'(\hat{\beta}_{GMM}-\beta)=0$$

Note that in this equation, n is commonly divided in all terms and 2 is multiplied in all terms. So we can practically get rid of them. So we can further rewrite the above as

$$(X'ZWZ'X)\beta-R(R'(X'ZWZ'X)^{-1}R)^{-1}\underbrace{R'\beta}_{=c}=X'ZWZ'y-R(R'(X'ZWZ'X)^{-1}R)^{-1}R'\hat{\beta}_{GMM}$$

Or

$$\begin{aligned}&(X'ZWZ'X)\beta=X'ZWZ'y-R(R'(X'ZWZ'X)^{-1}R)^{-1}(R'\hat{\beta}_{GMM}-c)\\ \implies\ &\hat{\beta}_{CGMM}=\hat{\beta}_{GMM}-(X'ZWZ'X)^{-1}R(R'(X'ZWZ'X)^{-1}R)^{-1}(R'\hat{\beta}_{GMM}-c)\end{aligned}$$

Note that if we premultiply R' to both sides, we can get

$$\begin{aligned}R'\hat{\beta}_{CGMM}&=R'\hat{\beta}_{GMM}-R'(X'ZWZ'X)^{-1}R(R'(X'ZWZ'X)^{-1}R)^{-1}(R'\hat{\beta}_{GMM}-c)\\&=R'\hat{\beta}_{GMM}-R'\hat{\beta}_{GMM}+c=c\end{aligned}$$

which shows that the CGMM satisfies the constraint.

3.7 Distance Test

When we looked at the Wald Test Statistics, we assumed a linear form of constraints for $r(\beta)=\theta$. If we are working with a possibly nonlinear form of $r(\beta)$, we can use an alternative criterion-based statistic. This is where the GMM Distance statistic comes in. The basic idea is to compare unrestricted and restricted estimators by contrasting the criterion functions. Again, we define

$$J(\beta)=n\bar{g}_n(\beta)'\widehat{\Omega}^{-1}\bar{g}_n(\beta)$$

where $\bar{g}_n(\beta)$ is $\frac{1}{n}\sum_{i=1}^{n} g(w_i, \beta)$ and $\widehat{\Omega}$ is the efficient weight matrix. With the unconstrained estimator and the constrained estimator, we can write

$$J(\hat{\beta}_{GMM}) = n\bar{g}_n(\hat{\beta}_{GMM})'\widehat{\Omega}^{-1}\bar{g}_n(\hat{\beta}_{GMM})$$
$$J(\hat{\beta}_{CGMM}) = n\bar{g}_n(\hat{\beta}_{CGMM})'\widehat{\Omega}^{-1}\bar{g}_n(\hat{\beta}_{CGMM})$$

Now we have all the ingredients to define the distance statistic D, defined as

$$D \equiv J(\hat{\beta}_{CGMM}) - J(\hat{\beta}_{GMM})$$

Note that $D \geq 0$ by construction. This is because $J(\hat{\beta}_{CGMM})$ is from the minimization problem that is constrained and the other is an unconstrained problem, which usually results in lower values.

As we did for hypothesis testing on Wald test statistic, we can use D for hypothesis tests of the following setup

$$H_0 : r(\beta) = \theta, \ \ H_1 : r(\beta) \neq \theta$$

Under H_0, D converges in distribution to χ^2_q. We can find a critical value c s.t. $\alpha = 1 - F(c)$, where F is the CDF for χ^2_q and reject H_0 if $D > c$. The idea is that if H_0 is true, then imposing the restriction does not alter the moment equations greatly, making it a sensible restriction. Otherwise, the moment equation changes greatly, making it unreasonable constraint.

In fact, we can show that when $r(\beta)$ is linear, D becomes identical to the Wald Test Statistic W. We have shown what $\hat{\beta}_{CGMM}$ is. Then we can write

$$\begin{aligned}
\bar{g}_n(\hat{\beta}_{CGMM}) &= \frac{Z'y - Z'X\hat{\beta}_{CGMM}}{n} \\
&= \frac{\overbrace{Z'y - Z'X(\hat{\beta}_{GMM}}^{=\bar{g}_n(\hat{\beta}_{GMM})} - \overbrace{(X'ZWZ'X)^{-1}R(R'(X'ZWZ'X)^{-1}R)^{-1}(R'\hat{\beta}_{GMM} - c))}^{=\hat{\beta}_{GMM}-\hat{\beta}_{CGMM}}}{n} \\
&= \bar{g}_n(\hat{\beta}_{GMM}) + \frac{Z'X(\hat{\beta}_{GMM} - \hat{\beta}_{CGMM})}{n}
\end{aligned}$$

Thus,

$$
\begin{aligned}
J(\hat{\beta}_{CGMM}) &= n\left(\bar{g}_n(\hat{\beta}_{GMM}) + \frac{Z'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})}{n}\right)' W \left(\bar{g}_n(\hat{\beta}_{GMM}) + \frac{Z'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})}{n}\right) \\
&= n\bar{g}_n(\hat{\beta}_{GMM})'W\bar{g}_n(\hat{\beta}_{GMM}) + \bar{g}_n(\hat{\beta}_{GMM})'WZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM}) \\
&+ (\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})'X'ZW\bar{g}_n(\hat{\beta}_{GMM}) + \frac{1}{n}(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})'X'ZWX'Z(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})
\end{aligned}
$$

where $\bar{g}_n(\hat{\beta}_{GMM})'WZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})$ terms can be erased because

$$
\begin{aligned}
\bar{g}_n(\hat{\beta}_{GMM})'WZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM}) &= \frac{y'ZWZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM}) - \hat{\beta}'_{GMM}X'ZWZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})}{n} \\
&= \frac{y'ZWZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM}) - y'ZWZ'X(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})}{n} = 0
\end{aligned}
$$

since $\hat{\beta}_{GMM} = (X'ZWZ'X)^{-1}X'ZWZ'y$. Therefore,

$$
J(\hat{\beta}_{CGMM}) = J(\hat{\beta}_{GMM}) + \frac{1}{n}(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})'X'ZWX'Z(\hat{\beta}_{GMM}-\hat{\beta}_{CGMM})
$$

So $J(\hat{\beta}_{CGMM}) - J(\hat{\beta}_{GMM})$ becomes a Wald Test Statistic.

3.8 Overidentification Tests

In this section, we generalize the overidentification test we learned in 2SLS setup to a GMM setting. It is generalized in the sense that we can allow for heteroskedasticity whereas for 2SLS overidentification test, we relied on the assumption of homoskedasticity. If $\dim(g_i) = l > k = \dim(\beta)$, it is possible that there exists no β s.t. $E[g(w_i, \beta)] = 0$ is satisfied. Therefore, the overidentifying restrictions become testable. Effectively, we are testing the hypothesis

$$
H_0 : E[g(w_i, \beta)] = 0 \text{ vs. } H_1 : E[g(w_i, \beta)] \neq 0
$$

Let β_0 be the true value for the parameter of interest. Then $\bar{g}_n(\beta_0) = \frac{Z'y - Z'X\beta_0}{n} = \frac{Z'e}{n}$ has a limiting distribution characterized by

$$
\sqrt{n}\bar{g}_n(\beta_0) \xrightarrow{d} N(0, \Omega), \quad \Omega \in \mathbb{R}^{l \times l}
$$

and thus $J(\beta_0) \xrightarrow{d} \chi^2_l$. Since we are trying to estimate what β_0 is, we need to use a GMM estimator to build our test statistic $J = J(\hat{\beta}_{GMM})$, which has a limiting distribution χ^2_{l-k} under H_0. Like before, we can find c s.t. $\alpha = 1 - F(c)$, and then reject H_0 when $J > c$. However, it should be noted that when the H_0 is rejected, we only know that some moment condition is violated. We cannot pick out which. Nevertheless, rejection of the H_0 is a bad sign.

We can apply overidentification test on a (strict) subset of instruments whose validity is uncertain. To do this, we can partition z_i into two sets - $z_{ai} \in \mathbb{R}^{l_a}$ and $z_{bi} \in \mathbb{R}^{l_b}$. We are uncertain about z_{bi} and want to test

$$H_0 : E(z_{bi}e_i) = 0, \text{ vs. } H_1 : E(z_{bi}e_i) \neq 0$$

We can construct the test statistic in a following manner. First, estimate the model by the efficient GMM with only the z_{ai} set of instruments and obtain the GMM criterion. This will be denoted as J_a. Then, estimate the model with the full set of instruments and obtain a separate GMM criterion, denoted as $J_{a,b}$. Then, create a test statistic

$$C = J_{a,b} - J_a \xrightarrow{d} \chi^2_{l-l_a=l_b}$$

The idea is similar to that of a distance test. If the C value is low, then the moment condition for the larger set of instruments are not too different from the moment condition from the smaller (but surer) set of instruments. Then, we find a critical value c for a significance level α and reject the null hypothesis if $C > c$.

One example of a subset overidentification test is an endogeneity test. Assume a following data generating process

$$y_i = x'_{1i}\beta_1 + x'_{2i}\beta_2 + e_i$$

where x_{1i} is exogenous but x_{2i} may not be. Then, we want to test

$$H_0 : E(x_{2i}e_i) = 0, \text{ vs. } H_1 : E(x_{2i}e_i) \neq 0$$

Let $z_i = (x_{1i} \;\; z_{2i})' \in \mathbb{R}^l$ and assume that $E(z_ie_i) = 0$. Then we can create the test statistics as follows. First, estimate the efficient GMM by using (x_{1i}, z_{2i}) to instrument (x_{1i}, x_{2i}). Then obtain the GMM criterion $\tilde{J}$. Next, work with a larger set by using (x_{1i}, x_{2i}, z_{2i}) to instrument (x_{1i}, x_{2i}). After this, we can get the GMM criterion $\widehat{J}$. The relevant test statistic is

$$C = \widehat{J} - \tilde{J} \xrightarrow{d} \chi^2_{k_2}$$

As before, we find a relevant critical value c for a significance level α and reject H_0 of exogeneity for x_2 if $C > c$.

Again, the idea is as follows. We know that x_{1i} and z_{2i} are exogenous. If x_{2i} is also exogenous, then the GMM criterion obtained by using all three should not be too different from GMM criterion from variables x_{1i} and z_{2i}[2]. If the difference between the two GMM criterion is sufficiently large, we have a reason to suspect that, given that x_{1i} and z_{2i} is exogenous for sure, that the moment condition involving x_{2i} variables are vastly different from moment conditions from exogenous variables. Thus, x_{2i} could be endogenous.

3.9 Conditional Moment Conditions

In some cases, the moment condition for the data generating process can be expressed in terms of conditional moments. Assume that the data generating process and conditional moment condition of interest is

$$y_i = m(x_i, \beta^0) + e_i \quad E(e_i(\beta)|z_i) = 0$$

where β^0 is the true value of β, z_i is the instrument and $m(x_i, \beta)$ can be either linear or nonlinear. Usually, these conditions are more stronger in the sense these sets of conditions imply that the unconditional moment conditions assumed throughout the discussion holds. In fact, we can show that if $E(e_i|z_i) = 0$, then $E(h(z_i)e_i) = 0$ for measurable function $h(\cdot)$. The other way may not hold.

Example 3.9.1. *Suppose that $E(z_ie_i) = 0$, then is it the case that $E(e_i|z_i) = 0$ hold? There is this counterexample: Let z_i and e_i be characterized as*

$$z_i = \begin{cases} 1 & \Pr(z_i = 1) = 1/2 \\ 2 & \Pr(z_i = 2) = 1/2 \end{cases}, \quad e_i = \begin{cases} 1 & (\textit{if } z_i = 1) \\ -1/2 & (\textit{if } z_i = 2) \end{cases}$$

Then $E[z_ie_i]$ is $\frac{1}{2}(1 \times 1) + \frac{1}{2}(2 \times (-1/2)) = \frac{1}{2} - \frac{1}{2} = 0$. As for $E[e_i|z_i]$, we get $E[e_i|z_i = 1] = 1$ and $E[e_i|z_i = 2] = -1/2$. Therefore, the conditional mean is not 0.

So how do we solve when given conditional moments? One is to use the idea that $E[e_i|z_i] = 0 \implies E[h(z_i)e_i] = 0$ to construct some number of instruments and solve a

[2] It can be slightly different, since the weighting matrix used in calculating the GMM criterion is usually different. This can lead to some discrepancies even if H_0 turns out to be true.

GMM with unconditional moments[3]. The other is to construct an optimal instrument as Chamberlain (1987) found. For this, we need to define these equations

$$R(z_i) = E\left[\frac{\partial e_i(\beta^0)}{\partial \beta} \mid z_i\right] \in \mathbb{R}^k$$
$$\sigma^2(z_i) = E[e_i(\beta^0)^2|z_i]$$

Then, the optimal instrument is defined by

$$A_i = -\frac{R(z_i)}{\sigma^2(z_i)} \in \mathbb{R}^k$$

yielding the optimal moment

$$g_i^*(\beta) = A_i e_i(\beta)$$

Thus, by working with the $E[g_i^*(\beta)] = 0$ gives us the best GMM estimator possible (best being the lowest possible variance). In class, we have shown that if we actually knew the true σ^2, this gets us the GLS estimator.

[3] There is no consensus on the optimal number of optimal instruments yet

Chapter 4

Panel Data

4.1 Basic Framework

To motivate the discussion going forward, consider the following setting. Let y and $x = (x_1, x_2, ..., x_k)$ be the observable factors. Denote α as an unobservable random variable that is incorporated into the data generating process additively. Then, we can write $E[y|x, \alpha]$ as

$$E[y|x, \alpha] = x'\beta + \alpha$$

We are interested in estimating β. To see whether we can find a consistent estimator for β, we need to see how α is correlated with x. If α is independent from x, the it is not different from the idiosyncratic error and β is consistently estimable. However, in most cases, α could be correlated with x. If this is the case, then we cannot find a consistent estimator for β unless we find a perfect proxy for α. However, if α is fixed across time for an individual and we have access to panel data, we can address this issue. Most discussion of panel data in class will focus on α representing an unobservable trait inherent for individuals.

In a panel data setting, the data now has two dimensions - dimensions across different unit of observation i and across time t. In maths,

$$(y_{it}, x_{it}) \text{ where } i = 1, ..., n, \text{ and } t = 1, ..., T$$

For our discussion, we will fix T and let the (y_{it}, x_{it}) be IID across $i = 1, ..., n$. Unless otherwise state, the asymptotics will be carried out by extending n to infinity.

To be more concrete, we can write the data generating process as

$$y_{it} = x_{it}'\beta + \alpha_i + u_{it} \quad ^{1}$$

where x_{it} is an observable variable that varies across individuals and time periods. α_i is the **individual (fixed) effect** that is unobservable. As such, we can define $v_{it} = \alpha_i + u_{it}$ and rewrite the data generating process as

$$y_{it} = x_{it}'\beta + v_{it}$$

Depending on whether α_i is correlated with x_{it} or not, we can categorize α_i into the following

- **Random Effects**: There is no correlation between the observables and α_i
- **Fixed Effects**: The correlation between x_{it} and α_i is nonzero.

In case you are looking into some old textbooks, the categorization is slightly different. If α_i is considered to be a parameter to be estimated, the old textbooks refers to this as fixed effects. If α_i is a random variable, it was called a random effect. Note that in the above discussion, α_i is random variable in both fixed and random effects. You should see that what matters now is whether α_i is correlated with x_{it} or not.

4.2 Strict Exogeneity

Before moving further, the following assumption is required to show whether the panel estimates are consistent or not

Assumption 4.2.1 (Strict Exogeneity)**.** *We say that the regressor x_{it} is* ***strictly exogenous*** *with respect to u_{it} if*

$$E[y_{it}|x_{i1},..,x_{iT},\alpha_i] = E[y_{it}|x_{it},\alpha_i]$$

which boils down to

$$E[u_{it}|x_{i1},..,x_{iT},\alpha_i] = E[u_{it}|x_{it},\alpha_i] = 0$$

[1] Note that we can be interested in the time fixed effect, usually denoted as γ_t. The analysis of such effects is largely similar to individual fixed effects.

In words, strict exogeneity implies that once x_{it}, α_i is controlled for, $x_{is}, s \neq t$ has no partial effect on y_{it}. The above condition also implies that

$$E[x_{is}u_{it}] = 0 \ (s,t) = 1, .., T$$

while this itself is a weaker condition that $E[u_{it}|x_{i1}, .., x_{iT}, \alpha_i] = 0$, it is sufficient for most of the consistency analysis. However, it is stronger than the exogeneity assumptions of the previous chapters. Previously it was enough for the regressors to be exogenous with the contemporaneous e_{it}. Now we assume x_{it} to be exogenous with all error terms. [2]

Also note that this condition is silent about the correlation between x_{it} and α_i.

4.3 Pooled OLS (POLS)

Pooled OLS estimation practically ignores the panel structure of the data and involves applying a typical OLS estimation on the data generating process with time and individual dimension. Given that our model is

$$y_{it} = x_{it}'\beta + \underbrace{\alpha_i + u_{it}}_{=v_{it}}$$

The estimator can be written as

$$\hat{\beta}_{POLS} = \left(\sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}x_{it}'\right)^{-1} \sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}y_{it}$$

Under the situation where $cov(x_{it}.\alpha_i) \neq 0$, we can see that $\hat{\beta}_{POLS}$ is inconsistent. Note that we can re-write the POLS estimator in terms of the true β

$$\hat{\beta}_{POLS} - \beta = \left(\sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}x_{it}'\right)^{-1} \sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}v_{it}$$

The trick is to show that $E[x_{it}v_{it}] \neq 0$ for $t = 1, .., T$. Since I can write $E[x_{it}(\alpha_i + u_{it})]$ instead, it can be seen that unless $cov(x_{it}, \alpha_i) = 0$, the $E[x_{it}\alpha_i]$ term will remain. Thus, POLS estimator is not consistent.

[2] Yes, you may think that this assumption is extreme. Later in the panel data discussion, we will talk about weaker version of this assumption - predetermined regressors.

4.4 First Difference (FD)

For this to work, we need $T \geq 2$. To obtain the first difference estimator, we need to subtract the original data generating process by one lag of y_{it}. Or

$$\Delta y_{it} = \Delta x_{it}'\beta + \Delta u_{it} \ \ (i = 1, ..., n \text{ and } t = 2, .., T)$$

where $\Delta y_{it} = y_{i,t} - y_{i,t-1}$ and similarly for other variables. Notice that since α_i is same across $t = 1, .., T$ (but different for each i), it vanishes. By taking an OLS, we can obtain

$$\hat{\beta}_{FD} = \left(\sum_{i=1}^{n} \sum_{t=2}^{T} \Delta x_{it} \Delta x_{it}' \right)^{-1} \sum_{i=1}^{n} \sum_{t=2}^{T} \Delta x_{it} \Delta y_{it}$$

We can show that this is a consistent estimator. Write

$$\hat{\beta}_{FD} - \beta = \left(\sum_{i=1}^{n} \sum_{t=2}^{T} \Delta x_{it} \Delta x_{it}' \right)^{-1} \sum_{i=1}^{n} \sum_{t=2}^{T} \Delta x_{it} \Delta u_{it}$$

We need to show that $E[\Delta x_{it} \Delta u_{it}] = 0$. This can be written

$$\begin{aligned} E[\Delta x_{it} \Delta u_{it}] &= E[(x_{it} - x_{i,t-1})(u_{it} - u_{i,t-1})] \\ &= E[x_{it}u_{it}] - E[x_{it}u_{i,t-1}] - E[x_{i,t-1}u_{it}] + E[x_{i,t-1}u_{i,t-1}] \\ &= 0 - 0 - 0 + 0 = 0 \end{aligned}$$

Therefore, $\hat{\beta}_{FD}$ is consistent. Another requirement for this to be defined is that $\left(\sum_{t=2}^{T} \Delta x_{it} \Delta x_{it}' \right)$ should be a full column matrix so that the inverse matrix is defined. This would effectively rule out time-constant regressors.

4.5 Within Estimators (WE)

This shares a similar idea with First Difference in the sense that it attempts to weed out α_i. This approach uses a different method. I will define new notations that involve variables that are averaged across time. Write

$$\bar{y}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it}$$

and similarly for other variables. By average over time across all variables, we can get a cross-sectional equation

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

The key is that even if y_i is average across time, we still get α_i itself. This is because $\frac{1}{T}\sum_{t=1}^{T}\alpha_i = (T\alpha_i)/T = \alpha_i$. So subtract the cross-sectional equation from the original data generating process to get

$$\tilde{y}_{it} = \tilde{x}_{it}'\beta + \tilde{u}_{it}, \quad (i = 1, .., n, \text{ and } t = 1, .., T)$$

where $\tilde{y}_{it} = y_{it} - \bar{y}_i$. The within estimator is obtained by taking an OLS to above equation

$$\hat{\beta}_{WE} = \left(\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{y}_{it}$$

To show consistency, rewrite the above as

$$\hat{\beta}_{WE} - \beta = \left(\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}$$

$E[\tilde{x}_{it}\tilde{u}_{it}]$ can be written as

$$E[\tilde{x}_{it}\tilde{u}_{it}] = E[x_{it}u_{it}] - E[x_{it}\bar{u}_i] - E[\bar{x}_i u_{it}] + E[\bar{x}_i\bar{u}_i]$$

Since $\bar{x}_i, \bar{u}_i$ incorporates regressors and errors from all time periods, applying strict exogeneity (and strict exogeneity only) reduces the above equation to 0. Therefore, $\hat{\beta}_{WE}$ is consistent. Note that for this estimator, we REALLY need a strict exogeneity assumption. Anything weaker than this could make this estimator inconsistent. In addition, we need that $E\left[\tilde{x}_{it}\tilde{x}_{it}'\right]$ be a full column rank for its inverse to be defined. Time-constant regressors are ruled out.

4.6 Least Square Dummy Variables (LSDV)

This comes from viewing α_i as a parameter to be estimated. Let Dk_i be the dummy variable that equals 1 if $i = k$ and 0 otherwise. The idea is to put a total of $N - 1$ of such dummy variables into the regression[3]. Therefore, we work with

$$y_{it} = x_{it}'\beta + D1_i\alpha_1 + ... + D(n-1)_i\alpha_{n-1} + u_{it}$$

[3] This is assuming that x_{it} contains a vector of 1's for a 'overall' constant.

What is going on here is that each individual has his/her own constant term. For the n'th individual, the constant term is represented by the β_0, the coefficient on the column vector of x_{it}. For $k(\neq n)$'th individual, the intercept term is $\beta_0 + \alpha_k$.

4.7 Some Interesting Topics on Fixed Effects Panel Estimators

- **WE vs FD**: When $T = 2$, it can be shown that they are numerically equal. I start with the WE, written as
$$\left(\sum_{i=1}^{n}\sum_{t=1}^{2}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\left(\sum_{i=1}^{n}\sum_{t=1}^{2}\tilde{x}_{it}\tilde{y}_{it}\right)$$

  Each term can be replaced by
$$\begin{aligned}\sum_{t=1}^{2}\tilde{x}_{it}\tilde{x}_{it}' &= \sum_{t=1}^{2}\left(x_{it}-\frac{x_{i1}+x_{i2}}{2}\right)\left(x_{it}-\frac{x_{i1}+x_{i2}}{2}\right)' \\ &= \sum_{t=1}^{2}x_{it}x_{it}' - \frac{x_{i1}x_{i1}'+x_{i1}x_{i2}'+x_{i2}x_{i1}'+x_{i2}x_{i2}'}{2} \\ &= \frac{x_{i1}x_{i1}'-x_{i1}x_{i2}'-x_{i2}x_{i1}'+x_{i2}x_{i2}'}{2} \\ &= \frac{(x_{i1}-x_{i2})(x_{i1}-x_{i2})'}{2} = \frac{\Delta x_{i2}\Delta x_{i2}'}{2}\end{aligned}$$

  and
$$\begin{aligned}\sum_{t=1}^{2}\tilde{x}_{it}\tilde{y}_{it} &= \sum_{t=1}^{2}\left(x_{it}-\frac{x_{i1}+x_{i2}}{2}\right)\left(y_{it}-\frac{y_{i1}+y_{i2}}{2}\right) \\ &= \sum_{t=1}^{2}x_{i}y_{i} - \frac{x_{i1}y_{i1}+x_{i1}y_{i2}+x_{i2}y_{i1}+x_{i2}y_{i2}}{2} \\ &= \frac{x_{i1}y_{i1}-x_{i1}y_{i2}-x_{i2}y_{i1}+x_{i2}y_{i2}}{2} \\ &= \frac{(x_{i1}-x_{i2})(y_{i1}-y_{i2})}{2} = \frac{\Delta x_{i2}\Delta y_{i2}}{2}\end{aligned}$$

  By combining the two, I get $\left(\sum_{i=1}^{n}\Delta x_{i2}\Delta x_{i2}'\right)^{-1}\sum_{i=1}^{n}\Delta x_{i2}\Delta y_{i2}$, which is equivalent to FD when $T = 2$.

  When $T \geq 3$, they are no longer equal. So it begs the question as to which estimator to use. This depends on the structure of the error terms. If u_{it} is free from serial correlation (or IID), then taking a first difference would introduce serial correlation. This is because

$$\begin{aligned} cov(\Delta u_{it}, \Delta u_{i,t-1}) &= E[u_{it}u_{it-1}] - E[u_{it}u_{it-2}] - E[u_{it-1}u_{it-1}] + E[u_{it-1}u_{it-2}] \\ &= 0 - 0 - var(u_{it-1}) + 0 \neq 0 \end{aligned}$$

As such, first difference in this situation suffers from inconsistency problems arising due to serial correlation.

There may be a case when Δu_{it} is serially uncorrelated. For instance, u_{it} could be a random walk process in the sense that

$$u_{it} = u_{it-1} + \eta_{it} \;\; (E[\eta_{it}] = 0, E[\eta_{it}\eta_{is}] = 0(s \neq t), var(u_{it}) = \sigma^2)$$

If we use a first difference estimator here, we get to obtain the most efficient estimator.

- **WE=LSDV** To show that they are numerically equal, it helps to know how these estimators can be written in terms of matrices and understand what Kronecker product is. Note the following

Definition 4.7.1 (Kronecker Product)**.** Let $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$. Then their Kronecker product $A \otimes B$ is defined as

$$A \otimes B = \begin{pmatrix} a_{11}B & ... & a_{1n}B \\ ... & ... & ... \\ a_{m1}B & ... & a_{mn}B \end{pmatrix} \in \mathbb{R}^{mp\times nq}$$

where $B = \begin{pmatrix} b_{11} & ... & b_{1q} \\ ... & ... & ... \\ b_{p1} & ... & b_{pq} \end{pmatrix}$. Some of its properties are

- $A \otimes (B + C) = A \otimes B + A \otimes C$ (B, C are of same size)
- $(A \otimes B)(C \otimes D) = AC \otimes BD$ (Assuming multiplication is implementable)
- $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$ (Assuming inverses are defined)
- $(A \otimes B)' = (A' \otimes B')$

I will first write WE in terms of matrices. for each individual i, we can stack up total of

T observations and write

$$\mathbf{y}_i = \mathbf{X}_i\beta + \alpha_i 1_T + \mathbf{u}_i \in \mathbb{R}^T$$

where 1_T is the $T \times 1$ vector of $1'$s as elements. Now I define $Q_T \equiv I_T - 1_T(1_T'1_T)^{-1}1_T'$ Note that $(1_T'1_T)^{-1} = T^{-1}$ and $1_T1_T'$ is the T-dimensional square matrix of $1'$s as elements. This is symmetric and idempotent. Also note that

$$Q_T 1_T = 0, Q_T\mathbf{y}_i = \mathbf{y}_i - \frac{1}{T}\begin{pmatrix} \sum_t y_{it} \\ ... \\ \sum_t y_{it} \end{pmatrix} = \begin{pmatrix} y_{i1} - \frac{1}{T}\sum_t y_{it} \\ ... \\ y_{iT} - \frac{1}{T}\sum_t y_{it} \end{pmatrix} = \tilde{\mathbf{y}}_i, Q_T\mathbf{X}_i = \tilde{\mathbf{X}}_i, Q_T\mathbf{u}_i = \tilde{\mathbf{u}}_i$$

This shows that pre-multiplying Q_T to $\mathbf{y}_i = \mathbf{X}_i\beta + \alpha_i 1_T + \mathbf{u}_i \in \mathbb{R}^T$ gives us the demeaned version of the original data generating process. So we need to take OLS on

$$Q_T\mathbf{y}_i = Q_T\mathbf{X}_i\beta + Q_T\mathbf{u}_i$$
$$\implies \hat{\beta}_{WE} = \left(\sum_{i=1}^n \mathbf{X}_i' Q_T \mathbf{X}_i\right)^{-1} \sum_{i=1}^n \mathbf{X}_i' Q_T \mathbf{y}_i$$

We can characterize the LSDV estimator from the same original data generating process. What we need to do is to stack $\mathbf{y}_i$ and $\mathbf{u}_i$ for each i and obtain a nT dimensional vector $\mathbf{y}$ and $\mathbf{u}$. Similarly, stack each $\mathbf{X}_i$ to get an $nT \times k$ matrix $\mathbf{X}$. Since each α_i exists for each individual i, the idea here is to use a Kronecker product and express the individual fixed effect as

$$\begin{pmatrix} 1_T & ... & 0 \\ ... & ... & ... \\ 0 & ... & 1_T \end{pmatrix}\begin{pmatrix} \alpha_1 \\ ... \\ \alpha_n \end{pmatrix} = (I_n \otimes 1_T)\alpha \in \mathbb{R}^{nT\times n} \times \mathbb{R}^{n\times 1} = \mathbb{R}^{nT\times 1}$$

Combine what we know to get

$$\mathbf{y} = \mathbf{X}\beta + (\underbrace{I_n \otimes 1_T}_{=D})\alpha + \mathbf{u}$$

We then use the Frisch-Waugh-Lovell theorem (Chapter 3, Hansen (2019)) to get

$$\hat{\beta}_{LSDV} = (\mathbf{X}' M_D \mathbf{X})^{-1}(\mathbf{X}' M_D \mathbf{y})$$

where $M_D = I_{nT} - D(D'D)^{-1}D'$. Rewrite $D(D'D)^{-1}D'$ using the properties mentioned above as

$$\begin{aligned} D(D'D)^{-1}D' &= (I_n \otimes 1_T)[(I_n \otimes 1_T)'(I_n \otimes 1_T)]^{-1}(I_n \otimes 1_T)' \\ &= (I_n \otimes 1_T)[(I_n \otimes 1_T')(I_n \otimes 1_T)]^{-1}(I_n \otimes 1_T') \\ &= (I_n \otimes 1_T)[(I_n \otimes 1_T'1_T)]^{-1}(I_n \otimes 1_T') \\ &= (I_n \otimes 1_T)[I_n \otimes (1_T'1_T)^{-1}](I_n \otimes 1_T') \\ &= I_n \otimes 1_T(1_T'1_T)^{-1}1_T' \end{aligned}$$

Thus, $M_D = I_{nT} - I_n \otimes 1_T(1_T'1_T)^{-1}1_T' = I_n \otimes I_T - I_n \otimes 1_T(1_T'1_T)^{-1}1_T'$. By using the first property mentioned with the Kronecker product, this is equal to $I_n \otimes (I_T - 1_T(1_T'1_T)^{-1}1_T') = I_n \otimes Q_T$. So $\hat{\beta}_{LSDV}$ is equal to

$$\hat{\beta}_{LSDV} = (\mathbf{X}'(I_n \otimes Q_T)\mathbf{X})^{-1}(\mathbf{X}'(I_n \otimes Q_T)\mathbf{y})$$

Since $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ ... \\ \mathbf{X}_n \end{pmatrix}$, $I_n \otimes Q_T = \begin{pmatrix} Q_T & 0 & 0 \\ ... & ... & ... \\ 0 & 0 & Q_T \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ ... \\ \mathbf{y}_n \end{pmatrix}$, we get

$$\hat{\beta}_{LSDV} = \left(\sum_{i=1}^{n} \mathbf{X}_i' Q_T \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' Q_T \mathbf{y}_i = \hat{\beta}_{WE}$$

and we are done.

4.8 Random Effects

Consider the following data generating process

$$y_{it} = x_{it}'\beta_1 + \alpha_i + u_{it}, \ \ (i = 1, .., n \text{ and } \ t = 1, .., T)$$

where we assume that $E(\alpha_i) = 0$, $E(\alpha_i u_{it}) = 0$, u_{it} is IID across i and t and independent of x_{it}. On top of this, a key assumption that separates random effects from fixed effects is α_i and x_{it} is now **uncorrelated**. The major consequence of this assumption is that even the POLS estimator becomes consistent. (We also still assume that strict exogeneity holds.)

To show this, I apply an OLS procedure on the data generating process above, while writing $v_{it} = \alpha_i + u_{it}$. This gets us

$$\hat{\beta}_{POLS} = \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} x_{it}' \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} y_{it}$$

which can alternatively be written as

$$\hat{\beta}_{POLS} - \beta_1 = \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} x_{it}' \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} v_{it}$$

We need to work with the asymptotics of $\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} v_{it}$. Since

$$\begin{aligned} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} v_{it} &= \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}(\alpha_i + u_{it}) \\ &= \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}\alpha_i + \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} u_{it} \\ &\overset{p}{\to} E(x_{it}\alpha_i) + E(x_{it} u_{it}) = 0 + 0 = 0 \end{aligned}$$

The last summation terms converge in probability to 0 now. The reason POLS in fixed effects setting was inconsistent was that this term converges in probability something nonzero. Therefore, POLS is now consistent as well.

However, this is not really the best we can do. The reason is that the composite error from different time periods are serially correlated. This can be seen by

$$\begin{aligned} cov(v_{it}, v_{is}) &= cov(\alpha_i + u_{it}, \alpha_i + u_{is}) \\ &= E[(\alpha_i + u_{it})(\alpha_i + u_{is})] - E[\alpha_i + u_{it}]E[\alpha_i + u_{is}] \\ &= E[\alpha_i^2 + \alpha_i u_{it} + \alpha_i u_{is} + u_{it} u_{is}] - (0+0)(0+0) \\ &= E[\alpha_i^2] + E[\alpha_i u_{it}] + E[\alpha_i u_{is}] + E[u_{it} u_{is}] \\ &= var(\alpha_i) \end{aligned}$$

If it is the case that $var(\alpha_i) = 0$, then α_i is a degenerate variable centered at 0 - effectively there is no individual fixed effect. Except for this case, the composite error terms would be serially correlated.

The estimation method that is the most efficient for this case would be the GLS. The key to constructing a GLS is to find a variance-covariance matrix for $\mathbf{v}_i$. What we can do is to stack the T observations for each individual and work with

$$\mathbf{y}_i = \mathbf{X}_i\beta_1 + \mathbf{v}_i$$

where $\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ ... \\ y_{iT} \end{pmatrix}$, $\mathbf{X}_i = \begin{pmatrix} x'_{i1} \\ ... \\ x'_{iT} \end{pmatrix}$, $\mathbf{v}_i = \begin{pmatrix} v_{i1} \\ ... \\ v_{iT} \end{pmatrix} = \begin{pmatrix} \alpha_i + u_{i1} \\ ... \\ \alpha_i + u_{iT} \end{pmatrix}$. To make algebra more straightforward, I will define 1_T as a T-dimensional column vector of 1's. Then,

$$\mathbf{v}_i = 1_T\alpha_i + \mathbf{u}_i$$

Then the variance can be calculated by the $E[(X - E(X))(X - E(X))']$ formula. Thus,

$$\begin{aligned} var(v_i) &= E[(\mathbf{v}_i - E(\mathbf{v}_i))(\mathbf{v}_i - E(\mathbf{v}_i))'] \\ &= E[\mathbf{v}_i\mathbf{v}_i'] \ (\because E(\mathbf{v}_i) = 0) \\ &= E\left[1_T\alpha_i^2 1_T' + \mathbf{u}_i\alpha_i 1_T' + 1_T\alpha_i\mathbf{u}_i + \mathbf{u}_i\mathbf{u}_i'\right] \\ &= 1_T 1_T' var(\alpha_i) + E[\mathbf{u}_i\mathbf{u}_i'] \end{aligned}$$

Suppose that u_{it} is homoskedastic, then $E[\mathbf{u}_i\mathbf{u}_i']$ becomes $\sigma_u^2 I_T$, the result derived in class.

The GLS estimator can be derived by the following: Note that V is a symmetric matrix and positive semidefinite. Then $V = Q\Omega Q'$ by spectral decomposition theorem[4] and $V^{-1/2}$ can be defined (It would be $Q\Omega^{-1/2}Q'$). The trick now is to pre-multiply $V^{-1/2}$ to get

$$V^{-1/2}\mathbf{y}_i = V^{-1/2}\mathbf{X}_i\beta_1 + V^{-1/2}\mathbf{v}_i$$

and the resulting GLS estimator would be

$$\left(\sum_{i=1}^{n}(V^{-1/2}\mathbf{X}_i)'(V^{-1/2}\mathbf{X}_i)\right)^{-1}\sum_{i=1}^{n}(V^{-1/2}\mathbf{X}_i)'(V^{-1/2}\mathbf{y}_i) = \left(\sum_{i=1}^{n}\mathbf{X}_i'V^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{X}_i'V^{-1}\mathbf{y}_i$$

Comment 4.8.1 (So when do we use fixed effects vs. random effects?). *One way of figuring out when to use random effect is through a Hausmann Test. Random effects can be considered*

[4] It is a theorem that states that if a matrix is symmetric and positive semidefinite, there exists real eigenvalues and real orthogonal eigenvectors. You can look at Gockenbach(2010) for more details.

as a special case of fixed effects. In this test, the null hypothesis would be random effects and the alternative would be fixed effects. Note that the ingredients we have here are similar to the Hausmann Test in the endogeneity test. Random effects estimators would only be consistent if H_0 is true whereas fixed effects are consistent in both H_0 and H_1. If you are planning to use the random effects estimator (and believe that the random effects assumptions are appropriate in your context), the Hausmann test can be used as a specification check for this assumption and provide evidence to support your approach.

4.9 Dynamic Panel Data: Endogeneity Problems

Now consider a setting where

$$y_{it} = \rho y_{it-1} + \alpha_i + u_{it}$$

where $t = 0, 1, ..., T$ and $i = 1, .., n$. Additionally, assume that $E[\alpha_i] = 0, E[u_{it}] = 0, E[\alpha_i u_{it}] = 0 \; \forall i, t$ and $E[u_{it}u_{is}] = 0$ for $t \neq s$. Also, the initial observation y_{i0} satisfies $E[y_{i0}u_{it}] = 0 \; (t \geq 1)$ and that $E|u_{it}|^{2+\delta} \leq M < \infty (\delta > 0)$. Also note that observations are independent across i.

The difference between this and the previous panel estimation is that the lagged dependent variable enters as the regressor. In this case, all of the previous methods - random effects and fixed effects - are inconsistent. I will show this for POLS. Let $v_{it} = \alpha_i + u_{it}$. The OLS estimates would be

$$\hat{\rho} = \left(\sum_{i=1}^{n} \sum_{t=0}^{T} y_{it-1}^2 \right)^{-1} \sum_{i=1}^{n} \sum_{t=0}^{T} y_{it-1} y_{it}$$

or

$$\hat{\rho} - \rho = \left(\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=0}^{T} y_{it-1}^2 \right)^{-1} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=0}^{T} y_{it-1} (\alpha_i + u_{it})$$

The problem is that since $y_{it-1} = \rho y_{it-2} + \alpha_i + u_{it-1}$, the $\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=0}^{T} y_{it-1} \alpha_i$ term does not converge in probability to 0.

Even if we get rid of α_i by using the first difference at $T = 2$, ρ estimates are still inconsistent. To show this, start with

$$\Delta y_{i2} = \rho \Delta y_{i1} + \Delta u_{i2}$$

We can now show that the regressor and the error term are correlated, since

$$\begin{aligned} cov(\Delta y_{i1}, \Delta u_{i2}) &= E[(y_{i1} - y_{i0})(u_{i2} - u_{i1})] \\ &= E[y_{i1}u_{i2}] - E[y_{i1}u_{i1}] - E[y_{i0}u_{i2}] + E[y_{i0}u_{i1}] \end{aligned}$$

Because $E[y_{i1}u_{i1}]$ contains terms from the 1st period, this contains term that is nonzero.

Even for the within estimator, which is written as

$$y_{it} - \frac{1}{T}\sum_{i=1}^{T} y_{it} = \rho\left(y_{it-1} - \frac{1}{T}\sum_{i=1}^{T} y_{it-1}\right) + u_{it} - \frac{1}{T}\sum_{i=1}^{T} u_{it}$$

The regressor contains $y_{i0}, .., y_{iT-1}$ and residuals contain $u_{i1}, .., u_{iT}$. There are overlapping time periods, implying that the regressor becomes endogenous.

4.10 Anderson-Hsiao Estimator

In estimating

$$y_{it} = \rho y_{it-1} + \alpha_i + u_{it}$$

Anderson and Hsiao (1982) proposed the following IV approach. First difference the above equation and obtain

$$\Delta y_{it} = \rho \Delta y_{it-1} + \Delta u_{it}$$

where possible values of i remain the same but $t = 2, ..., T$. The suggested IV is that Δy_{it-1} be instrumented with y_{it-2}. We can check that this is a valid IV.

- **Relevancy**: Note that $\Delta y_{it-1} = y_{it-1} - y_{it-2}$. So this term contains y_{it-2} and thus relevancy is satisfied.

- **Exogeneity**: Note that

$$\begin{aligned} cov(y_{it-2}, \Delta u_{it}) &= E[y_{it-2}, u_{it} - u_{it-1}] \\ &= E[y_{it-2}u_{it}] - E[y_{it-2}u_{it-1}] \\ &= 0 - 0 = 0 \end{aligned}$$

The last line is justified as follows. Since we are assuming $E[y_{i0}u_{it}] = 0$ for $t \geq 1$, we can expand to

$$\begin{aligned} E[y_{i1}u_{it}]\ (t \geq 2) &= E[(\rho y_{i0} + \alpha_i + u_{i1})u_{it}] \\ &= \rho E[y_{i0}u_{it}] + E[\alpha_i u_{it}] + E[u_{i1}u_{it}] \\ &= \rho \times 0 + 0 + 0 = 0 \end{aligned}$$

Likewise,

$$\begin{aligned} E[y_{i2}u_{it}]\ (t \geq 3) &= E[(\rho y_{i1} + \alpha_i + u_{i2})u_{it}] \\ &= \rho E[y_{i1}u_{it}] + E[\alpha_i u_{it}] + E[u_{i2}u_{it}] = 0 \end{aligned}$$

Therefore, we can generalize to $E[y_{is}u_{it}] = 0$ for $t \geq s+1$.

We can generalize using matrix notation. Define

$$\Delta \mathbf{y}_i = \rho \Delta \mathbf{y}_{i,-1} + \Delta \mathbf{u}_i$$

where $\Delta \mathbf{y}_i = \begin{pmatrix} \Delta y_{i2} \\ ... \\ \Delta y_{iT} \end{pmatrix}$, $\Delta \mathbf{y}_{i,-1} = \begin{pmatrix} \Delta y_{i1} \\ ... \\ \Delta y_{iT-1} \end{pmatrix}$ and $\Delta \mathbf{u}_i = \begin{pmatrix} \Delta u_{i2} \\ ... \\ \Delta u_{iT} \end{pmatrix}$. The matrix of instruments Z_i would be

$$Z_i = \begin{pmatrix} y_{i0} & 0 & 0 & ... \\ 0 & y_{i1} & 0 & ... \\ ... & ... & ... & ... \\ 0 & 0 & ... & y_{iT-2} \end{pmatrix} \in \mathbb{R}^{(T-1)\times(T-1)}$$

This is a just identified case in the sense that the number of IV is equal to the number of endogenous variables. So we solve the sample analogue of

$$\begin{aligned} E[Z_i'\Delta \mathbf{u}_i] = 0 &\iff E[Z_i'(\Delta \mathbf{y}_i - \rho \Delta \mathbf{y}_{i,-1})] = 0 \\ &\implies \frac{1}{n}\sum_{i=1}^{n}(Z_i'\Delta \mathbf{y}_i - \rho Z_i'\Delta \mathbf{y}_{i,-1}) = 0 \\ &\iff \rho \frac{1}{n}\sum_{i=1}^{n} Z_i'\Delta \mathbf{y}_{i,-1} = \frac{1}{n}\sum_{i=1}^{n} \rho Z_i'\Delta \mathbf{y}_i \\ &\iff \hat{\rho} = \left(\sum_{i=1}^{n} Z_i'\Delta \mathbf{y}_{i,-1}\right)^{-1} \sum_{i=1}^{n} Z_i'\Delta \mathbf{y}_i \end{aligned}$$

4.11 Arellano-Bond Estimator: Without Regressors

This approach is similar to that of Anderson-Hsiao in the sense that we start with the first differentiation. The difference is in the instruments used. Arellano and Bond (1991) suggests that to instrument for Δy_{it-1}, we use $y_{i0}, ..., y_{it-2}$ as instruments. To see why they are valid

- **Relevancy:** It should be clear why y_{it-2} is relevant. As for others, since $y_{it-1} = \rho y_{it-2} + u_{it-1}$ and $y_{it-2} = \rho y_{it-3} + u_{it-2}$ We can write recursively that

$$y_{it-1} = \rho^2 y_{it-3} + \rho u_{it-2} + u_{it-1}$$

... and so on. Therefore, we can verify relevancy.

- **Exogeneity:** Note that $cov(y_{is}, \Delta u_{it})$ for any $s < t$ is 0, as we have shown above. So exogeneity holds as well.

The generalized approach using matrix will be similar except for the instrument matrix Z_i.

$$Z_i = \begin{pmatrix} y_{i0} & 0 & 0 & ... \\ 0 & (y_{i0}, y_{i1}) & 0 & ... \\ ... & ... & ... & ... \\ 0 & 0 & ... & (y_{i0}, ..., y_{iT-2}) \end{pmatrix} \in \mathbb{R}^{(T-1)\times \frac{T(T-1)}{2}}$$

Unlike Anderson-Hsiao estimator, this is an overidentified case. So we would need to use a GMM criterion with a weight matrix W_n. This would result in

$$\hat{\rho} = \arg\min_{\rho} \left\{ n \times \frac{1}{n}\sum_{i=1}^{n}(Z_i'\Delta\mathbf{y}_i - \rho Z_i'\Delta\mathbf{y}_{i,-1})' W_n \frac{1}{n}\sum_{i=1}^{n}(Z_i'\Delta\mathbf{y}_i - \rho Z_i'\Delta\mathbf{y}_{i,-1}) \right\}$$
$$\implies \left[\left(\sum_{i=1}^{n} \Delta\mathbf{y}_{i,-1}' Z_i\right) W_n \left(\sum_{i=1}^{n} Z_i'\Delta\mathbf{y}_{i,-1}\right)\right]^{-1} \left(\sum_{i=1}^{n} \Delta\mathbf{y}_{i,-1}' Z_i\right) W_n \left(\sum_{i=1}^{n} Z_i'\Delta\mathbf{y}_i\right)$$

The remaining question is to now select a matrix W_n that would lead to the lowest variance possible. If $g(Z_i, \rho)$ is the moment condition, the following would qualify as the most efficient weighting matrix.

$$W_n = E[g(Z_i, \rho)g(Z_i, \rho)']^{-1}$$

In our context, $g(Z_i, \rho)$ would be equivalent to $Z_i'\Delta u_i$. So we would need to find a sample analogue of

$$E[Z_i'\Delta\mathbf{u}_i\Delta\mathbf{u}_i'Z_i] \implies \frac{1}{n}\sum_{i=1}^{n} Z_i\Delta\mathbf{u}_i\Delta\mathbf{u}_i'Z_i$$

For further simplification, we can assume two types of settings

- u_{it} is **IID and homoskedastic**: In such case, where $E[u_{it}^2] = \sigma_u^2$, we can write

$$E[\Delta \mathbf{u}_i \Delta \mathbf{u}_i] = E \begin{bmatrix} \Delta u_{i2}^2 & \Delta u_{i2}\Delta u_{i3} & ... & \Delta u_{i2}\Delta u_{iT} \\ \Delta u_{i2}\Delta u_{i3} & \Delta u_{i3}^2 & ... & \Delta u_{ie}\Delta u_{iT} \\ ... & ... & ... & ... \\ \Delta u_{i2}\Delta u_{iT} & \Delta u_{i3}\Delta u_{iT} & ... & \Delta u_{iT}^2 \end{bmatrix} = \begin{bmatrix} 2\sigma_u^2 & -\sigma_u^2 & ... & 0 \\ -\sigma_u^2 & 2\sigma_u^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 2\sigma_u^2 \end{bmatrix}$$

  Define matrix $H \in \mathbb{R}^{(T-1)\times(T-1)}$ to have 2's in the diagonal elements, -1's in the immediate off-diagonals, and 0 everywhere else. Then $E[\Delta \mathbf{u}_i \Delta \mathbf{u}_i'] = \sigma_u^2 H$. This implies that the weighting matrix that we are looking for is

$$E[Z_i' \Delta \mathbf{u}_i \Delta \mathbf{u}_i' Z_i]^{-1} = \left(\frac{1}{n} \sum_{i=1}^{n} Z_i H Z_i \right)^{-1}$$

  where σ_u^2 is taken out since scaling W_n by a scalar does not affect the value of the estimator. However, note that it may affect the test statistics for the overidentification test we will conduct later.

- u_{it} is **heteroskedastic**: For this, we take an approach similar to the two-step GMM estimator we did some weeks ago. The optimal weighting matrix in this case would be

$$W_n = \left(\frac{1}{n} \sum_{i=1}^{n} Z_i \Delta \tilde{\mathbf{u}}_i \Delta \tilde{\mathbf{u}}_i' Z_i \right)^{-1}$$

  where $\Delta \tilde{\mathbf{u}}_i = \Delta \mathbf{y}_i - \tilde{\rho} \Delta \mathbf{y}_{i,-1}$, a residual from the preliminary estimator $\tilde{\rho}$. The preliminary estimator could be either from $W_n = I_{T(T-1)/2}$ or from the one-step GMM estimator that we derived earlier.

Because we are using more moment conditions that the number of endogenous variables, this is when we could test for an overidentification restriction. Suppose that W_n is the efficient weighting matrix. Then, similar to the GMM overidentification test, we are testing

$$H_0 : E[g(Z_i, \rho)] = 0, \quad H_1 : E[g(Z_i, \rho)] \neq 0$$

We can construct the following test statistic (Sargan-Hansen overidentification statistic)

$$J = n \bar{g}_n(\hat{\rho})' W_n \bar{g}_n(\hat{\rho})$$

Under H_0, J has a limiting distribution $\chi^2_{\left(\frac{T(T-1)}{2}-1\right)}$. We lose one degree of freedom since we have used one estimator for ρ.

4.12 Arellano-Bond Estimator: With Regressors

Now we generalize further by including regressors $x_{it} \in \mathbb{R}^k$. We can write the data generating process as

$$y_{it} = \rho y_{it-1} + \beta' x_{it} + \alpha_i + u_{it}$$

There are two types of assumption we can put on the regressors

Assumption 4.12.1 (Predetermined)**.** *We say that the regressors are* ***predetermined*** *if*

$$E[u_{it}|x_{i0},...,x_{it}] = 0 \text{ for each } t = 0,...,T$$

Assumption 4.12.2 (Strictly Exogenous)**.** *We say that the regressors are* ***strictly exogenous*** *if*

$$E[u_{it}|x_{i0},...,x_{iT}] = 0 \text{ for all } t = 0,...,T$$

Since there is an unobservable term α_i which we allow to be correlated with x_{it} and y_{it-1}, one thing we can do is to take the first difference

$$\Delta y_{it} = \rho \Delta y_{it-1} + \beta' \Delta x_{it} + \Delta u_{it}$$

However, note that the equation is endogenous because of the two reasons

- Δy_{it-1} is endogenous as

$$\begin{aligned} E[\Delta y_{it-1} \Delta u_{it}] &= E[(y_{it-1} y_{it-2})(u_{it} - u_{it-1})] \\ &= E[y_{it-1}u_{it}] - E[y_{it-1}u_{it-1}] - E[y_{it-2}u_{it}] + E[y_{it-2}u_{it-1}] \end{aligned}$$

  Since we still assume $E[y_{i0}u_{it}] = 0$ for $t \geq 1$, (and thus $E[y_{it}u_{is}] = 0$ for $s \geq t+1$), we can still get rid of other terms except $E[y_{it-1}u_{it-1}]$. The remaining term has a nonzero correlation, so Δy_{it-1} becomes endogenous

- Δx_{it} is endogenous under predetermined case since

  $$\begin{aligned} E[\Delta x_{it} \Delta u_{it}] &= E[(x_{it} - x_{it-1})(u_{it} - u_{it-1})] \\ &= E[x_{it} u_{it}] - E[x_{it} u_{it-1}] - E[x_{it-1} u_{it}] + E[x_{it-1} u_{it-1}] \end{aligned}$$

  While other terms are 0 by predetermined assumption, this assumption is silent about $E[x_{it} u_{it-1}]$. This is where the nonzero correlation is from.

Given the two possible assumptions on the regressors, there are two sets of IV/GMM estimation that we can use.

- **Predetermined Regressors**: Rewrite the differenced equation and stack in the vector form for each individual by writing

  $$\Delta \mathbf{y}_i = \Delta \mathbf{w}_i \delta + \Delta \mathbf{u}_i$$

  where $\Delta \mathbf{w}_i = [\Delta \mathbf{y}_{i,-1}, \ \Delta \mathbf{x}_i] \in \mathbb{R}^{(T-1) \times (k+1)}$, $\Delta \mathbf{x}_i = \begin{pmatrix} x_{i2} \\ ... \\ x_{iT} \end{pmatrix}$, $\delta = \begin{pmatrix} \rho \\ \beta' \end{pmatrix}$ and the other two variables are the same as before. For notational convenience, I define

  $$\mathbf{y}_{it} = \begin{pmatrix} y_{i0} \\ ... \\ y_{it} \end{pmatrix}, \mathbf{x}_{it} = \begin{pmatrix} x'_{i0} \\ ... \\ x'_{it} \end{pmatrix}$$

  The possible instruments for the endogenous variable Δx_{it} is any regressor up to period $t-1$, or $\mathbf{x}'_{it-1}$. The exogeneity is satisfied since the error term is $u_{it} - u_{it-1}$, regressors $x_{i1}, ..., x_{it-1}$ is not correlated to this by predeterminedness assumption. Since $\mathbf{x}'_{it-1}$ includes x_{it-1}, the IV's are relevant as well. Thus, we can write

  $$Z_i = \begin{pmatrix} y_{i0}, \mathbf{x}'_{i1} & 0 & 0 & ... & 0 \\ 0 & \mathbf{y}'_{i1}, \mathbf{x}'_{i2} & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \mathbf{y}'_{iT-2}, \mathbf{x}'_{iT-1} \end{pmatrix}$$

  Then we take a GMM approach by using the moment condition $E[Z_i' \Delta \mathbf{u}_i] = 0$, which results in the following estimator

$$\hat{\delta} = \left[\left(\sum_{i=1}^{n} \Delta \mathbf{w}_i' Z_i \right) W_n \left(\sum_{i=1}^{n} Z_i' \Delta \mathbf{w}_i \right) \right]^{-1} \left(\sum_{i=1}^{n} \Delta \mathbf{w}_i' Z_i \right) W_n \left(\sum_{i=1}^{n} Z_i' \Delta \mathbf{y}_i \right)$$

- **Strictly exogenous**: The difference between predetermined and strictly exogenous case is that the entire elements in $\mathbf{x}_{iT}'$ is a valid IV. Strict exogeneity prevents any correlation between error terms of all time periods and there are elements in $\mathbf{x}_{iT}'$ that are included in Δy_{it} for each t. So we can write the Z_i matrix as

$$Z_i = \begin{pmatrix} y_{i0}, \mathbf{x}_{iT}' & 0 & 0 & ... & 0 \\ 0 & \mathbf{y}_{i1}', \mathbf{x}_{iT}' & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \mathbf{y}_{iT-2}', \mathbf{x}_{iT}' \end{pmatrix}$$

We still work with the same moment condition $E[Z_i' \Delta \mathbf{u}_i] = 0$. The resulting estimator for δ shares the same expression as in the predetermined case, but with different matrix for Z_i.

4.13 Regressing with an 'Overall' Intercept

Consider the following model

$$y_{it} = \mu + \rho y_{it-1} + x_{it}' \beta + \alpha_i + u_{it}$$

where μ term represents an 'overall' constant - this term applies commonly to all i's and t's in the regression. When we estimated ρ and β before, we relied on first-differencing. The cost of doing this is the loss of time-invariant terms. This would now include μ. If we are desperate to know what μ is, how do we estimate for them?

There are two ways to do this. One is a two step approach in the sense that we estimate the other parameters first and the back out μ. The other is doing it simultaneously using additional moment conditions.

Before proceeding, it should be noted that $E(\alpha_i) = 0$ is implicitly assumed. If not, we can change the above equation in the following fashion:

$$y_{it} = \underbrace{\mu + E(\alpha_i)}_{\mu_0} + \rho y_{it-1} + x_{it}'\beta + \underbrace{\alpha_i - E(\alpha_i)}_{=\alpha_{0i}} + u_{it}$$

In this setup, $E(\alpha_{0i}) = 0$. and we estimate for μ_0. After that, we can find μ.

- **Two-step Approach**: We first obtain the estimate $\hat{\rho}$ and $\hat{\beta}$ using an Arellano-Bond estimator. Then, we can rearrange the equation into

  $$y_{it} - \hat{\rho} y_{it-1} - x_{it}'\hat{\beta} = \mu + \alpha_i + u_{it}$$

  Given that the mean of α_i and u_{it} is zero, we can use the sample mean of the left-hand-side to obtain the estimate for μ. Since μ applies to all i and t, we need to sum over all i's and t's. Therefore,

  $$\hat{\mu} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(y_{it} - \hat{\rho} y_{it-1} - x_{it}'\hat{\beta} \right)$$

- **Simultaneous Approach**: Note that $\alpha_i + u_{it}$ has a zero mean. For a fixed i, we can define $\mathbf{y}_i, \mathbf{y}_{i,-1}$ and $\mathbf{x}_i$ as stacked-up vector for individual i with T observations. We obtain

  $$\begin{aligned} \mathbf{y}_i =& 1_T\mu + \rho \mathbf{y}_{i,-1} + \mathbf{x}_i\beta + \underbrace{1_T\alpha_i + \mathbf{u}_i}_{=\mathbf{v}_i} \\ =& 1_T\mu + \mathbf{w}_i\delta + \mathbf{v}_i \end{aligned}$$

  using the fact that $\alpha_i + u_{it}$ has a zero mean yields this moment condition

  $$E[\mathbf{v}_i] = E[1_T\alpha_i + \mathbf{u}_i] = E[\mathbf{y}_i - \mathbf{w}_i\delta - 1_T\mu] = 0$$

  Along with the $E[Z_i'\Delta\mathbf{u}_i] = 0$ condition, we can jointly write

  $$\begin{bmatrix} E(\mathbf{v}_i) \\ E(Z_i'\Delta\mathbf{u}_i) \end{bmatrix} = E\left[\begin{bmatrix} I_T & 0 \\ 0 & Z_i \end{bmatrix}' \begin{bmatrix} \mathbf{v}_i \\ \Delta\mathbf{u}_i \end{bmatrix} \right]$$

Define $\bar{Z}_i = \begin{bmatrix} I_T & 0 \\ 0 & Z_i \end{bmatrix}$, $\bar{W}_i = \begin{bmatrix} I_T & \mathbf{w}_i \\ 0 & \Delta\mathbf{w}_i \end{bmatrix}$, $\bar{Y}_i = \begin{bmatrix} \mathbf{y}_i \\ \Delta\mathbf{y}_i \end{bmatrix}$. Then $\begin{bmatrix} \mathbf{v}_i \\ \Delta\mathbf{u}_i \end{bmatrix} = \bar{Y}_i - \bar{W} \begin{bmatrix} \mu \\ \delta \end{bmatrix}$. The moment condition and the resulting estimator for $\begin{bmatrix}\mu\\ \delta\end{bmatrix}$ is ($\bar{V}_n$ is the weighting matrix)

$$E\left[\bar{Z}_i'(\bar{Y}_i - \bar{W}_i)\begin{bmatrix}\mu\\ \delta\end{bmatrix}\right] = 0$$

$$\begin{bmatrix}\hat{\mu}\\ \hat{\delta}\end{bmatrix} = \left[\sum_i(\bar{W}_i'\bar{Z}_i)\bar{V}_n\sum_i\bar{Z}_i'\bar{W}_i\right]^{-1}\sum_i(\bar{W}_i'\bar{Z}_i)\bar{V}_n\sum_i\bar{Z}_i'\bar{Y}_i$$

Note that in this manner, we can technically estimate for α_i. However, since T is small in usual datasets, we may not be able to guarantee the consistency of such estimator.

4.14 Subset of Regressors Uncorrelated with Fixed Effects

Return to the model where

$$y_{it} = \mu + \rho y_{it-1} + x_{it}'\beta + \alpha_i + u_{it}$$

The difference is now that a subset of x_{it} is exogenous with respect to α_i. Also assume that regressors are strictly exogenous. Denote such variable as $x_{it}^{(1)}$. Then the following holds

$$E[x_{is}^{(1)}\alpha_i] = 0 \implies E[x_{is}^{(1)}v_{it}] = 0 \implies E[x_{is}^{(1)}(y_{it} - w_{it}'\delta - \mu)] = 0\ \forall s,t$$

For notational convenience, define $h_i = vec(x_i^{(1)}) = \begin{bmatrix}x_{i1}^{(1)}\\ ...\\ x_{iT}^{(1)}\end{bmatrix}$. Then the above moment condition is equivalent to

$$E[(I_T \otimes h_i)\mathbf{v}_i] = 0$$

We can also make use of the fact that $E[\mathbf{v}_i] = 0$, which is $E[(I_T \otimes 1)\mathbf{v}_i]$. This would give us the required moment condition for equations in levels (to back out μ), expressed as

$$E\left[\left(I_T \otimes \begin{bmatrix}1\\ h_i\end{bmatrix}\right)\mathbf{v}_i\right] = 0$$

Another set of required moment condition comes from Z_i matrix defined for strictly exogenous case with $\mathbf{x}_{iT}'$ replaced with $vec(x_i^{(2)})'$. This gives us $E[Z_i'\Delta\mathbf{u}_i] = 0$. The GMM estimator has the same expression as in the previous case but with I_T in $\bar{Z}_i$ replaced with $\left(I_T \otimes \begin{bmatrix} 1 \\ h_i \end{bmatrix}\right)$.

4.15 Nonlinear Moments

Consider the model

$$y_{it} = \rho y_{it-1} + \underbrace{\alpha_i + u_{it}}_{=v_{it}}$$

The assumptions on our variables were

- $E(\alpha_i) = 0, E(u_{it}) = 0, E(\alpha_i u_{it}) = 0 \ \forall i, t$
- $E(u_{it}u_{is}) = 0 \ t \neq s$
- $E(y_{i0}u_{it}) = 0 \ \forall i$ and $t = 1, .., T$

These three assumptions imply $E[\Delta v_{it-1}v_{it}] = 0 \ t = 3, .., T$. The rationale is as follows. Note that $\Delta v_{it-1} = u_{it-1} - u_{it-2} = \Delta u_{it-1}$. Also $v_{it} = \alpha_i + u_{it}$. Since fixed effects is not correlated with the error terms and there is no serial correlation, $E[\Delta v_{it-1}v_{it}] = 0$ holds. In other words,

$$E[(\Delta y_{it-1} - \rho \Delta y_{it-2})(y_{it} - \rho y_{it-1})] = 0$$

The sample analogue of this would be

$$\frac{1}{n}\sum_{i=1}^{n}(\Delta y_{it-1}y_{it} - \rho \Delta y_{it-1}y_{it-1} - \rho \Delta y_{it-2}y_{it} + \rho^2 \Delta y_{it-2}y_{it-1})$$

Notice the ρ^2 term here. Because of this, the moment condition becomes nonlinear (quadratic, to be exact). If GMM estimation is used, the objective function involves ρ^4 and FOC would involve ρ^3. So it is tricky to work with.

In addition, assume that there is a time series homoskedasticity, or $E[u_{it}^2] = \sigma_u^2$. Then the following $T - 1$ additional moments

$$E[v_{it}^2] - E[v_{it-1}^2] = 0, \ t = 2, .., T$$
$$\iff E[(y_{it} - \rho y_{it-1})^2] - E[(y_{it-1} - \rho y_{it-2})^2] = 0$$

Similar to the previous case, the objective function involves ρ^4 and FOC has ρ^3. This is again, tricky to work with. There is one way to 'linearize' the moment conditions. Which is...

4.16 Mean Stationarity

If a distribution of a certain variable y_t does not change over time, we call this a stationary distribution. Mean stationarity refers to the situation where the mean of a variable is time-invariant. One case where this can hold is as follows. Suppose

$$y_{i0} = \frac{\alpha_i}{1-\rho} + e_{i0}$$

Then $E[y_{i0}|\alpha_i] = \frac{\alpha_i}{1-\rho}$. Under the data generating process

$$y_{it} = \rho y_{it-1} + \alpha_i + u_{it}$$

we can show that means stationarity holds [5]. Specifically

$$\begin{aligned} E[y_{i1}|\alpha_i] &= E[\rho y_{i0} + \alpha_i + u_{i1}|\alpha_i] \\ &= \rho E[y_{i0}|\alpha_i] + \alpha_i + E[u_{i1}|\alpha_i] \\ &= \frac{\rho\alpha_i}{1-\rho} + \alpha_i = \frac{\alpha_i}{1-\rho} \end{aligned}$$

We can reiterate and get the same result for $E[y_{it}|\alpha_i]$ for $t = 1, ..., T$. Therefore, mean stationarity holds.

So how do the nonlinear moments become linearized? Let's start with the first nonlinear moment $E[(\Delta y_{it-1} - \rho\Delta y_{it-2})(y_{it} - \rho y_{it-1})] = 0$. By mean stationarity, we get

$$E[y_{it} - y_{it-1}|\alpha_i] = E[\Delta y_{it}|\alpha_i] = 0 \implies E[\alpha_i \Delta y_{it}] = 0 \;\forall t$$

This result, along with $E[y_{i1}u_{is}] = 0$ for $s \geq 2$ implies that $E[\Delta y_{i1}u_{is}] = 0$ for $s \geq 2$. Thus, $E[\Delta y_{i1}v_{is}] = 0$ $(s \geq 2)$. So we have $E[\Delta y_{i1}v_{i2}] = 0$. For $s \geq 3$, we can use the nonlinear moment condition $E[\Delta v_{is-1}v_{is}] = 0$ to back out

$$\begin{aligned} E[\Delta v_{is-1}v_{is}] = 0 &\implies E[(\Delta y_{is-1} - \rho\Delta y_{is-2})v_{is}] = 0 \\ (s=3) &\implies E[(\Delta y_{i2} - \rho\Delta y_{i1})v_{i3}] = 0 \\ &\quad = E[\Delta y_{i2}v_{i3}] = 0 \; (\because E[\Delta y_{i1}v_{is}] = 0 \; (s \geq 2)]) \end{aligned}$$

Repeat the similar process to ultimately get $E[\Delta y_{it-1}v_{it}] = 0$

[5] This is what question 3 in 2019 Spring Midterm is about

So the lagged differences of the instruments qualify as instruments for equation in levels. This is a Blundell-Bond estimator, or what is known as a system GMM estimation. Combine the above moment condition with the usual $E[Z_i' \Delta \mathbf{u}_i] = 0$ to get a joint moment condition

$$E\left[Z_i^{+'} \begin{pmatrix} \Delta \mathbf{u}_i \\ \mathbf{v}_i \end{pmatrix} \right] = 0$$

with $Z_i^+ = \begin{bmatrix} Z_i & 0 & 0 & ... & 0 \\ 0 & \Delta y_{i1} & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \Delta y_{iT-1} \end{bmatrix}$. Here, $\mathbf{v}_i$ has $t = 2, ..., T$.

As for the homoskedastic case, we can use the mean stationarity condition to obtain a linear moment condition

$$\begin{aligned} E[v_{it}^2] - E[v_{it-1}^2] &= E[(y_{it} - \rho y_{it-1})v_{it} - (y_{it-1} - \rho y_{it-2})v_{it-1}] \\ &= E[y_{it}v_{it} - y_{it-1}v_{it-1}] = 0 \; (t = 2, ..., T) \end{aligned}$$

The remaining terms can be written as

$$\begin{aligned} E[-\rho y_{it-1}(\alpha_i + u_{it}) + \rho y_{it-2}(\alpha_i + u_{it-1})] &= E[-\rho \Delta y_{it-1}\alpha_i + \rho y_{it-2}u_{it-1} - \rho y_{it-1}u_{it}] \\ &= -0 + 0 - 0 = 0 \end{aligned}$$

The mean stationarity justifies the first zero. Therefore, we can use more instruments, namely

$$Z_{iH}^+ = \begin{bmatrix} Z_i & 0 & 0 & ... & 0 & 0 & 0 & 0 & .. & 0 \\ 0 & 0 & 0 & .. & 0 & y_{i1} & 0 & 0 & .. & 0 \\ 0 & \Delta y_{i1} & 0 & ... & 0 & -y_{i2} & y_{i2} & 0 & .. & 0 \\ ... & ... & ... & ... & ... & .. & ... & ... & ... & y_{T-1} \\ 0 & 0 & 0 & ... & \Delta y_{iT-1} & 0 & 0 & 0 & .. & -y_T \end{bmatrix}$$

Therefore, using the combined moment condition

$$E\left[Z_{iH}^{+'} \begin{pmatrix} \Delta \mathbf{u}_i \\ \mathbf{v}_i \end{pmatrix} \right] = 0$$

we can obtain a GMM estimator of ρ. Here, $\mathbf{v}_i$ starts from $t = 1$.

4.17 Factor Analysis and Interactive Fixed Effects

Consider the following model

$$y_{it} = \mu_i + x_{it}'\beta + \lambda_i' f_t + u_{it}$$

where λ_i is a vector of factor loadings and f_t is a vector of factors. Each can be written

$$\lambda_i = \begin{bmatrix} \lambda_{i1} \\ ... \\ \lambda_{ir} \end{bmatrix}, f_t = \begin{bmatrix} f_{1t} \\ ... \\ f_{rt} \end{bmatrix}$$

where r is usually a small number. We will assume that only y_{it} and x_{it} is observable.

In fact, we can see that this format is a generalized version of the additive fixed effect we have seen so far. For one thing, by writing

$$\lambda_i = \begin{bmatrix} \alpha_i \\ 1 \end{bmatrix}, f_t = \begin{bmatrix} 1 \\ \delta_t \end{bmatrix}$$

we can back out the two-way fixed effects model of the following form

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

Generally, we can capture unobserved individual traits that can vary with time. Or we can also capture entity-level responses to a common shock at certain time.

Estimating the above model with regressors can be done as follows. If it is the case that we know β, then we can write

$$y_{it} - x_{it}'\beta = \mu_i + \lambda_i' f_t + u_{it}$$

and estimate pure factor models. If we know what $\mu_i + \lambda_i' f_t$ is, we write

$$y_{it} - \mu_i - \lambda_i' f_t = x_{it}'\beta + u_{it}$$

which becomes a standard model. If we need to determine the parameters of interest all at once, we can do a LASSO-type regularized regression in the following sense.

$$\min \sum_i \sum_t (y_{it} - x_{it}'\beta - l_{it})^2 + \tau \|L\|_*$$

where l_{it} is the $\lambda_i' f_t$ and L is the matrix of l_{it}'s. Note that we are applying a nuclear norm here. In this context, what we are doing is to minimize the sum squared residuals but with the penalty that applies when the rank of L is large. Basically, we are treating β, λ_i, f_t as a parameter to be estimated, whereas for the static dynamic model, our interest was on β.

A more compact way to write this is to stack observations in the time series format. Define

$$\mathbf{y}_t = \begin{bmatrix} y_{1t} \\ ... \\ y_{nt} \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ ... \\ \mu_n \end{bmatrix}, \Lambda = \begin{bmatrix} \lambda_1' \\ ... \\ \lambda_n' \end{bmatrix}, \mathbf{u}_t = \begin{bmatrix} u_{1t} \\ ... \\ u_{nt} \end{bmatrix}$$

Then the pure factor model can be written as

$$\mathbf{y}_t = \mu + \Lambda f_t + \mathbf{u}_t$$

We impose these assumptions:

- $E(f_t) = 0, var(f_t) = \Sigma_f \in \mathbb{R}^{r \times r}$
- $E(\mathbf{u}_t) = 0, var(\mathbf{u}_t) = \Psi$, where Ψ is a diagonal matrix

This implies that $E[\mathbf{y}_t] = 0$ and $var[\mathbf{y}_t] = \Lambda \Sigma_f \Lambda' + \Psi$. Also, Ψ being diagonal implies that the error across entities are uncorrelated. The goal is to identify each parameters in the $var[\mathbf{y}_t]$ term. For this we need to put restrictions, r^2 of them. The reason is as follows.

Suppose that $\Lambda \Sigma_f \Lambda'$ is identifiable. So how do we determine individual parameters? This is tricky since for any matrix $A \in \mathbb{R}^{r \times r}$ s.t. $\det(A) \neq 0$,

$$\begin{aligned} \Lambda \Sigma_f \Lambda' &= \Lambda A A^{-1} \Sigma_f (A A^{-1})' \Lambda' \\ &= \underbrace{\Lambda A}_{\Lambda^*} \underbrace{A^{-1} \Sigma_f A^{-1'}}_{\Sigma_f^*} \underbrace{A' \Lambda'}_{\Lambda^{*'}} \\ &= \Lambda^* \Sigma_f^* \Lambda^{*'} \end{aligned}$$

We can get many observationally equivalent models. This is known as a rotational indeterminancy. Restrictions are needed to prevent this problem

Typically, three sets of assumptions are imposed

- **Classical:** We impose $\Sigma_f = I_r$. That will take care of $\frac{r(r+1)}{2}$ restrictions (Σ_f is symmetric). To take care of the rest, we impose $\Lambda' \Psi^{-1} \Lambda$ is diagonal - i.e off diagonals are symmetric (diagonal elements are still free). This takes care of $\frac{(r-1)r}{2}$ restrictions. Thus,

$$var(\mathbf{y}_t) = \Lambda\Lambda' + \Psi$$

- **Triangular:** Split Λ into $\Lambda = \begin{bmatrix} \Lambda_1 \in \mathbb{R}^{r \times r} \\ \Lambda_2 \in \mathbb{R}^{(n-r) \times r} \end{bmatrix}$. Then we impose $\Sigma_f = I_r$ and Λ_1 be a lower triangular matrix. So the upper diagonal matrix of Λ_1 is restricted to 0. The sum of the number of both sets of restriction is r^2. This implies that

$$\begin{aligned} y_{1t} &= \lambda_{11} f_{1t} + u_{1t} \\ y_{2t} &= \lambda_{21} f_{1t} + \lambda_{22} f_{2t} + u_{2t} \\ &.. \\ y_{rt} &= \lambda_{r1} f_{1t} + ... + \lambda_{rr} f_{rt} + u_{rt} \end{aligned}$$

and $var(\mathbf{y}_t) = \Lambda\Lambda' + \Psi$.

- **Measurement error:** This involves an unrestricted Σ_f and $\Lambda_1 = I_r$. Since Λ_1 is

$$\Lambda_1 = \begin{bmatrix} \lambda_{11} & \lambda_{12} & ... & \lambda_{1r} \\ \lambda_{21} & \lambda_{22} & ... & \lambda_{2r} \\ ... & ... & ... & ... \\ \lambda_{r1} & \lambda_{r2} & ... & \lambda_{rr} \end{bmatrix}$$

imposing $\Lambda_1 = I_r$ uses up all r^2 restrictions. Then we have

$$\begin{aligned} y_{1t} &= f_{1t} + u_{1t} \\ y_{2t} &= f_{2t} + u_{2t} \\ &.. \\ y_{rt} &= f_{rt} + u_{rt} \end{aligned}$$

and hence the name measurement error restriction. Then

$$var(\mathbf{y}_t) = \begin{bmatrix} \Sigma_f & \Sigma_f \Lambda_2' \\ \Lambda_2 \Sigma_f & \Lambda_2 \Sigma_f \Lambda_2' \end{bmatrix} + \Psi$$

These assumptions solve rotational indeterminancy. To fully identify all parameters, additional identification assumptions are required. For this, I recommend referring to Professor Bai's lecture notes on Factor Analysis, available upon request.

Chapter 5

Bootstrap Methods and Multiple Testing

5.1 Refresher on the Properties of the Estimators

Before we get to bootstrap, a reminder on the key properties of the estimators come in handy. Assume that we have data $(y_1, ..., y_n)$ distributed according to some probability measure P. The definition of a probability space is as follows.

Definition 5.1.1 (Probability Measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space, with Ω referring to sets of states of the world and $\mathcal{F}$ referring to $\sigma-$algebra of subsets of Ω (or collection of events). The probability measure P is a $[0, 1]$ valued function summarizing the likelihood of an event that satisfies

- $P(\Omega) = 1, P(\emptyset) = 0$
- $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ for a countable disjoint sequence of elements of $\mathcal{F}$.

For instance, a normal distribution with mean μ and variance σ^2 can count.

Normally, we do not have an exact knowledge of what P did the data $(y_1, ..., y_n)$ came from. We may know that the data is generated by a distribution F_θ for some $\theta \in \Theta$. However, there is a clear virtue in pinpointing a particular θ (denoted as θ_0). Knowing this allows us to discover the population from which our data is from. For instance, if $F_\theta = N(\theta, 1)$ and if we can pinpoint a particular θ_0, we know the mean and the variance of our data.

Therefore, if we do not know the exact θ_0 yet, we estimate what this might be. An **estimator** is a function of the observations $(y_1, ..., y_n)$, often denoted as $\hat{\theta}_n$. For instance, a mean, defined as $\frac{\sum_{i=1}^{n} y_i}{n}$, or the median is an estimator. While there are many possible estimators for a given data, not all of them are equal. Ideally, we would like the estimators to satisfy these properties

- **Unbiased**: $E(\hat{\theta}_n) = \theta_0$, where $E(\cdot)$ is the expectation under F_{θ_0}
- **Consistent**: $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$
- **Asymptotically Normal**: $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma_0)$ as $n \to \infty$. (Σ_0 is PD matrix)

In addition, we would like to carry out inference on our estimators reliably. To do this, we need to find a good way to approximate the finite-sample distribution of $\hat{\theta}_n$.

All of this is easy when we have a classical linear model, specified as

$$y = Xb_0 + u$$

where $u \sim N(0, \sigma_0^2)$. Assuming that the observations are IID conditional on $X = x$, the model has distribution $y \sim N(xb_0, \sigma_0^2)$. The OLS estimator has a finite-sample distribution that is characterized as follows

$$\begin{aligned} \hat{b}_n &= b_0 + (X'X)^{-1}X'u \\ \implies \sqrt{n}(\hat{b}_n - b_0) &= \left(\frac{X'X}{n}\right)^{-1} \frac{X'u}{\sqrt{n}} \\ &\sim N(0, \sigma_0^2(X'X)^{-1}) \end{aligned}$$

Since CLT implies that $\frac{X'u}{\sqrt{n}} \sim N(0, \sigma_0(X'X))$. If we are unsure of what σ_0^2 exactly is, we need to estimate for it using

$$\hat{\sigma}_n^2 = \frac{1}{n-p}\sum_{i=1}^{n} \hat{u}_i^2$$

where $p = \dim(X)$ and $\hat{u}_i = y_i - x_i\hat{b}_n$. This is an unbiased estimator for σ_0^2. We can see this from the following.

$$\begin{aligned} \hat{u} &= y - X\hat{b}_n \\ &= y - X(X'X)^{-1}X'y \\ &= Xb_0 + u - Xb_0 - X(X'X)^{-1}X'u = Mu \end{aligned}$$

where $M = I - X(X'X)^{-1}X$ and is idempotent and symmetric. Then,

$$\begin{aligned}
E[\hat{u}'\hat{u}] &= E[u'Mu] \\
&= E[tr(u'Mu)] \ (\because u'Mu \text{ is scalar.}) \\
&= E[tr(Muu')] \ (tr(AB) = tr(BA)) \\
&= tr[E(Muu')] \ (E(tr(A)) = E(A_{11} + ... + A_{nn}) = E(A_{11}) + ... + E(A_{nn}) = tr(E(A))) \\
&= tr[ME(uu')] \ (\text{If we condition } X = x, \ M \text{ is not stochastic}) \\
&= tr[M\sigma_0^2 I_n] = \sigma_0^2 tr(M) \\
&= \sigma_0^2 tr(I - X(X'X)^{-1}X') = \sigma_0^2[tr(I) - tr(X(X'X)^{-1}X')] \\
&= \sigma_0^2(n - tr((X'X)^{-1}X'X)) = \sigma_0^2(n-p)
\end{aligned}$$

Therefore, $E\left[\frac{1}{n-p}\hat{u}'\hat{u}\right] = E\left[\frac{1}{n-p}\sum_{i=1}^{n}\hat{u}_i^2\right] = \sigma_0^2$.

For a more honest inference that captures how reliable the point estimates are, we would need a confidence interval. We can estimate the coverage probability for b_0 of a region $B \in \mathbb{R}^p$ by

$$\Pr(N(\hat{b}_n, \hat{\sigma}_n^2(X'X)^{-1}) \in B)$$

In words, it estimates the proportion of times that the region B contains the distribution of b_0.

Suppose we are interested in finding the 95% confidence interval for component k in $\hat{b}_n$, denoted as $\hat{b}_{kn}$. Applying above setup, the relevant standard error is $\hat{\sigma}_{kn} = \hat{\sigma}_n\sqrt{(X'X)_k^{-1}}$ and the interval would have upper and lower bounds of $\hat{b}_{kn} \pm 1.96\hat{\sigma}_{kn}$. We can test the null hypothesis that $b_{k0} = c$ with the rejection criterion

$$\frac{|\hat{b}_{kn} - c|}{\hat{\sigma}_{kn}} > 1.96$$

which gives a test with size (reject true null hypothesis wrongly by) 5% and power (correctly reject wrong hypothesis) that converges to 1.

Comment 5.1.1 (What is a confidence interval?). *There are some cautions when interpreting the confidence interval. Let (l, u) be the 95% confidence interval after a sample is drawn. A 95% confidence interval can be interpreted as saying that* ***there is a 95% probability that the (l, u) confidence interval encompasses the true value****. Notice that this is a probability statement* *about the confidence interval and not the true value. True value of β_{k0} should not be treated as a random variable. Therefore, it is not necessarily true to say that there is a 95% probability that b_{0k} is included in a particular (l, u) confidence interval. Another way of seeing this is if the sample has been drawn N times, the fraction of calculated confidence intervals (that differs for each draw) which encompasses the true value is 95%, or that $N \times 0.95$ confidence intervals encompass true parameter.*

5.2 Beyond Classical Setup

In most cases, we only know the asymptotics. Suppose that $\hat{\theta}_n$ is a CAN (consistent and asymptotically normal) estimator s.t.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \sim N(0, \Sigma_0)$$

In a finite-sample context, we need to rely on an approximation of a following sense: Let $\widehat{\Sigma}_n$ be a consistent estimator for Σ_0. Then we can write the above equation as

$$\hat{\theta}_n = \theta_0 + \frac{1}{\sqrt{n}} \widehat{\Sigma}_n^{-1/2} N(0, I_n)$$

There are cases where above approximation can be very misleading. One example is from the White Misspecification Test.

To show this, we need to formulate Fisher's Information Matrix. Let L be a likelihood function, which is a function of y, x, and parameter θ. Define

$$I = E\left[\left(\frac{\partial \log L}{\partial \theta}\right) \cdot \left(\frac{\partial \log L}{\partial \theta'}\right)\right]$$

Which is the same as

$$J = -E\left[\frac{\partial^2 \log L}{\partial \theta \partial \theta'}\right]$$

By Cramer-Rao Lower Bound, we can estimate the variance of MLE as I^{-1} or J^{-1}. Alternatively, we can also use a Sandwich formula $J^{-1} I J^{-1}$ for more robust estimate. White (1982) says that comparing the results gives an asymptotic test for misspecification. It turns out that the asymptotic approximation is unreliable to the extent that the test rejects well-specified models an order of magnitude more often than it should.

Comment 5.2.1 (Why is $I = J$?). *To show that expressions in I and J are equal, I define $L(\theta; x) = f(x;\theta)$ as a likelihood function. Notice that*

$$1 = \int f(x;\theta)dx$$

Firstly, differentiate with respect to θ to get

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} f(x;\theta)dx \\ &= \int \frac{\partial}{\partial\theta} \log f(x;\theta) \times f(x;\theta)dx = E\left[\frac{\partial \log f(x;\theta)}{\partial\theta}\right] \end{aligned}$$

Differentiate above with respect to θ' to get

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta'}\left(\frac{\partial}{\partial\theta}\log f(x;\theta) \times f(x;\theta)\right) dx \\ &= \int \frac{\partial^2}{\partial\theta\partial\theta'} \log f(x;\theta)dx + \int \frac{\partial \log f(x;\theta)}{\partial\theta}\frac{\partial f(x;\theta)}{\partial\theta'}dx \\ &= \int \frac{\partial^2}{\partial\theta\partial\theta'} \log f(x;\theta)dx + \int \frac{\partial \log f(x;\theta)}{\partial\theta}\frac{\partial \log f(x;\theta)}{\partial\theta'}f(x;\theta)dx \\ &= E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(x;\theta)\right] + E\left[\frac{\partial}{\partial\theta}\log f(x;\theta) \cdot \frac{\partial}{\partial\theta'}\log f(x;\theta)\right] \end{aligned}$$

Take the first term on RHS of the last line to LHS, and we get $I = J$. (Hogg et al. (2014))

To overcome such problems, we can rely on resampling techniques. That is, we pretend that the sample is the entire population and that we use as much as we can by re-drawing within the samples. While this does not necessarily get us the better estimates, it may be possible to obtain a better confidence interval.

5.3 Bootstrap

In doing the bootstrap, we are looking for alternative ways to assign standard errors (and ultimately the test statistics and confidence intervals) that are *at least as good as* and ideally *better than* the typical asymptotic approximation that we normally do. **Bootstrap** estimators are usually as good as asymptotic approximations in the sense that bootstrap estimators are also *consistent* and better in the sense that it has *smaller approximation error*.

The procedure works as follows for the t-statistic :

1. Start with an IID sample $(X_1, ..., X_n)$. Estimate the model and compute the t-statistic
2. Make n draws *with replacement*. Re-estimate the model and compute the new t-statistic.
3. Repeat step 2 many times
4. Since there are different t-statistics for each sample, there will be a distribution of t-statistics. Use this t-distribution to get a reliable critical value

To generalize the discussion, I will use the following framework. We have a model and a data $(X_1, ..., X_n)$ generated from an unknown distribution F_0. We can characterize the test statistic as $T_n(X_1, ..., X_n)$. Define the finite sample distribution of the test statistic under F_0 as

$$G_n(t, F_0) = \Pr(T_n \leq t)$$

Two properties below are nice properties for T_n to have:

Definition 5.3.1 (Pivotal and Asymptotically Pivotal)**.**

- **Pivotal**: We say T_n is pivotal if G_n itself does not depend on the unknown parameters of F_0.
- **Asymptotically Pivotal**: We say T_n is asymptotically pivotal if

  $$G_\infty(t, F_0) = \lim_{n \to \infty} G_n(t, F_0)$$

  does not depend on F_0. (so that F_0 can be omitted in the arguments.)

Since we do not know F_0, we need an estimator for it - an empirical distribution function which puts a probability mass $1/n$ on each observation

Definition 5.3.2 (Empirical Distribution Function)**.** The **empirical distribution function**

(or EDF) for the data $X_1, .., X_n$ is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$$

Using the EDF, we can estimate $G_n(t, F_n)$ instead of $G(t, F_0)$ in this way

1. Select a large B, say 10,000. (This depends on your computer's processing capacity)
2. For each $b \in \{1, .., B\}$ we do as follows
   - For each $i = 1, .., n$ pick $j = 1, ..., n$ randomly with replacement and let $X_i^b = X_j$. In other words, draw n times with replacement from $(X_1, .., X_n)$ and redefine the 'shuffled' sample as $(X_1^b, ..., X_n^b)$.
   - Then, compute the new test statistic $T_n^b = T_n(X_1^b, ..., X_n^b)$.
3. Then, we will end up with B numbers of T_n^b. Then $G_n(t, F_n)$ can be written as

$$G_n(t, F_n) = \frac{1}{B}\sum_{b=1}^{B} 1(T_n^b \leq t)$$

Once we are done with this, we find the bootstrapped critical values. Suppose that our test is of size 5% and we are doing a two-sided test. We find lower(upper) bounds for our critical value c^- (c^+) s.t. $0.025B$ values of T_n^b are smaller(larger) than that critical value. We then reject the null hypothesis if the T_n from the original $(X_1, .., X_n)$ does not belong within the bounds set by this procedure. Bootstrapped confidence intervals are obtained by inverting the test - we take all values not rejected by the above procedure and construct an interval.

Now we show why bootstrapped estimators are *at least as good as* and sometimes ideally *better than* normal asymptotic approximations. Using the technic of Edgeworth Expansion, we can write

$$\begin{aligned}
G_n(t, F_0) &= G_\infty(t, F_0) + \frac{1}{\sqrt{n}}g_1(t, F_0) + \frac{1}{n}g_2(t, F_0) + \frac{1}{n\sqrt{n}}g_3(t, F_0) + O\left(\frac{1}{n^2}\right) \\
G_n(t, F_n) &= G_\infty(t, F_n) + \frac{1}{\sqrt{n}}g_1(t, F_n) + \frac{1}{n}g_2(t, F_n) + \frac{1}{n\sqrt{n}}g_3(t, F_n) + O\left(\frac{1}{n^2}\right)
\end{aligned}$$

uniformly over t almost surely. (From Horowitz, 2011). Suppose that $G_\infty(t, \cdot)$ is continuous at F_0. Then

$$\begin{aligned} G_n(t, F_n) - G_n(t, F_0) = G_\infty(t, F_n) - G_\infty(t, F_0) + \frac{1}{\sqrt{n}}(g_1(t, F_0) - g_1(t, F_n)) \\ + \frac{1}{n}(g_2(t, F_0) - g_2(t, F_n)) + \frac{1}{n\sqrt{n}}(g_3(t, F_0) - g_3(t, F_n)) + O\left(\frac{1}{n^2}\right) \end{aligned}$$

Notice that $G_\infty(t, F_n) - G_\infty(t, F_0)$ is $O\left(\frac{1}{\sqrt{n}}\right)$ since $F_n - F_0 = O\left(\frac{1}{\sqrt{n}}\right)$.[1] Therefore, the size of the bootstrap approximation error is $O\left(\frac{1}{\sqrt{n}}\right)$ and the bootstrap error is consistent. Note that normal asymptotic approximation also has the same size of approximation error.

If T_n is asymptotically pivotal, $G_\infty(t, F_n) - G_\infty(t, F_0)$ term vanishes. The leading term is $\frac{1}{\sqrt{n}}(g_1(t, F_0) - g_1(t, F_n))$ and the size of the approximation error is $O\left(\frac{1}{n}\right)$ - improvement.

5.4 Types of Bootstraps

- **Parametric Bootstrap** : Suppose that there is a parametric model for $y = g(x, \theta_0, u)$ and that you have faith in this form. Then 1) hold x_i's fixed and bootstrap on the residuals $\hat{u}_i$, which gets you the bootstrapped $y_i = x\hat{\theta}_0 + \hat{u}_i$. Another way to do it is to draw from the estimated parametric distribution of u. Note that if you want to know the distribution of T_n under H_0, then assume H_0 when resampling. Each step contributes to making bootstrap powerful.

- **Bias Corrected Bootstrap** : To get the idea, suppose that X is a random vector where $\mu = E(X)$. Assume that the we are interested in the true value of θ, which is $\theta_0 = g(\mu)$ (g is known and continuous). We can consistently estimate θ_0 by $\theta_n = g(\bar{X})$. However, when we take a Taylor expansion of $\theta_n = g(\bar{X})$ about $\bar{X} = \mu$, [2]

  $$\theta_n = g(\mu) + g'(\mu)(\bar{X} - \mu) + \frac{1}{2}g''(\mu)(\bar{X} - \mu)^2 + O(n^{-2})$$

  As such, the bias term is

  $$B_n = E[\theta_n] - \theta_0 = g'(\mu)E(\bar{X} - \mu) + \frac{1}{2}E(g''(\mu)(\bar{X} - \mu)^2) + O(n^{-2}) = \frac{1}{2}E(g''(\mu)(\bar{X} - \mu)^2)$$

  which is nonzero and $O(n^{-1})$ unless g is a linear function.

[1] Refer to Dvoretzky-Kiefer-Wolfowitz inequality

[2] For simplicity, I discuss the scalar case

In the bootstrap setup, let $\bar{X}^*$ denote the mean of the 'shuffled' sample. The analogue for the above is

$$\theta_n^* = \theta_n + g'(\bar{X})(\bar{X}^* - \bar{X}) + \frac{1}{2}g''(\bar{X})(\bar{X}^* - \bar{X})^2 + O(n^{-2})$$

then, take the empirical expectation to get

$$B_n^* = E^*[\theta_n^* - \theta_n] = \frac{1}{2}E^*(g''(\bar{X})(\bar{X}^* - \bar{X})^2) + O(n^{-2})$$

Here, we have $E[B_n^*] = B_n + O(n^{-2})$. Thus, the bias corrected estimator of θ_0 is now $\theta_n - B_n^* = 2\theta_n - E^*[\theta_n^*]$. The bias is

$$E[\theta_n - B_n^*] - \theta_0 = O(n^{-2})$$

- **Variance of an Estimator** : This is the most common use for the bootstrap. The bootstrap variance of the estimator $\hat{\theta}_n$ is defined as

$$\tilde{V}_n = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_n^b - \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_n^i\right)^2$$

  Once you have a computer, it is not too difficult to compute. However, the resulting standard error depends on the choice of B.

- **Block Bootstrap** : This is used when we suspect that the data is serially correlated. What we do is to use m-size blocks of consecutive observations. Then we can use $B = n + 1 - m$ blocks in total. This works when $m/n \to 0, m \to \infty$ as $n \to \infty$

- **Wild Boostrap** : This is useful when we have a heteroskedastic error terms. Suppose we have a sample b, and there is an i'th observation in that sample. The trick is to use $u_i^* = \epsilon_i \hat{u}_i$ where $E[\epsilon_i] = 0, E[\epsilon_i^2] = 1$. As in class, you can use $\epsilon_i = \pm 1$. Mammen (1993) shows that choosing

$$\epsilon_i = \begin{cases} \frac{1+\sqrt{5}}{2} \text{ Prob:} \frac{\sqrt{5}-1}{2\sqrt{5}} \\ \frac{1-\sqrt{5}}{2} \text{ Prob:} \frac{\sqrt{5}+1}{2\sqrt{5}} \end{cases}$$

  leads to a valid and very efficient bootstrap inference for linear models.

- **Subsampling**[3] : Consider the same setup as bootstrap - we are still interested in the distribution of $T_n(X_1, ..., X_n)$. Instead of resampling with the size n, we resample with size m and do a bootstrap for B times. So there are a total of $N_{n,m} = \begin{pmatrix} n \\ m \end{pmatrix}$ total number of subsets that can be found. I will denote one of those subsample as $W_k, k \in \{1, ..., N_{n,m}\}$. Then we are going to make use of $T_m(W_k)$. We are trying to estimate $G_n(t, F_n)$ with T_m, but on a rescaled version. Define

$$J_n(t, F_0) = \Pr(r_n(T_n - T_\infty) \leq t) \text{ (You can define } J_m(t, F_0) \text{ similarly)}$$

where r_n is a scaling parameter independent of F_0 and chosen such that $J_n(t, F_0)$ converges in distribution to a nondegenerate distribution $J_\infty(t, F_0)$ (say, $\sqrt{n}$). We estimate this using

$$L_{n,m} = \frac{1}{N_{n,m}} \sum_{i=1}^{N_{n,m}} 1[r_m(T_m - T_n) \leq t]$$

The $L_{n,m}$ is intended to be an estimator of $J_m(t, F_0)$. Then, as $m/n \to 0, r_m/r_n \to 0$ and $m \to \infty$ as $n \to \infty$, we have

$$r_n(T_n - T_\infty) \simeq r_m(T_m - T_n)$$

which works better than bootstrap as long as $J_\infty(t, F_0)$ is continuous with respect to t

5.5 Multiple Hypothesis Testing

Suppose that we are testing S hypothesis where S is some large number and each hypothesis is mutually independent. For instance, one might run multiple regressions on several investment strategies while comparing to some benchmark and wishes to find which strategies are outperforming the said benchmark. Let $S = 1000$. Then, the probability of rejecting any one of the 1000 hypothesis is calculated by $1 - (0.95)^{1000} \simeq 1$[4], in other words, you get something significant just out of pure luck. In doing so, the most dangerous consequence is running in to a **false discovery** - a rejection of a true null hypothesis[5].

The best approach that prevents this mistake is finding either an empirical or theoretical

[3] Usually carried without replacements. For the similar version that executes this with replacement, search *m* out of *n* bootstrap

[4] I did this on a Google calculator.

[5] Discovery refers to any rejection of a null hypothesis.

motivations for the hypothesis. In a context with larger datasets, we need to come up with a technical method for testing multiple hypothesis.

5.6 Setting up for Multiple Testing

We are running S simultaneous tests. For a particular test $s \leq S$, we are testing the null H_{0s} versus the alternative H_{1s}, we know that the p-value is

$$p_s = \Pr(\text{Reject } H_{0s} | H_{0s} \text{ is true})$$

This is also the mathematical definition of the false discovery. Some other key terms are

- **False discovery proportion** (FDP): It is the proportion of the false discoveries out of all discoveries, or
- **False discovery rate** (FDR): $E[FDP]$
- **Familywise error rate** (FWE): It is the probability of one or more false discoveries. A k-FWE calculates probability of k or more false discoveries.

There are two strands of methods. One controls for the FWE. The other controls for the FDR. We go over each strands

5.7 Controlling FWE: Bonferroni Method and Holm Method

Bonferroni method is a classical method based on Bonferroni inequalities. To see how this works, let A_s be the event that $p_s < w$ (and for which H_{0s} is rejected) for some value w associated with the significance level α, where $s = 1, ..., S$. Let I_0 be indices for the true hypothesis, with $n(I_s) = S_0 \leq S$. The goal is to determine the value of w that makes the $FWE \leq \alpha$. Then the following logic works

$$FWE = \Pr(\cup_{s \in I_0} A_s) \leq \sum_{s \in I_0} \Pr(A_s) = S_0 w \leq S w$$

So for the FWE, we can let $w = \alpha / S$.

For the general case of k-FWE, we take this approach:

Theorem 5.7.1 (From Lehmann and Romano (2005)[a]). *For testing H_{0s} where $s = 1, .., S$, if p_s is $U[0, 1]$, then following the procedure where H_{0s} is rejected if $p_s \leq \frac{k\alpha}{S}$ controls k-FWE*

Proof. Let I_0 be the collection of indices of the true null hypotheses and $n(I_0) = S_0$. Let N be the number of rejections of true null hypothesis. Then we can write

$$\begin{aligned}
\Pr(N \geq k) &\leq \frac{E(N)}{k} \ (\because \text{ Markov Inequality}) \\
&= \frac{E\left[\sum_{s\in I_0} 1\left(p_s \leq \frac{k\alpha}{S}\right)\right]}{k} \\
&= \sum_{i\in I_0} \frac{\Pr\left(p_s \leq \frac{k\alpha}{S}\right)}{k} \\
&\leq \sum_{i\in I_0} \frac{k\alpha/S}{k} = S_0\frac{\alpha}{S} \leq \alpha
\end{aligned}$$

□

[a] Lehmann, E. L. and J. P. Romano (2005) Generalizations of the Familywise Error Rate, *The Annals of Statistics*, 33(3), 1138-1154

While we do not need to make any assumptions about distribution of the p-values, the problem is that this method tends to be conservative - there are too few rejections.

Holm method is a step-down method in the sense that ordering of the different p-values are involved. The procedure is as follows

1. Order the p-values s.t. $p_1 < ... < p_S$, and denote each null hypothesis as $H_{01}, ..., H_{0S}$.

2. Reject H_{0j} if $p_j \leq \frac{\alpha}{S-j+1}$ for $j = 1, 2, ..., S$

That is, we start with the hypothesis with lowest p-value and compare its p-value against α/S. If this is rejected, move on to the next hypothesis and compare its p-value against $\alpha/(S-1)$. In the last hypothesis, we would be comparing against α.

The way this works is as follows: Let k be the index of the first true null hypothesis that is rejected. Since we ordered hypothesis according to the size of the p-value, the first $k-1$ rejected hypotheses are false null hypothesis. Since there are total of $S - S_0$ false hypotheses, we can claim $k - 1 \leq S - S_0$. This equation implies $\frac{1}{S-k+1} \leq \frac{1}{S_0}$. Since kth hypothesis is still rejected, $p_k \leq \frac{\alpha}{S-k+1} \leq \frac{\alpha}{S_0}$. So, there exists a true hypothesis whose p-value is bounded above by α/S_0. Then

$$\begin{aligned} FWE &= \Pr(\text{Reject at least one true hypothesis}) \\ &\leq \Pr(p_s \leq \alpha/S_0 \text{ for some true hypothesis}) \\ &\leq \sum_{s\in I_0} \Pr(p_s \leq \alpha/S_0) = S_0 \times (\alpha/S_0) = \alpha \end{aligned}$$

where I use the fact that if null hypothesis is true and the statistic is continuous, p-value has $U[0,1]$ distribution. As for the general case, we can do as follows

Theorem 5.7.2 (From Lehmann and Romano (2005)[a]). *For testing H_{0s} where $s = 1,..,S$, if p_s is $U[0,1]$, then following the procedure where H_{0s} is rejected if*

$$\alpha_s = \begin{cases} \frac{k\alpha}{S} & (s \leq k) \\ \frac{k\alpha}{S-s+k} & (s \geq k) \end{cases} \quad (\alpha_s \textit{ is the cutoff p-value for sth null})$$

controls k-FWE. This is equivalent to using $\frac{k\alpha}{\min\{S-s+k,S\}}$ as cutoff.

Proof. Assume that $n(I_0) = S_0 \geq k$. Let j be index of the kth true null hypothesis rejected. Then, $j-k$ of the hypotheses so far ar false, thus,

$$j-k \leq S - S_0 \implies k \leq j \leq S - S_0 + k$$

The Holm procedure would reject k true null hypothesis iff $p_s \leq \alpha_s$ for $s = 1,...,j$. Since we have $j \geq k$, it must be that

$$p_j \leq \alpha_j = \frac{k\alpha}{S-j+k} \leq \frac{k\alpha}{S_0}$$

This implies that when at least k false rejections are made, there exists null hypotheses whose p-values is bounded above by $\frac{k\alpha}{S_0}$. We can then apply Theorem 2.1 to show that the k-FWE is bounded above by α. □

[a] Lehmann, E. L. and J. P. Romano (2005) Generalizations of the Familywise Error Rate, *The Annals of Statistics*, 33(3), 1138-1154

Note that Holm method is not as conservative as Bonferroni method in that more rejections are likely to happen. They share the common feature in that no assumption about the (in)dependence structure of the hypotheses are required.

5.8 Controlling FDR: Benjamini & Hochberg Method

The goal of this approach is to set the false discovery rate to be below some $\gamma \in [0,1]$. The procedure is as follows

1. Order p-values from the lowest to the highest: $p_1 < ..., < p_S$
2. Make the following adjustment:

$$\tilde{p}_s = \min\left\{\frac{Sp_j}{j} | j \geq s\right\}$$

3. We move from the Sth hypothesis. If $\tilde{p}_S \leq \gamma$, then all S hypothesis are rejected and the process ends. Otherwise, move to $S-1$th hypothesis.
4. At some point, you will find i s.t. $\tilde{p}_i \leq \gamma$. Then reject the first i hypotheses.

So how can we do this? We do not reject H_{0s} iff for all $j \geq s, p_j \geq \frac{j\gamma}{S}$ $(\geq \frac{s\gamma}{S})$. So what we can do is to plot the ordered p-values and the threshold line $p(s) = \frac{s\gamma}{S}$. If a p-value of the sth hypothesis is larger than $\frac{s\gamma}{S}$, then that hypothesis can not be rejected. In this manner, we can find s^*, the largest index s.t. the p-value is below the $p(s)$ plot. Then, we reject all H_{0s} s.t. $s \leq s^*$.

We can show that in suitable conditions, this method yields $FDR = \frac{S_0}{S}\gamma$. This is still conservative method in that $S_0 \leq S$. We can do better by estimating what S_0 would be and using $\frac{S}{S_0}\gamma$ as cutoff instead.

Chapter 6

Quantile Regression

6.1 Introduction: A Check Function Approach

The usual linear regression, which has the form

$$y = X\beta + u$$

where $E(Xu) = 0$ (or more restrictively, $E(u|X) = 0$) is aimed at capturing the conditional mean of y given X at a certain value. However, there is no reason to restrict our attention to just a conditional mean. In fact, we an do more. For instance, what is the conditional median? What about the top 10%, or 1%? **Quantile regression** aims to capture different values of β depending on the location of the conditional distribution.

Rigorously speaking, the quantile regression seeks to estimate the conditional quantile

$$q_\tau(y|X) = X\beta_\tau$$

where $\tau \in [0, 1]$ is the percentile of our choice satisfying $F_{y|X}(X\beta_\tau|X) = \tau$. To characterize the conditional quantile function, we need to use the **check function**. Since we have

$$F_{y|X}(X\beta_\tau|X) = \Pr(y \leq X\beta_\tau|X) = \tau$$

we can write as

$$\tau - \Pr(y \leq X\beta_\tau|X) = 0$$

Since $\Pr(y \leq X\beta_\tau|X)$ is equal to $E[1(y - X\beta_\tau \leq 0)] = E[1(u \leq 0)]$, we can obtain the condition

$$E[\tau - 1(y - X\beta_\tau \leq 0)|X] = 0$$

and using the law of iterated expectations, this also implies $E[(\tau - 1(y - X\beta_\tau \leq 0))X] = 0$. The check function can be defined as

$$\rho_\tau(u) = u(\tau - 1(u \leq 0))$$

6.2 Usage of Check Function

To see what this check function is, I will talk about two specific cases

- Median: Let $\tau = 1/2$. Then the check function becomes

$$\rho_{1/2}(u) = \begin{cases} -\frac{1}{2}u & (u \leq 0) \\ \frac{1}{2}u & (u > 0) \end{cases} = \frac{1}{2}|u| = \frac{1}{2}|y - X\beta_{1/2}|$$

  This becomes equivalent to solving the least absolute deviation problem.

- $\tau = 1/3$: Then the check function becomes

$$\rho_{1/3}(u) = \begin{cases} -\frac{2}{3}u & (u \leq 0) \\ \frac{1}{3}u & (u > 0) \end{cases}$$

  which has a kink at $u = 0$ and is asymmetric.

For all values of τ the check function has a kink at $u = 0$. However, the check function is convex and can be shown to be a lipschitz function.

Theorem 6.2.1 (Convex Functions are Lipschitz Functions?). *Let f be a convex function and K be a closed, bounded set contained in the relative interior of the domain of f. Then f is Lipschitz continuous on K.*

Proof. Suppose that f is convex on $[a, b]$. Select c, d s.t. $a < c < d < b$ and let $t_1 \in [d, b]$, $t_2 \in [a, c]$ and $x_1, x_2 \in [c, d]$. Since f is convex we can write

$$\frac{f(c) - f(t_2)}{c - t_2} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(t_1) - f(d)}{t_1 - d}$$

Then, let $L = \max\left[\left|\frac{f(c)-f(t_2)}{c-t_2}\right|, \left|\frac{f(t_1)-f(d)}{t_1-d}\right|\right]$, we then have

$$|f(x_2) - f(x_1)| \leq L|x_2 - x_1|$$

□

6.3 Obtaining Quantile Regression Estimator

The minimization problem that should be solved in quantile regression is

$$\min_{\beta} E[\rho_\tau(y - X\beta_\tau)|X]$$

which can be written as

$$E[\rho_\tau(y - X_i\beta_\tau)|X] = (\tau - 1)\int_{-\infty}^{a}(y-a)f_{Y|X}(y|x)dy + \tau\int_{a}^{\infty}(y-a)f_{Y|X}(y|x)dy$$

where $a = X\beta_\tau$. Take the first order condition w.r.t. a to get

$$-(\tau - 1)\int_{-\infty}^{a} f_{Y|X}(y|x)dy - \tau\int_{a}^{\infty} f_{Y|X}(y|x)dy = 0$$
$$\iff -\tau + F(a|X) = 0$$

Thus, the β_τ that solves

$$X\beta_\tau = a = F^{-1}_{Y|X}(\tau|X)$$

is the β_τ that we are looking for. We can also solve for

$$\min_{\beta} E[\rho_\tau(y - X\beta_\tau)X]$$

or in a sample analogue

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - X_i\beta_\tau)X_i$$

to find the quantile estimator β_τ. For suitable conditions, the resulting estimator is CAN.

Example 6.3.1 (Levin, 2001)**.** *As part of the education production function literature, the effect of class size on various outcomes for the student is controversial (Lazear, 2001). This paper, while controlling for a large number of observable characteristics and endogeneity in class size variable, estimates education production function using a quantile regression. In doing so, it attempts to analyze the conventional claims that reducing class size improves learning directly via and increased allocation of teaching per student. While it does not find statistically significant impact of class size once peer effect is controlled for, peer effect has a positive and significant effect for those in the lower portion of the achievement distribution for math and language scores.*

Example 6.3.2 (Autor, Houseman, & Kerr 2017[a])**.** *Using the Detroit's welfare-to-work program, this paper studies the the effect of two government employment programs - direct hire assistance and temporary-help job placements - on distribution of participant's earnings over a 7-quarter period. The paper finds that for the low-tail of the earnings distribution, neither programs are effective. The direct-hire increases earnings for the high-tail but temporary-help placements negatively affect the distribution for the same group. Autor and Houseman (2010) have studied the same program 7 years ago without using the quantile approach and find that on average, the same program lead to earnings gain. The key takeaway is that by using quantile regression, you can unmask the effect at a different quantile that you cannot find out through conditional expectations.*

[a] Technically, this paper uses Chernozhukov-Hansen Instumental Variables Quantile Regression, which is not part of our curriculum yet. You may learn this in Microeconometrics course next year.

Chapter 7

Nonparametric and Semiparametric Models

7.1 Nonparametric Regression Models: Introduction

Consider a setting where we observe the data (y_i, x_i) which is i.i.d. and drawn from a DGP $P_0(y|x)$. Unlike in the parametric setup, we are interested in backing out the whole or part of the data generating process $P_0(y|x)$ **without any modeling assumptions**. In essence, we are interested in an estimation problem involving an unknown function. This approach is called a **nonparametric** approach. We normally use nonparametric approach to conduct a diagnostic checking of an estimated parametric model, to conveniently display key features of the dataset, and to conduct an inference under very weak assumptions.

For discrete-valued Y and X, this is relatively easy. We can back out the data generating process of interest

$$\hat{P}(y \in A|x \in B) = \frac{n^{-1}\sum_{i=1}^{n} 1(y_i \in A, x_i \in B)}{n^{-1}\sum_{i=1}^{n} 1(x_i \in B)}$$

Provided that $P_0(x \in B) \neq 0$, then the above estimator converges to the true $P_0(y|x)$ as $n \to \infty$.

When we are working with continuous variables to get an estimate of $P_0(Y \leq y|x)$, one way we could approach is to define A in the above example as $A \equiv (-\infty, y], B \equiv [x - \epsilon, x + \epsilon]$ and use $\lim_{\epsilon \to 0} \hat{P}(A|B)$ to back out the DGP. This can be problematic when we have too many dimensions of X (and by too many, the number required for this to happen is not that large). Therefore, the method to conduct nonparametric estimation in this context is kernel density function approach.

7.2 Kernel Density Estimation

Consider the problem of estimating the probability density function $f(x)$ of a random scalar variable X at $X = x$. We are putting a minimal number of restriction on what $f(x)$ could be. Let $\{X_1, ..., X_n\}$ be a random sample of X. If f is a smooth function, we can approximate f by

$$f(x) \simeq \frac{\int_{x-h}^{x+h} f(u)du}{2h}$$

for a small h (which will later be our bandwith). The numerator on the right hand side can be estimated using a sample analogue of the following form

$$\frac{1}{n}\sum_{i=1}^{n} 1\{x - h \leq X_i \leq x + h\}$$

Combining the two expressions, we can approximate $f(x)$ with

$$\begin{aligned}
\hat{f}(x) &= \frac{n^{-1}\sum_{i=1}^{n} 1[x - h \leq X_i \leq x + h]}{2h} \\
&= \frac{1}{2nh}\sum_{i=1}^{n} 1[x - h \leq X_i \leq x + h] \\
&= \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2} 1\left[\left|\frac{x - X_i}{h}\right| \leq 1\right] \\
&= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)
\end{aligned}$$

where $K(u) = \frac{1}{2}1(|u| \leq 1)$ and $K_h = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. In general, $K(\cdot)$ is a **kernel density estimator**. Specifically for this example, I have used a uniform kernel.

As for the type of kernel density estimators, there can be many. The requirements are that K is a real-to-real function, positive and smooth and integrates to 1 on its support. For instance, you can use a Gaussian Kernel, $K = \phi$ or an Epanechinkov kernel, defined as $K(u) = \frac{3}{4}\max\{1 - u^2, 0\}$. However, the selection of K does not matter in most cases.

7.3 Choice of the bandwidth h

While we may have some freedom in choosing K, same cannot be said for the bandwidth h. This is because our choice of h directly affects the trade-off between the bias and the variance. To see this.

$$\begin{aligned}
E[\hat{f}(x)] &= E\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right] \\
&= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right) f(s)ds \\
&= -\frac{1}{h}\int_{-\infty}^{\infty} K(t)f(x-ht)(-hdt) \ (\because \frac{x-s}{h} = t \text{ transformation}) \\
&= \int_{-\infty}^{\infty} K(t)f(x-ht)dt \\
&= \int_{-\infty}^{\infty} K(t)\left[f(x) - f'(x)ht + \frac{f''(x)h^2t^2}{2} + o(h^2)\right] dt \\
&= f(x) - 0 + \frac{1}{2}\int_{-\infty}^{\infty} K(t)h^2t^2f''(x)dt + o(h^2)
\end{aligned}$$

Where $\int_{-\infty}^{\infty} K(t)dt = 1$ justifies $f(x)$ and $\int_{-\infty}^{\infty} tK(t)dt = 0$ (since kernel is symmetric around 0) justifies second term in the last line being 0.[1] As such, the bias is captured by

$$E[\hat{f}(x)] - f(x) = \frac{1}{2}\int_{-\infty}^{\infty} K(t)h^2t^2f''(x)dt$$

Therefore, smaller bandwidth h reduces the bias.

However, for variance, we have

$$\begin{aligned}
Var[\hat{f}(x)] &= E[\hat{f}^2(x)] - (E[\hat{f}(x)])^2 \\
&= E\left[\frac{1}{n^2h^2}\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right)^2\right] - (E[\hat{f}(x)])^2 \\
&= E\left[\frac{1}{n^2h^2}\left(\sum_{i=1}^{n} K^2\left(\frac{x-X_i}{h}\right) + 2\sum_{i<j} K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)\right)\right] - (E[\hat{f}(x)])^2 \\
&= \frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\left(\frac{x-s}{h}\right) f(s)ds + \frac{n(n-1)}{n^2h^2}\left(\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right) f(s)ds\right)^2 - \frac{1}{h^2}\left(\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right) f(s)ds\right)^2
\end{aligned}$$

Then, we use the Taylor expansion and the variable transformation $\frac{x-s}{h} = t$ on each of the terms. As a result, we can show that the first term can be written

$$\frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\left(\frac{x-s}{h}\right) f(s)ds = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)f(x-ht)dt \simeq \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)f(x)dt$$

so that $\frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)f(x)dt$ is the leading term. The other two terms will be cancelled out as

[1] In the lecture slides, $O(h^2) = \frac{f''(x)h^2t^2}{2} + o(h^2)$

$n \to \infty$, as

$$\left(\frac{n(n-1)}{n^2h^2} - \frac{1}{h^2}\right)\left(\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right) f(s)ds\right)^2 = -\frac{1}{nh^2}\left(\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right) f(s)ds\right)^2$$

Since the first term can be written as $O\left(\frac{1}{nh}\right)$, smaller h increases the variance.

So what determines the optimal h? The answer is the h that minimizes the loss function, or asymptotic mean integrated squared error (AMISE), defined as

$$\int E(\hat{f}(x) - f(x))^2 dx$$

where the bias-variance decomposition of MSE kicks in as follows

$$\begin{aligned}
E[(\hat{f}(x) - f(x))^2] &= E[(\hat{f}(x) - E(\hat{f}(x)) + E(\hat{f}(x)) - f(x))^2] \\
&= E[(\hat{f}(x) - E(\hat{f}(x)))^2 + (E(\hat{f}(x)) - f(x))^2 + 2(\hat{f}(x) - E(\hat{f}(x)))(E(\hat{f}(x)) - f(x))] \\
&= E[(\hat{f}(x) - E(\hat{f}(x)))^2] + E[(E(\hat{f}(x)) - f(x))^2] \\
&= \text{Variance} + \text{Bias}^2
\end{aligned}$$

From the discussion about the bias and variance above, we can write

$$\text{Variance} + \text{Bias}^2 = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t) f(x) dt + \frac{h^4}{4}(f''(x))^2 \left(\int_{-\infty}^{\infty} K(t)t^2 dt\right)^2$$

Thus, the final version of AMISE can be written as

$$\int (\text{Variance} + \text{Bias}^2) dx = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t) dt + \frac{h^4}{4}\int_{-\infty}^{\infty} (f''(x))^2 \left(\int_{-\infty}^{\infty} K(t)t^2 dt\right)^2 dx$$

For a shorthand notation, we can write $A = \frac{1}{4}\int_{-\infty}^{\infty}(f''(x))^2\left(\int_{-\infty}^{\infty} K(t)t^2 dt\right)^2 dx$, $B = \int_{-\infty}^{\infty} K^2(t)dt$ to get

$$AMISE = Ah^4 + \frac{B}{nh}$$

Then, the minimization problem becomes $\min_h AMISE$. Therefore, we find h satisfying

$$4Ah^3 - Bn^{-1}h^{-2} = 0 \iff h^5 = \frac{B}{4An} \iff h = \left(\frac{B}{4An}\right)^{1/5}$$

In this framework, the bias and standard errors are both in $n^{-2/5}$ and AMISE will be in $n^{-4/5}$. Therefore, we may not have a CAN estimator at $n^{-1/2}$, which is what is normally the case.

Even bigger problem arises from $\int f''(x)dx$ term in A, as we are not sure of $f(x)$ to begin with, we do not know what $f''(x)$ would be. There are several ways to deal with this.

- **Silverman's Rule of Thumb**: If K is a normal kernel and $f(x)$ is normal with some variance σ^2, Silverman's rule of thumb implies the benchmark

$$h = 1.06\sigma n^{-1/5}$$

  where σ can be replaced with sample standard deviation s. For a more robust version, we can use

$$h = 0.9 \min\{s, IQ/1.34\} n^{-1/5}$$

  where IQ is an interquartile distance.

- **Cross-validation**: This method is similar to the original method in the sense that we are selecting some h that minimizes certain criterion function, $CV(h)$. In this case, we are using a leave-one-out estimator. This is usually calculated as

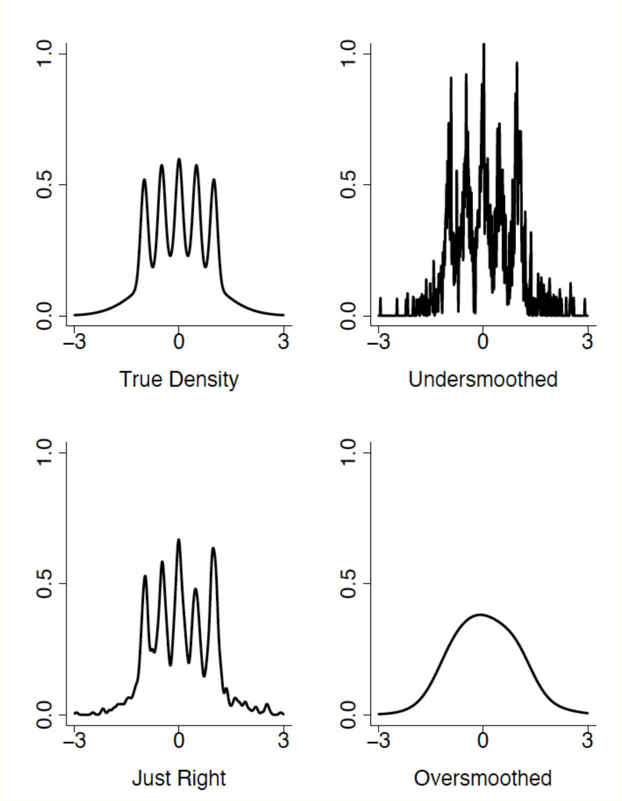
$$\hat{f}_{-i}(x) = \frac{1}{nh} \sum_{j \neq i} K\left(\frac{x - x_j}{h}\right)$$

  for every observation i. This is computationally burdensome and the resulting h converges very slowly ($n^{-1/10}$).

- **Local Bandwidth**: You may care only about $f(x)$ at a given point. The key idea is to make h larger in a low density area, as low density could imply larger spread in the distribution of $f(x)$. Making h larger would decrease variance. In a wiggly region, it is better to take h smaller. The distribution can take many values and bias can be problematic. Reducing h would minimize worries about biases.

Comment 7.3.1 (So why worry about bandwidth to begin with?)**.** *If h is smaller than the optimal bandwidth length, we say that there is an **undersmoothing**. If otherwise, we say **oversmoothing** is present. When there is an undersmoothing, there are too much more detail than necessary and in some areas, leading to computational difficulty. When there is an oversmoothing, we risk masking some important detail about the distribution of $f(x)$. We may end up with a normal distribution even if it really is not.*

Below is a figure that I borrowed from a statistics lecture note from CMU.[a]



[a]https://www.stat.cmu.edu/ larry/=sml/densityestimation.pdf

7.4 Curse of Dimensionality

Now assume that x is not necessarily a scalar, but of dimension d. Then we can use a d-dimensional kernel K and estimate $f(x)$ with

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where K can be a d-product of a uni-dimensional kernels. In addition, we are not necessarily confined to using a same bandwith for all d kernels. For instance, if you know that the variance of x_1 is larger than x_2 you may want higher h_1 relative to h_2 to avoid variance of the kernel density estimator for the first covariate from being too large. We may want to sphericize if the kernel estimators are correlated.

A bigger concern has to do with the computation cost from using a large d. This is what **curse of dimensionality** refers to. If we return to calculating $E[\hat{f}(x)^2]$ term in the variance component of AMISE,

$$
\begin{aligned}
E[\hat{f}(x)^2] &= E\left[\frac{1}{n^2h^{2d}}\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right)^2\right] \\
&= E\left[\frac{1}{n^2h^{2d}}\left(\sum_{i=1}^{n} K^2\left(\frac{x-X_i}{h}\right) + 2\sum_{i<j} K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)\right)\right] \\
&= \frac{1}{nh^{2d}}\int_{-\infty}^{\infty} K^2\left(\frac{x-s}{h}\right)f(s)ds + \frac{n(n-1)}{n^2h^{2d}}\left(\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right)f(s)ds\right)^2
\end{aligned}
$$

Again, the leading term is

$$
\frac{1}{nh^{2d}}\int_{-\infty}^{\infty} K^2\left(\frac{x-s}{h}\right)f(s)ds \simeq \frac{1}{nh^d}\int K^2(t)f(x-ht)dt
$$

since s is actually d-dimensional. Thus, the variance is now $O\left(\frac{1}{nh^d}\right)$. What this means is that $AMISE = Ah^4 + \frac{B}{nh^d}$ and the optimal h is solved as

$$
h = \left(\frac{Bd}{4An}\right)^{1/(4+d)}
$$

Thus, h will be in $n^{-\frac{1}{4+d}}$ and convergence occurs in $n^{-\frac{2}{4+d}}$ - at an even slower rate that when $d = 1$, which was already slower than parametric convergence rate. AMISE would be in order $n^{-\frac{4}{4+d}}$

7.5 Local Constant and Polynomial Estimation

Given the data (y_i, x_i), we are attempting to capture $E[g(y,x)|x] = m(x)$ for some $g(y,x)$. For instance, we can attempt to estimate the conditional expectation by letting $g(y,x) = y$. Note that the conditional expected value can be written as

$$
\begin{aligned}
E[y|x] &= \int y f_{Y|X}(y|x)dy \\
&= \int y\frac{f_{Y,X}(y,x)}{f_X(x)}dy = \frac{\int y f_{Y,X}(y,x)dy}{\int f_{Y,X}(y,x)dy}
\end{aligned}
$$

We can obtain the nonparametric estimator for the conditional expected value by replacing $f_{Y,X}$ with its kernel estimator. For simplicity, I will assume that the dimension of both y and

x is 1. The numerator becomes

$$\begin{aligned}
\int y\hat{f}(y,x)dy &= \int y \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) K\left(\frac{y-Y_i}{h}\right) dy \\
&= \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \int yK\left(\frac{y-Y_i}{h}\right) dy \\
&= \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \int (Y_i+sh)K(s)\,(hds)\ (\because s = \frac{y-Y_i}{h}) \\
&= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i\ (\because \int K(s)ds = 1, \int sK(s)ds = 0)
\end{aligned}$$

The denominator can be written as $\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$. Thus, the estimator for the conditional expectation becomes

$$\frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$$

Effectively we are putting weight $\frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$ on each observation Y_i. This is also a **local constant estimation** in the sense that when solving the following minimization problem

$$\hat{f}(x) = \arg\min_{a} \frac{1}{nh}\sum_{i=1}^{n}(Y_i - a)^2 K\left(\frac{x-X_i}{h}\right)$$

The first order condition on a yields the following results

$$\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right) = a\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \implies a = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$$

As for the optimal h, it is possible to use cross-validation method and minimizing AMISE. This method can be unreliable if $f(x) \to 0$.

We can also do something more general. For instance, a **local linear estimation**. This is when we regress $g(y,x)$ on a constant (a) and a linear term ($b(x - X_i)$). In mathematical expression, we solve

$$\min_{a,b} \frac{1}{nh}\sum_{i=1}^{n}(Y_i - a - b(x-X_i))^2 K\left(\frac{x-X_i}{h}\right)$$

and obtain that $\hat{a} = \hat{g}$ and $\hat{b}$ is an estimate of $\frac{\partial g(x)}{\partial x}$. If it happens that the true functional form is linear, then this estimate does not produce a bias. In addition, local linear estimation performs better than local constant estimation in the boundaries of the support for X.

We can do even more with **local polynomial estiation** by regressing $g(y,x)$ on a constant, $x - X_i$, $(x - X_i)^2$ and so on.

7.6 Semi-nonparametrics

- Flexible approach: Suppose that we are sure that $f(x)$ can be characterized by $f_{m,\sigma}$ where m, σ indexes some properties of the density function f. Then, by Weierstrass approximation theorem, we can choose a family of positive functions which increases in complexity $P_\theta^1, P_\theta^2, ...$ and maximize over the loglikelihood

$$\sum_{i=1}^{n} \log f_{m,\sigma}(X_i) P_\theta^M(X_i)$$

Theorem 7.6.1 (Weierstrass Approximation Theorem)**.** *If f is a continuous real-valued function on $[a,b]$ and if any $\epsilon > 0$ is given, then there exists a polynomial P on $[a,b]$ s.t.*

$$|f(x) - P(x)| < \epsilon$$

for all $x \in [a,b]$

- Mixture of normals: Suppose that $Y|X$ is drawn from the two distributions

  - $N_1(m_1(x,\theta), \sigma_1^2(x,\theta))$ with probability $q_1(x,\theta)$
  - $N_2(m_2(x,\theta).\sigma_2^2(x,\theta))$ with probability $q_2(x,\theta)$

  Then, we apply a maximum likelihood of the following form

$$\min_\theta \sum_i \sum_k q_k(x_i,\theta)[(y_i - m_k(x_i,\theta))'\sigma_k(x_i,\theta)^{-1}(y_i - m_k(x_i,\theta)) + \log\det \sigma_k(x_i,\theta)]$$

  As with other MLE estimators, there could be more than one local maxima. Furthermore, the optimal number of k is difficult to determine in practice.

- Series estimator: Let $\{P_k(x_i)|k = 1, 2...\}$ be the orthonormal basis for a smooth function. By orthonormal, it means that

  $$\int P_k(x)^2 dx = 1, \int P_k(x)P_m(x) = 0 \ (k \neq m)$$

  These could be polynomials of degree k, sine functions and so on. What we do here is to run a linear regression that has the following form

  $$y_i = \sum_{k=1}^{M} P_k(x_i)\theta_k + \epsilon_i$$

  The $\sum_{k=1}^{M} P_k(x_i)\theta_k$ part is a series approximation to $g(x)$. However, depending on the number of M that we choose, the curse of dimensionality can kick in.

7.7 Semiparametric Regresion

Semiparametrics can be thought of as a middle ground between nonparametric and parametric regression. Suppose that we partition the covariates into two unoverlapping spaces - X and W. Also assume that $E(\epsilon|X, W) = 0$. One example of a semiparametric regression is a **partially linear regression** which has the following form

$$Y_i = X_i\beta + g(W_i) + \epsilon_i$$

where $\beta \in \dim(X_i)$ represents a coefficient for the linearly regressed terms and $g(\cdot)$ is a nonparametric portion of the regression.

To estimate β, we use the fact that

$$\begin{aligned} E[Y_i|W_i] &= E[X_i|W_i]\beta + E[g(W_i)|W_i] + E[\epsilon_i|W_i] \\ &= E[X_i|W_i]\beta + g(W_i) + E[\epsilon_i|W_i] \\ &= E[X_i|W_i]\beta + g(W_i) + E[E[\epsilon_i|X_i, W_i]|W_i] = E[X_i|W_i]\beta + g(W_i) \end{aligned}$$

Given this, we can write

$$Y_i - E[Y_i|W_i] = \{X_i - E[X_i|W_i]\}\beta + \epsilon_i$$

Then we follow this procedure

1. Nonparametrically estimate $E[X_i|W_i]$ and $E[Y_i|W_i]$. Then define $\tilde{X}_i = X_i - \hat{E}[X_i|W_i]$ and $\tilde{Y}_i = Y_i - \hat{E}[Y_i|W_i]$, where $\hat{E}$ are nonparametric estimators

2. Regress $\tilde{Y}$ onto $\tilde{X}$ to get an estimate of β

3. We can estimate $g(\cdot)$ by nonparametrically regressing $Y_i - X_i\hat{\beta}$ onto W_i

As far as approximation is concerned, β follows the properties of parametric estimators. That is, it converges at rate $n^{-1/2}$ regardless of the dimensions of X_i, W_i. Thus there is no curse of dimensionality here. Unfortunately, estimating $g(\cdot)$ follows the same properties as nonparametric estimators. It has a slower convergence rate, which becomes even slower with more dimensions of W_i.

There is another caveat to this method. Identification of β requires an exclusion restriction in the sense that none of the components in X_i is perfectly predictable by W_i components. If otherwise, $X_i = E[X_i|W_i]$. Thus, β cannot be identified. This would effectively rule out including a constant in the X_i part of the regression. If we do have $X_i \neq E[X_i|W_i]$, we have

$$\begin{aligned} &(X_i - E[X_i|W_i])'(Y_i - E[Y_i|W_i]) = (X_i - E[X_i|W_i])'[(X_i - E[X_i|W_i])\beta + \epsilon_i] \\ \implies\ &E\{(X_i - E[X_i|W_i])'(X_i - E[X_i|W_i])\}\beta = E\{(X_i - E[X_i|W_i])'(Y_i - E[Y_i|W_i])\} \end{aligned}$$

The value of β that solves the above equation is the estimator we are looking for.

Example 7.7.1 (Horowitz, Lee (2002))**.** *The paper reviews several semiparametric methods for estimating conditional mean functions and show that semiparametrics allow more flexibility than parametric modeling and more precision than nonparametric models. For fun (at least for a baseball nerd like me), this paper tests this idea on a data of salaries, runs, tenure of baseball players in 1987.*

Example 7.7.2 (Ucal et al (2010))**.** *This paper analyzes whether and to what extent the inflow of FDI is affected before and after the occurence of a financial crisis in developing countries using generalized partial linear models. The results indicate that FDI inflows decrease in the years after a financial crisis and an upturn in FDI inflows the year before a financial crisis hit the country.*

Chapter 8

Treatment Effects: Random Assignment and Estimation under Unconfoundedness

8.1 Introduction to Treatment Effects: Causality and Treatment Assignment

Now, our interest is in finding out whether some X variable **causes** Y, not just whether they are correlated. Generally, correlation is not enough to argue causation. Suppose that you find that $cov(X, Y) \neq 0$. This can happen because

- X do cause Y, which is good for us. But..
- Y could also cause X. So there is a reverse causality bias here
- Z mutually affects X and Y. This is an omitted variable bias and leads to nonzero correlation even if X and Y has no connection whatsoever.

In any study examining program evaluation or quasi-experimental setting, the key issue is the assignment of the treatment, which is considered binary for the moment. There are three possible cases, which are

- **Random Assignment**: This would be a case when we can guarantee that the assignment to the treatment arms (treatment and control) are determined by chance - like flipping a coin or random draw of numbers. This rarely happens in social science, as people can voluntarily opt-in and opt-out of treatment.

- **Selection on Observables**: The treatment assignment is effectively random once we condition on some observable covariates. This is also known as ignorability or unconfoundedness assumptions.

- **Selection on Unobservables**: The assignment depends fundamentally on unobservables, or in other words, we cannot break down the dependence structure of assignment using observed variables. For instance, even in a clinical trials that involve voluntary participation, it is very likely that (lack of) risk averseness determines participation. How do we control for risk averseness?

8.2 Theoretical Framework for Treatment Effect

We will consider a binary treatment variable - whether individual i received a treatment or not (in the treatment group or control group). Define a variable D_i s.t.

$$D_i = \begin{cases} 1 & \text{If treated} \\ 0 & \text{If not treated} \end{cases}$$

In the above notation, i indexes the unit of the treatment - an individual, household, county, firm, and etc. For each unit i, there are two possible outcomes. One is the outcome without treatment, $Y_i(0)$. The other is the outcome with treatment, $Y_i(1)$. This can be seen as a counterfactual framework in the sense that if we get $Y_i(1)$ for unit i, we cannot get $Y_i(0)$ and vice versa. We always have a missing data problem in this regard. A useful way of writing this is

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

The left hand side is an outcome for unit i, treated or untreated. *This is observed for every i.* The right hand side is meant to capture that the observed outcome for individual i is either *one* of $Y_i(1)$ or $Y_i(0)$. This framework is sometimes called **potential outcome framework** and will come in handy when we analyze various treatment effect results.

We are interested in how an outcome for unit i changes between treatment and control. In other words,

$$TE_i = Y_i(1) - Y_i(0)$$

Another parameter of interest could be the treatment effect averaged over those who share the common covariate value, which is

$$TE(x) = E(TE_i|X_i = x)$$

We could also be interested in the treatment effect averaged over the population. This is the average treatment effect, defined as

$$ATE = E(TE_i) = E(Y_i(1) - Y_i(0))$$

Of course there are more, most of which we will learn over the course of this sequence. There are average treatment effect on the treated, ATE for the untreated, etc.

Because of the fundamental problem of a missing data, we are forced to make an assumption about the things we cannot observe. Take the ATE for example. We can write

$$\begin{aligned}E[Y_i(1) - Y_i(0)] &= E[Y_i(1)] - E[Y_i(0)]\\ &= \{\Pr(D_i = 1)E[Y_i(1)|D_i = 1] + (1 - \Pr(D_i = 1))E[Y_i(1)|D_i = 0]\}\\ &- \{\Pr(D_i = 1)E[Y_i(0)|D_i = 1] + (1 - \Pr(D_i = 1))E[Y_i(0)|D_i = 0]\}\end{aligned}$$

Here is the problem: We can get what $E[Y_i(1)|D_i = 1]$ and $E[Y_i(0)|D_i = 0]$ are from the data. This is because [1]

$$\begin{aligned}E[Y_i|D_i = 1] &= E[1 \cdot Y_i(1) - (1 - 1) \cdot Y_i(0)|D_i = 1] = E[Y_i(1)|D_i = 1]\\ E[Y_i|D_i = 0] &= E[0 \cdot Y_i(1) - (1 - 0) \cdot Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 0] \qquad \text{(TE)}\end{aligned}$$

So we can get $E[Y_i(1)|D_i = 1]$ and $E[Y_i(0)|D_i = 0]$ from the data - take the expected value of observed Y_i conditional on $D_i = 1$ and 0. We cannot do the same for $E[Y_i(1)|D_i = 0]$ and $E[Y_i(1)|D_i = 0]$. This forces us to make an assumption about what the two unobserved values could be.

Before going on, it would be useful to characterize the counterfactual outcome in a more econometrics-friendly context. Define the counterfactual outcomes as

$$Y_i(D_i = d) = \mu(X_i, d) + \epsilon_i(d)$$

where d can take either $0, 1$. What we want to learn involves understanding the joint distri-

[1] You might have noticed that I rarely tag equations in my recitation notes. So when I do, it means I am going back to this very often.

bution of the variables $Y_i(0)$ and $Y_i(1)$. Take the treatment effect at $X_i = x$, written as

$$TE(x) = \mu(x,1) - \mu(x.0)$$

This is the effect of the treatment on the outcome Y_i for $X_i = x$. The interpretation is that if we shift everyone with $X_i = x$ from the control group to the treatment group, the average outcome increases by $TE(x)$. We can also calculate ATE as

$$\begin{aligned}ATE &= E[Y_i(1) - Y_i(0)] = E[\mu(X_i,1) - \mu(X_i,0) + \epsilon_i(1) - \epsilon_i(0)] \\ &= E\left[E[\mu(X_i,1) - \mu(X_i,0) + \epsilon_i(1) - \epsilon_i(0)|X_i]\right] = E\left[E[\mu(X_i,1) - \mu(X_i,0)|X_i]\right] \\ &= E[\mu(X_i,1) - \mu(X_i,0)] = E[TE(X_i)]\end{aligned}$$

8.3 Randomized Assignment

We are back to having to assume what $E[Y_i(1)|D_i = 0]$ and $E[Y_i(1)|D_i = 0]$ are. A **random assignment** assumes that the outcome is independent of the treatment status. More formally

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \tag{RA}$$

This implies that

$$E[Y_i(1)] = E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$$

and similarly for $E[Y_i(0)]$. If we go back to the (TE) equation, the problem was that we were unable to make any statement about $E[Y_i(1)|D_i = 0]$ and $E[Y_i(1)|D_i = 0]$. Now what we are doing is to equate

$$\begin{aligned}E[Y_i|D_i = 1] = E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0] \\ E[Y_i|D_i = 0] = E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]\end{aligned}$$

and effectively make the statements about the previously unobserved values. This allows us to rewrite the ATE as

$$\begin{aligned}E[Y_i(1) - Y_i(0)] &= E[Y_i(1)] - E[Y_i(0)] \\ &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \ (\because \text{RA}) \\ &= E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \ (\because \text{TE})\end{aligned}$$

Therefore, we can estimate ATE by mapping $E[Y_i(1)]$ to $E[Y_i|D_i = 1]$, $E[Y_i(0)]$ to $E[Y_i|D_i = 0]$.

8.4 Conditional Independence Assumption

In this setup, we are assuming that conditional on X_i, the outcomes and D_i are independent. Formally, we can write

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i|X_i \tag{CIA}$$

We can alternatively conceptualize this framework in the following way: Define D_i as

$$D_i = 1(u_i < p(X_i))$$

where $u_i \equiv U[0,1]$ determines selection and $p(X_i)$ can be interpreted as a propensity score - how much are you likely to be treated given that you have covariates specified as some X_i. This framework, along with the (CIA), allows us to say the following:

$$\begin{aligned}
E[Y_i(1)|X_i = x] = E[Y_i(1)|D_i = 1, X_i = x] &= E[Y_i(1)|D_i = 0, X_i = x] \ (\because \text{CIA}) \\
&\implies E[\mu(x,1) + \epsilon_i(1)|D_i = 1, X_i = x] = E[\mu(x,1) + \epsilon_i(1)|D_i = 0, X_i = x] \\
&\implies E[\mu(x,1)|D_i = 1, X_i = x] - E[\epsilon_i(1)|D_i = 1, X_i = x] \\
&= E[\mu(x,1)|D_i = 0, X_i = x] + E[\epsilon_i(1)|D_i = 0, X_i = x] \\
&\implies E[\epsilon_i(1)|D_i = 1, X_i = x] = E[\epsilon_i(1)|D_i = 0, X_i = x] \ (\because \mu(x,1) \text{ is same for both cases}) \\
&\implies E[\epsilon_i(1)|u_i < p(x), X_i = x] = E[\epsilon_i(1)|u_i > p(x), X_i = x] \\
&\implies (\epsilon_i(1), \epsilon_i(0)) \perp\!\!\!\perp u_i|X_i
\end{aligned} \tag{CIA2}$$

Thus, conditional on observed values for the covariate X_i, the unobserved elements of the selection rule u_i are independent of the unobserved elements of the error process $(\epsilon_i(1), \epsilon_i(0))$.

The caveat, however, is that $p(x) \in (0,1)$. If $p(x) = 1$, then for every possible u_i, $D_i = 1$ - everyone gets treated. Therefore, there is no meaningful statement we can make about the $D_i = 0$ case. We can say similarly for $p(x) = 0$ case. In some textbooks, this is also known as **overlap** assumption - each unit in the defined population should have some chance of being treated and some chance of being in the control group. [2]

[2] If both (CIA) and overlap is satisfied, we say that have **strong ignorability** (Rosenbaum, Rubin 1983).

Under CIA, we can derive the treatment effect as follows

$$\begin{aligned}
E[Y_i|D_i=1,X_i=x]-E[Y_i|D_i=0,X_i=x] &= E[\mu(x,1)+\epsilon_i(1)|D_i=1,X_i=x]-E[\mu(x,0)+\epsilon_i(0)|D_i=0,X_i=x] \\
&= E[\mu(x,1)|D_i=1,X_i=x]+E[\epsilon_i(1)|D_i=1,X_i=x] \\
&- E[\mu(x,0)|D_i=0,X_i=x]-E[\epsilon_i(0)|D_i=0,X_i=x] \\
&= \mu(x,1)+E[\epsilon_i(1)|X_i=x]-\mu(x,0)-E[\epsilon_i(0)|X_i=x] \ (\because \text{CIA2}) \\
&= \mu(x,1)-\mu(x,0)
\end{aligned}$$

Note that

$$\begin{aligned}
E[Y_i(1)-Y_i(0)|X_i=x] &= E[Y_i(1)|X_i=x]-E[Y_i(0)|X_i=x] \\
&= (\Pr(D=1|X_i=x)\cdot E[Y_i(1)|D_i=1,X_i=x]+(1-\Pr(D=1|X_i=x))E[Y_i(1)|D_i=0,X_i=x]) \\
&- (\Pr(D=1|X_i=x)\cdot E[Y_i(0)|D_i=1,X_i=x]+(1-\Pr(D=1|X_i=x))E[Y_i(0)|D_i=0,X_i=x]) \\
&= E[Y_i(1)|D_i=1,X_i=x]-E[Y_i(0)|D_i=0,X_i=x] \ (\because \text{CIA}) \\
&= E[Y_i|D_i=1,X_i=x]-E[Y_i|D_i=0,X_i=x]
\end{aligned}$$

Thus, under (CIA) we can back out the ATE for $X_i = x$ using the observables. This assumption allows us to map $E[Y_i(1)|X_i=x]$ to $E[Y_i|D_i=1,X_i=x]$ and $E[Y_i(0)|X_i=x]$ to $E[Y_i|D_i=0,X_i=x]$, which is impossible without it.

This is called a **regression adjustment** method. The treatment effect for $X_i = x$ using a regression adjustment can be obtained by utilizing the fact that

$$\mu(x,1)=E[Y_i|D_i=1,X_i=x],\ \mu(x,0)=E[Y_i|D_i=0,X_i=x]$$

and regressing on the subsample of each treatment arm to get $\hat{\mu}(x,1)$ and $\hat{\mu}(x,0)$. Thus

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}(x,1)-\hat{\mu}(x,0))$$

8.5 Inverse Probability Weighting

We can obtain the average treatment effect using a different approach. This is an **inverse probability weighting**. The steps are as follows

1. Estimate the propensity score $p(X_i)$ by computing

$$\hat{p}_n(X_i)=\Pr(D_i=1|X_i=x)$$

2. Use the magic formula to estimate $E(TE(x)|x \in A)$: That is,

$$\frac{\sum_{D_i=1,x_i\in A} a_i Y_i}{\sum_{D_i=1,x_i\in A} a_i} - \frac{\sum_{D_i=0,x_i\in A} b_i Y_i}{\sum_{D_i=0,x_i\in A} b_i}$$

where $a_i = \frac{1}{\hat{p}_n(x_i)}, b_i = \frac{1}{1-\hat{p}_n(x_i)}$. The above can also be written as

$$\frac{\sum_{i=1,x_i\in A}^{N} \frac{D_i Y_i}{\hat{p}(X_i)}}{\sum_{i=1,x_i\in A}^{N} \frac{D_i}{\hat{p}(X_i)}} - \frac{\sum_{i=1,x_i\in A}^{N} \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}}{\sum_{i=1,x_i\in A}^{N} \frac{1-D_i}{1-\hat{p}(X_i)}}$$

The reason this works is as follows: Notice that

$$\begin{aligned} D_i Y_i &= D_i(D_i Y_i(1) + (1-D_i)Y_i(0)) = D_i Y_i(1) \\ (1-D_i)Y_i &= (1-D_i)Y_i(0) \end{aligned}$$

Thus, we have

$$\begin{aligned} E\left[\frac{D_i Y_i}{p(X_i)}\right] &= E\left[E\left[\frac{D_i Y_i(1)}{p(X_i)}|X_i\right]\right] \\ &= E\left[\frac{E[D_i|X_i]E[Y_i(1)|X_i]}{p(X_i)}\right] \\ &= E\left[\frac{p(X_i)E[Y_i(1)|X_i]}{p(X_i)}\right] = E[E[Y_i(1)|X_i]] = E[Y_i(1)] \end{aligned}$$

so we have $E\left[\frac{D_i Y_i}{p(X_i)}\right] = E[Y_i(1)], E\left[\frac{D_i Y_i}{p(X_i)}|X_i\right] = E[Y_i(1)|X_i]$. We can make the similar arguments for $Y_i(0)$ and get $E\left[\frac{(1-D_i)Y_i}{1-p(X_i)}\right] = E[Y_i(0)], E\left[\frac{(1-D_i)Y_i}{1-p(X_i)}|X_i\right] = E[Y_i(0)|X_i]$. Thus, the ATE for $X_i \in A$ can be written as

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{p(X_i)} - \frac{(1-D_i)Y_i}{1-p(X_i)}\right) \; \forall X_i \in A$$

In most cases, the propensity score is estimated. So we use the estimator involving the $\hat{p}$, which is what we have in the lecture notes. It is said that normalizing the weights to one in finite samples improves the mean squared error properties of the estimator (Imbens, Rubin (2019) - *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*)

8.6 Matching Estimation

The idea behind matching estimation is to impute the values for something we cannot get from the data: Namely, $Y_i(0)$ for those in $D_i = 1$ and $Y_i(1)$ for those in $D_i = 0$. In other words, you try to find a 'close' match for a particular unit i in a different treatment arm. When we say we find k-closest neighbors for unit i in $D_i = 0$, we find k individuals in $D_i = 1$ that has close traits (X_i) to individual i, or k individuals with the smallest values of $||x_j - x_i||$. Then, we construct a counterfactual $Y_i(1)$ by taking a (weighted) average over the Y_i's of the k individuals found in the other group. The treatment effect would than be $Y_i(1)$ -Y_i, where $Y_i(1)$ is calculated as in previous sentence.

Let's try with $k = 1$ - we find the one individual in the opposite treatment arm that has the similar value of X_i. Then, the average treatment effect can be written as

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i(1) - \hat{Y}_i(0))$$

where

$$\hat{Y}_i(1) = \begin{cases} Y_i & (D_i = 1) \\ \text{imputed value} & (D_i = 0) \end{cases}, \ \hat{Y}_i(0) = \begin{cases} \text{imputed value} & (D_i = 1) \\ Y_i & (D_i = 0) \end{cases}$$

There are some caveat in this approach

- The covariate X_i that is used to find the closest match in the other treatment arm should not affect the assignment of the treatment. In other words, if the assignment of treatment vs. control depends on some subset of X_i, you should not use that subset as a standard for finding the closest match.

- The overlap condition becomes critical. If it is not satisfied, i.e. for some X_i, $p(X_i) = 1$ or 0, then we are unable to find a closest match in the other treatment arm because for individuals with that covariate value, all of them are either in control or treatment group and not spread around.

- Who are the neighbors? The answer might depend on what distance measure we use. Moreover, how many of those in the treatment can be considered neighbors? There is a trade-off in the sense that using larger k would force us to put someone who is not 'close' as neighbors and using small k may cause difficulty in imputing counterfactual values.

8.7 Difference-in-Difference Estimation

This involves a specific framework where we can clearly define a 'before and after' - denoted as t_0 and t_1. No one is treated at t_0 but there is a subset of people ($G_i = 1$) that are treated at t_1. Those in $G_i = 0$ are never treated in either time period. If we define

$$Y_{i,t} = G_i Y_i(1,t) + (1 - G_i) Y_i(0,t) \ \ (t \in \{t_0, t_1\})$$

and impose

$$Y_i(G_i, t_0) = Y_i(t_0) \ \text{ for both } G_i = 1 \text{ and } G_i = 0$$

then we will be able to observe $(Y_{i,t_1}, Y_{i,t_0}, G_i.X_i)$ for every unit i. What we do not observe is $Y_i(1,t_1)$ for those in $G_i = 0$ and $Y_i(0,t_1)$ for those in $G_i = 1$. The analogue to the conditional independence assumption in this context is a **parallel trend assumption**, defined as

$$(Y_i(1,t_1) - Y_i(t_0), Y_i(0,t_1) - Y_i(t_0)) \perp\!\!\!\perp G_i | X_i$$

To see why they are equal, consider a setting where $D_i = G_i$ and write

$$\begin{aligned} Y_i = Y_{i,t_1} - Y_{i,t_0} &= D_i(\underbrace{Y_i(1,t_0) - Y_i(1,t_0)}_{Y_i(1)}) + (1 - D_i)(\underbrace{Y_i(0,t_1) - Y_i(0,t_0)}_{Y_i(0)}) \\ &= D_i Y_i(1) + (1 - D_i) Y_i(0) \end{aligned}$$

Therefore, we can rewrite the parallel trend assumption as

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i | X_i$$

which is the same as the conditional independence assumption.

To apply this in regression, we can write

$$Y_{it} = \beta_0(X_i) + \beta_1(X_i) \cdot 1(t = t_1) + \beta_2(X_i) \cdot G_i + \beta_3(X_i) \cdot G_i \cdot 1(t = t_1) + \epsilon_{it}$$

where X_i is a set of covariates, which can include a constant. In this context, the treatment effect for $X_i = x$ would be

$$\begin{aligned}
TE(x) &= E[Y_i(1) - Y_i(0)|X_i = x] \\
&= E[(Y_i(1,t_1) - Y_i(1,t_0)) - (Y_i(0,t_1) - Y_i(0,t_0))|X_i = x] \\
&= x \cdot E\{[(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2)] - [(\beta_0 + \beta_1) - (\beta_0)]|X_i = x\} \\
&= x \cdot E[(\beta_1 + \beta_3) - \beta_1|X_i = x] \\
&= \beta_3 x
\end{aligned}$$

So β_3 would be our parameter of interest.

One difficulty arises from testing the parallel trends assumption. Loosely speaking, what we can do is to select a time period $\tilde{t} < t_0$ and find out if the difference $y_i(t_0) - y_i(\tilde{t})$ is independent with G_i. If not independent, we say that there is an autonomous trend.

Chapter 9

Treatment Effects: Selection on Unobservables

9.1 Violation of Conditional Independence Assumptions

There are cases where the assignment, even if we condition on observables, are not random. So in the regressional framework of $y_i(d) = \mu(X_i, d) + \epsilon_i(d)$, the error terms is not independent of $D_i|X_i$. The conditional independence assumption can be broken because:

- Participants self-select based on expected benefit: Think about a job training program for plumbing. Then, maybe those who are more healthy and suffer less in terms of costs are likely to join. If health is not perfectly observed, we risk breaking the conditional independence assumption

- Participants may be selected, consciously or not, to join: Think of the clinical trial where participation is voluntary. In such case, individuals who are more risk-loving are more likely to join. In other words, participants and nonparticipants systematically differ on risk-averseness - something that is not usually observed.

- There may be equilibrium effects: A tuition subsidy program that intends to increase the number of people entering college may have a spillover effect by increasing supply of college graduates at the labor market, leading to a decrease in college premium. This may induce students to enter college less. If TE is the (rate of) college entrance, such equilibrium effect may be influential.

Mathematically, what happens is that when we calculate $E[Y_i|D_i = 1, x] - E[Y_i|D_i = 0, x]$, we end up with

$$\begin{aligned}E[Y_i|D_i=1,x]-E[Y_i|D_i=0.x]&=\mu(x.1)-\mu(x,0)+E[\epsilon_i(1)|D_i=1,x]-E[\epsilon_i(0)|D_i=0,x]\\&=TE+E[\epsilon_i(1)|D_i=1,x]-E[\epsilon_i(0)|D_i=0,x]\end{aligned}$$

The error term no longer can be erased from the equation since CIA assumption is not applicable. The difference between the error term is the **selection bias** (Also appears in Angrist, Pischke 2009). This also means that the error terms and the u_i in $D_i = 1(u_i < p(x_i))$ can covary. The result is that the treatment effect estimated from here can be inaccurate.

There are two possible solutions. Old method relies on **Heckman correction**. Recent focus is on IV to derive **marginal treatment effects** and **localized average treatment effect**.

9.2 A Brief Discussion of the Heckman Correction

Suppose that we are in the situation where we have a data generating process

$$Y_i=\max\{X_i\beta+\sigma\eta_i,0\},\eta_i\sim N(0,1)$$

So we only see Y_i if $X_i\beta+\sigma\eta_i>0$. We are able to observe D_i, specified as

$$D_i=\begin{cases}1 & \text{if } \eta>-\frac{X_i\beta}{\sigma}\\0 & \text{otherwise}\end{cases}$$

Then, for the observed sample, we are likely to have an η_i that is positively selected. So the idiosyncratic error is no longer random and we have a biased estimates, shown as (assuming X_i is observable)

$$\begin{aligned}&E[Y_i|Y_i>0]=X_i\beta+E[\sigma\eta_i|Y_i>0]\\\implies\ &E[\eta_i|Y_i>0]=E\left[\eta_i|\eta_i>-\frac{X_i\beta}{\sigma}\right]=\int_{-X_i\beta/\sigma}^{\infty}t\phi(t|\eta_i>-X_i\beta/\sigma)dt\\&=\int_{-X_i\beta/\sigma}^{\infty}t\frac{\phi(t)}{\Pr(\eta_i>-X_i\beta/\sigma)}dt=\frac{1}{1-\Phi(-X_i\beta/\sigma)}\int_{-X_i\beta/\sigma}^{\infty}t\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt\\&=\frac{1}{1-\Phi(-X_i\beta/\sigma)}\left[\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\right]_{-X_i\beta/\sigma}^{\infty}=\frac{\phi(-X_i\beta/\sigma)}{1-\Phi(-X_i\beta/\sigma)}\\&=\frac{\phi(X_i\beta/\sigma)}{\Phi(X_i\beta/\sigma)}=m(-X_i\beta/\sigma)\neq 0\end{aligned}$$

So the error has nonzero conditional mean. The ratio of the pdf over cdf $\frac{\phi(\cdot)}{\Phi(\cdot)}$ is the inverse Mill's ratio (IMR). Heckman's correction uses the IMR to correct for the bias. The challenge is to identify β/σ. We do this by

1. Use probit to regress D_i onto X_i. Then we can obtain $\widehat{\beta/\sigma}$.
2. Define $\hat{f}(x) = m(-X_i\widehat{\beta/\sigma})$. We include $\hat{f}(x)$ in the control variable and run

$$Y_i = X_i\beta + \gamma\hat{f}(x) + \epsilon_i$$

on the participant sample.

The problem is that we assume the normality of the error term and the linearity of the DGP, which is not always true. Thus, it is not always an ideal way to deal with selection on unobservables.

9.3 Instrumental Variables

In this setup, you assume that there exist variables Z_i that affect the treatment D_i but not the outcomes (at least on its own). It should satisfy

- **Relevancy**: It effects the propensity score $p(x,z) = \Pr(D_i = 1|X_i = x, Z_i = z)$
- **Validity**: Distribution of the counterfactual outcomes and u_i does not depend on $Z_i|X_i$. To put it in mathematical notation,

$$(Y_i(1), Y_i(0), u_i) \perp\!\!\!\perp Z_i|X_i$$

Comment 9.3.1 (Some remarks about the above condition)**.** *Note that u_i is tied together with the potential outcomes. This implies that u_i and the outcome can be correlated, which is allowed in a non-CIA setup. What matters is that the changes in Z_i not affect u_i*

Comment 9.3.2 (Violation of validity)**.** *Let Z_i change from z to $z' > z$ and that the propensity rises with z. Suppose that for a fixed $X_i = x$, $p(x,z) > u_i$. So $D_i = 1$ in this case. If we have $p(x,z') < u_i'$, then $D_i = 0$. In addition, we may have the change of z itself affecting the outcome directly with the violation of the validity condition.*

Comment 9.3.3 (So who are we identifying?). *If u is independent of Z_i given X_i, then by moving from $p(x,z)$ to some $p(x,z')$, we can find a group of individuals who do not get treated at $p(x,z)$ but do get treated at $p(x,z')$. We refer to them as* ***compliers****. If people were participating at $p(x,z)$ they participate at $p(x,z')$. We then have a group of* ***always takers****. Conversely, if people do not get treated at $p(x,z')$, they don't get treated at lower propensity scores - the* ***never takers****.*

We do not identify ***defiers*** *- they have $u_i < p(x,z)$ but $u_i' > p(x,z') > p(x,z)$. The only way this can happen is if u_i changes with z_i - a violation of validity. This is sometimes called as monotonicity condition or no two-way movement condition.*

- **Full support**: The support of $p(x,z)$ conditional on x extends to all of $[0,1]$. This implies that change in Z induces large variations in the propensity score.

So what concerns do we run into in terms of identification? We can identify $p(x,z)$ by calculating $\Pr(D_i = 1|X_i = x, Z_i = z)$. We can also write

$$\begin{aligned} E[Y_i|D_i = 1, X_i = x, Z_i = z] &= \mu(x,1) + E[\epsilon_i(1)|u_i < p(x,z), X_i = x, Z_i = z] \\ &= \mu(x,1) + E[\epsilon_i(1)|u_i < p(x,z), X_i = x] \ (\because \text{validity}) \end{aligned}$$

Define $K_1(p(x,z)) = E[\epsilon_i(1)|u_i < p(x,z), X_i = x]$ to be some unknown function of the conditional expectation of $\epsilon_i(1)$. We can also work similarly to define

$$E[Y_i|D_i = 0, X_i = x, Z_i = z] = \mu(x,0) + E[\epsilon_i(0)|u_i > p(x,z), X_i = x]$$

and $K_0(p(x,z)) = E[\epsilon_i(0)|u_i > p(x,z), X_i = x]$. Note that the left hand sides for both sets of equations can be identified by naively taking an average of Y_i's conditional on some the treated (untreated) for the group $X_i = x, Z_i = z$. The estimated result breaks down into

$$\mu(x,1) - \mu(x,0) + \underbrace{K_1(p(x,z)) - K_0(p(x,z))}_{\Delta(p(x,z))} = TE(x) + \Delta(p(x,z))$$

The $\Delta(p(x,z))$ is the control function that stands for the selection bias. By the relevance condition, the change in z affects the value of $p(x,z)$ for a fixed x. So changing z allows us to identify how $\Delta(p(x,z))$ *changes*. However, we do not know what the *exact value* of $\Delta(p(x,z))$ is for the initial value of z we started. Therefore, the true parameter of interest, which is TE, cannot be uncovered.

So how can we progress on? Note that the K_d functions are conditional expectations of $\epsilon_i(d)$ on $X_i = x$ and selection rule $u_i(<,>)p(x,z)$. In particular, we can get that

$$\begin{aligned} K_1(1) &= E[\epsilon_i(1)|u_i < 1, X_i = x] \\ &= E[\epsilon_i(1)|X_i = x] \text{ } (\because \text{Everyone is treated}) \\ &= 0 \\ K_0(0) &= E[\epsilon_i(1)|u_i > 0, X_i = x] \\ &= E[\epsilon_i(1)|X_i = x] \text{ } (\because \text{No one is treated}) \\ &= 0 \end{aligned}$$

Heckman and Vytlacil show that if we are given a continuous instrument z, we can do a much better job of identifying the treatment effect.

9.4 Marginal Treatment Effects

The **marginal treatment effect** at $p(x,z) = p$ is defined as the treatment effect on individuals whose $u_i = p(x,z)$. We can write

$$MTE(p) = E[Y_i(1) - Y_i(0)|u_i = p]$$

The conditional expectation above is not directly obtainable from the data. Heckman and Vytlacil show that

$$MTE(p) = \frac{\partial E[Y_i|p(x,z) = p]}{\partial p}$$

which is obtainable from the data. This is done by

1. Estimate $p(x,z) = \Pr(D_i = 1|X_i = x.Z_i = z)$

2. Regress Y_i on the estimated $p(x,z)$ in a flexible setting - preferably not just linearly but with some nonlinearities

3. Take a derivative with respect to p. (or local linear estimator)

4. For treatment effects, evaluate the $E[Y_i|p(x,z),x]$ at $p(x,z) = 1$ and $p(x,z) = 0$ and identify the difference. (You can obtain $E[Y_i|\cdot]$ by getting the predicted values).

Intuitively, what is going on with MTE is as follows: By changing p slightly by dp, we are able to identify the marginal compliers who move from not being treated to being treated. We are finding out how their outcome changes as they move from non-participation to participation into the treatment.

To see why the above result holds, we define

$$G(p) = E[Y_i \cdot 1(p(x,z) = p)]$$

Since $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0) = 1(u_i < p(x,z))Y_i(1) + 1(u_i > p(x,z))Y_i(0)$, we can rewrite the above as

$$\begin{aligned} G(p) &= E[Y_i(1) \cdot 1(u_i < p(x,z)) \cdot 1(p(x,z) = p) + Y_i(0) \cdot 1(u_i > p(x,z)) \cdot 1(p(x,z) = p)] \\ &= E[Y_i(1) \cdot 1(u_i < p) \cdot 1(p(x,z) = p)] + E[Y_i(0) \cdot 1(u_i > p) \cdot 1(p(x,z) = p)] \\ &= G_1(p) + G_0(p) \end{aligned}$$

By the validity condition and the fact that $u_i \sim U[0,1]$ (and thus $f(u_i) = 1$), we can write

$$\begin{aligned} E[Y_i(1) \cdot 1(u_i < p) \cdot 1(p(x,z) = p)] &= E[Y_i(1) \cdot 1(u_i < p)] \Pr(p(x,z) = p) \\ &= \int_0^p E[Y_i(1)|u = t]dt \Pr(p(x,z) = p) \end{aligned}$$

And similarly,

$$E[Y_i(0) \cdot 1(u_i > p) \cdot 1(p(x,z) = p)] = \int_p^1 E[Y_i(0)|u = t]dt \Pr(p(x,z) = p)$$

So

$$E[Y_i \cdot 1(p(x,z) = p)] = G(p) = \left(\int_0^p E[Y_i(1)|u = t]dt + \int_p^1 E[Y_i(0)|u = t]dt\right) \Pr(p(x,z) = p)$$

with the fact that $E[Y_i \cdot 1(p(x,z) = p)] = E[Y_i|p(x,z) = p] \cdot \Pr(p(x,z) = p)$ this implies that

$$E[Y_i|p(x,z) = p] = \frac{G(p)}{\Pr(p(x,z) = p)} = \int_0^p E[Y_i(1)|u = t]dt + \int_p^1 E[Y_i(0)|u = t]dt$$

Then, by Leibniz's integral rule

$$\frac{\partial E[Y_i|p(x,z) = p]}{\partial p} = E[Y_i(1)|u = p] - E[Y_i(0)|u = p] = MTE(p)$$

Hence, the treatment effect can be recovered by

$$TE = \int_0^1 MTE(p)dp = E[Y_i|p(x,z)=1] - E[Y_i|p(x,z)=0]$$

9.5 Some Large Caveats for MTE

For the above methods to work, we need the Z_i instruments to satisfy

- Z belongs in the treatment equation (relevancy): $D_i = 1(u_i < p(X_i, Z_i))$
- Z does not belong in the outcome equation (exclusivity): $Y_i(d) = \mu(x,d) + \epsilon_i(d)$
- In other words, we get $(\epsilon_i(1), \epsilon_i(0), u_i) \perp\!\!\!\perp Z_i|X_i$
- Continuity: $p(x,z)$ changes w.r.t z in a continuous way (This is because we need to be able to take derivatives)

For the range of $p(x,z)$ available, the above condition allows us to estimate the marginal treatment effect. For any work you see that uses marginal treatment effect, there includes a graph that maps marginal treatment effect on the vertical axis and the 'resistance' parameter u_i on the horizontal axis.

With this framework, we can also test if conditional independence assumption holds. Recall that conditional independence assumption is satisfied when

$$(\epsilon_i(1), \epsilon_i(0)) \perp\!\!\!\perp u_i|X_i$$

In cases where this is true, then the outcomes are independent of u_i conditional on X_i. Thus,

$$MTE(x,p) = E[Y_i(1) - Y_i(0)|X_i = x, u_i = p] = E[Y_i(1) - Y_i(0)|X_i = x]$$

Since $MTE(x,p) = \frac{\partial E[Y_i|X_i=x, p(x,z)=p]}{\partial p}$, the above condition implies that $E[Y_i|X_i = x, p(x,z) = p]$ is linear in p. Thus, it is highly recommended to put polynomial terms of $p^k, k = 1,2,3,...$ when you estimate marginal treatment effects. Then, test to find whether the nonlinear terms have coefficient zero.

So what do we gain by testing the form of $MTE(x,p)$? How it varies with p may inform how well the treatment is designed. Ideally, we want the treatment to be applied to those who are more likely to be benefitted by it. If p is higher for those with larger benefits, then the design is well-targeted.

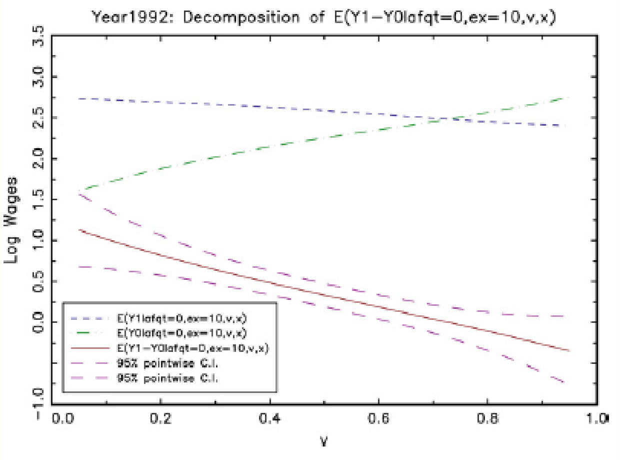
Definition 9.5.1 (Policy Relevant Treatment Effect)**.** Policy relevant treatment effect is the mean effect of going from a baseline policy to a new policy per net person shifted (Heckman, Vytlacil 2001, Carneiro, Heckman, Vytlacil 2010). This can be written as

$$\frac{E[Y_i|\text{New Policy}] - E[Y_i|\text{Old Policy}]}{E[D_i|\text{New Policy}] - E[D_i|\text{Old Policy}]}$$

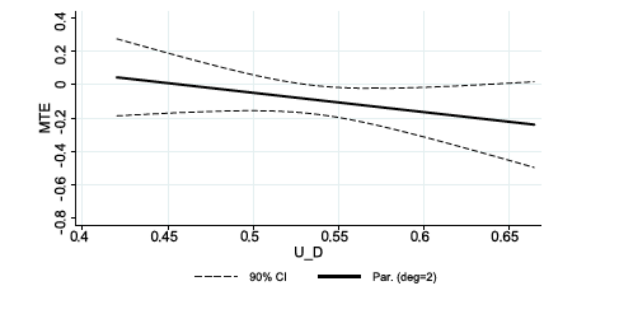
$$= \int_0^1 MTE(u)\omega(x,u)du \quad (\omega(x,u) = F_{P_{old}|X}(u|x) - F_{P_{new}|X}(u|x))$$

Example 9.5.1 (Carneiro & Lee, 2019)**.** *The paper estimates the impact of attending college on log wage distributions. The paper finds that individuals more likely to attend college (and have low resistance parameters) are more likely to have higher college wage ($Y_i(1)$) over high school wage ($Y_i(0)$). The opposite holds true for people with high school degree. They have a MTE figure (at figure 3) that maps MTE as a function of resistance parameters.*

Example 9.5.2 (Johnson & Taylor, 2019)**.** *The paper shows that causal impact of migration decreases longevity. This is even with the consideration that migrants are more likely to be educated and have higher baseline earnings compared to non-migrants. They use the MTE to (with railcar traffic at the town of origin as one of their IVs). They document that those who have lower latent ability (high U_d) suffer more from migrating out, reflected in the downward sloping MTE*



(a) Carneiro, Lee (2009)



(b) Johnson, Taylor (2019)

9.6 Local Average Treatment Effect

The marginal treatment effect looks at 'marginal' compliers: It is interested in the small group of individuals who changes treatment status from $D_i = 0$ to $D_i = 1$ as $p(x,z)$ changes slightly due to Z_i. Instead of changing the propensity score by a tiny bit (thanks to the continuity of the instrument Z_i), we can look at finite differences - like Z_i having just two possible values. Then we are going to be looking at a larger group of compliers than in the marginal treatment effect context. The treatment effect for this larger group of compliers is called **local average treatment effect**.

We will setup the framework as follows:

- Z_i will be a binary instrumental variable. This of this as an eligibility rule.
- $D_i(z)$ can be characterized as $D_i(z) = 1(u_i < p(x,z))$. Note that as z rises, so will $p(x,z)$. This is the relevance condition
- Z_i itself has no bearing, at least directly, on the outcome. So $Y_i(d) = \mu(x,d) + \epsilon_i(d)$. So we still have $(\epsilon_i(1), \epsilon_i(0), u_i) \perp\!\!\!\perp Z_i | X_i$.

The formal way to define local average treatment effect is as follows

$$LATE(x) = E[Y_i(1) - Y_i(0) | p(x,z) < u_i < p(x,z').X_i = x]$$

We can identify LATE by using the following (I skip $X_i = x$)

$$\begin{aligned}
E[Y_i|Z_i = z'] - E[Y_i|Z_i = z] &= E[1(u_i < p(z'))Y_i(1) + 1(u_i > p(z'))Y_i(0)|Z_i = z'] \\
&- E[1(u_i < p(z))Y_i(1) + 1(u_i > p(z))Y_i(0)|Z_i = z] \\
&= E[1(u_i < p(z'))Y_i(1) + 1(u_i > p(z'))Y_i(0)] \\
&- E[1(u_i < p(z))Y_i(1) + 1(u_i > p(z))Y_i(0)] \\
&= E[(1(u_i < p(z')) - 1(u_i < p(z))Y_i(1) - (1(u_i > p(z')) - 1(u_i > p(z))Y_i(0)] \\
&= E[(1(u_i < p(z')) - 1(u_i < p(z))(Y_i(1) - Y_i(0))] \\
&= \Pr[1(u_i < p(z')) - 1(u_i < p(z)) = 1]E[Y_i(1) - Y_i(0)|1(u_i < p(z')) - 1(u_i < p(z)) = 1]
\end{aligned}$$

Note that we do not have to worry about always takers and never takers as their $1(u_i < p(z')) - 1(u_i < p(z)) = 0$. Since we rule out defiers using the monotonicity assumption, we need not worry about $1(u_i < p(z')) - 1(u_i < p(z)) = -1$. Moreover, $1(u_i < p(z')) - 1(u_i < p(z)) = 1$ holds iff $p(z) < u_i < p(z')$. So to continue,

$$\begin{aligned}
E[Y_i|Z_i = z'] - E[Y_i|Z_i = z] &= \Pr[1(u_i < p(z')) - 1(u_i < p(z)) = 1]E[Y_i(1) - Y_i(0)|1(u_i < p(z')) - 1(u_i < p(z)) = 1] \\
&= \Pr[p(z) < u_i < p(z')]E[Y_i(1) - Y_i(0)|p(z) < u_i < p(z')]
\end{aligned}$$

Therefore, we are able to back out the definition of the LATE and can identify them as

$$LATE(x,z,z') = \frac{E[Y_i|Z_i = z'] - E[Y_i|Z_i = z]}{\Pr(p(z) < u_i < p(z'))} = \frac{E[Y_i|Z_i = z'] - E[Y_i|Z_i = z]}{p(z') - p(z)}$$

Or in terms of the propensity score (and by bringing $X_i = x$ back in)

$$LATE(x, p(z), p(z')) = \frac{E[Y_i|p = p(x.z')] - E[Y_i|p = p(x,z)]}{p(x,z') - p(x,z)}$$

We can go further: Estimate propensity scores with

$$p(x,z) = E[D_i|X_i = x, Z_i = z]$$

and get

$$LATE(x,z,z') = \frac{E[Y_i|x,z'] - E[Y_i|x,z]}{E[D_i|x,z'] - E[D_i|x,z]}$$

To obtain this from the regression, we follow these steps:

1. Regress D on Z and other covariates X to get $\hat{D} = \hat{p}(x,z)$
2. Regress Y on other covariates X and $\hat{D}$.

The LATE estimator can then be obtained here is called the Wald estimator.

9.7 Some Words on Monotonicity Condition

When we wrote $D_i = 1(u_i < p(x,z))$ and assume that $P(x,z)$ rises with z for fixed x, monotonicity condition implies that D_i can only increase. We are ruling out a case where D_i can change from 1 to 0. We also throw away information on the always takers and never takers as well. Both LATE and MTE can only tells us about compliers while throwing away defiers. This condition, along with the validity condition, cannot be tested empirically. Therefore, the applicability of this condition should be argued using intuition or external facts.

Chapter 10

Treatment Effects: Regression Discontinuity

10.1 Why do We Want Regression Discontinuity?

Regression discontinuity (commonly abbreviated to RD or RDD) approach actually started from Psychology in 1960s and was brought into the mainstream of applied economics very recently. Whenever you learn applied microeconomics class in your second year, you are bound to read or present about a paper that uses this research design.

As its name suggests, this approach exploits the discontinuities in the treatment assignment. So the key difference between regression discontinuity and previous frameworks is that RDD does not require the overlap condition in the sense that $p(X_i) \in (0, 1)$. In fact, one of the design has $p(X_i) = 1$ for $X_i \geq c$ and 0 otherwise. Another thing that RDD should satisfy is that for every other covariates, the distribution should remain continuous even through the discontinuity.

This is usually presented visually through graphs - one with the covariate that determines assignment to the treatment, usually called the forcing (running) variable W_i and the treatment status D_i, the other with the W_i and all other covariates X_i. For instance, if we are in the public finance context and study the effect of a particular welfare program with a means-tested benefit assignment mechanism, we could use the means-tested criteria as our W_i. If we are to study the effect of a specific educational program offered to those below a particular test score, that score could be the W_i. Here are some works that deal use RDD.

Example 10.1.1 (Lee, 2007)**.** *This paper focuses on the elections won and lost by a very narrow margin. The logic is that in such case, winning and losing is effectively randomized. The running variable is the Democrat vote share − Republican vote share, or a margin of victory/loss. One of the findings is that Democrats who barely win an election are much more likely to run for office and succeed in the next election, compared to Democrats who barely lose. In general this paper finds positive relationship between margin of victory and electoral outcomes, both in terms of vote shares and wins by individuals and parties.*

Example 10.1.2 (Howell, 2017)**.** *This paper assess the impact of R&D subsidies from the government to new ventures. The rank on the application to the R&D is used as a running variable, (and this is a sharp RDD and a discrete running variable*[a]*). The paper finds that an early-stage award approximately doubles the probability that a firm receives subsequent venture capital (summarized in the figure below) and has large, positive impacts on patenting and revenue. These effects are stronger for more financially constrained firms.*

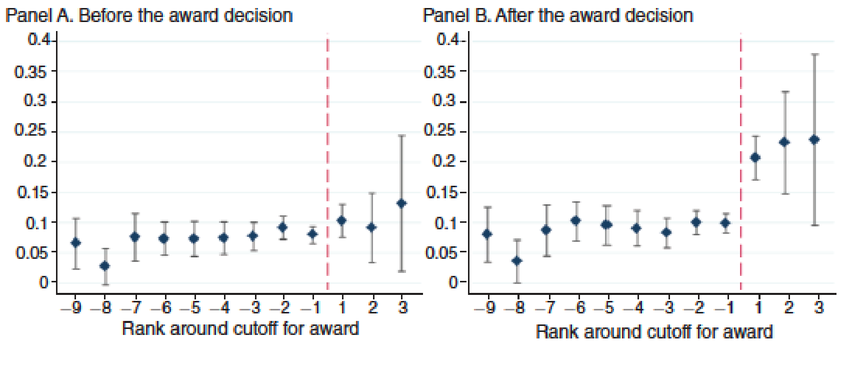


Figure 3. Probability of Venture Capital before and after Grant by Rank

Figure 10.1: Reception of venture capital, before and after

Example 10.1.3 (Ost, Pan & Weber, 2018)**.** *This paper estimates the return to college using admin data on college enrollment and earnings from Ohio. The running variable here is a GPA, in that this paper uses the fact that colleges dismiss low-performing students on certain GPAs. RDD is used to measure effects of college on earnings. This paper finds that dismissal leads to a short-run increase in earnings and tuition savings, which are wiped out by gains from college graduates 8 years after. They also carry out the test that the result is not driven by manipulation.*

[a] Because of the discreteness, McCrary test is infeasible.

10.2 Framework

We will start with a **sharp RDD**. You can think of this case as a case where the assignment into the treatment is determined by

$$D_i = 1(W_i \geq c(X_i))$$

where $c(X_i)$ can be a known function of X_i or just a constant. We also keep the potential outcome framework from before, but we will slightly change this to

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + D_i(Y_i(1) - Y_i(0)) \\ &= Y_i(0) + 1(W_i \geq c(X_i)) \cdot (Y_i(1) - Y_i(0)) \end{aligned}$$

We also keep the regressional framework for the potential outcomes, $Y_i(d) = \mu(X_i, W_i, d) + \epsilon_i(d)$. So we have

$$E[Y_i(d)|X_i, W_i] = \mu(X_i.W_i.d)$$

One key assumption that we will carry through this section is the continuity of the conditional expectation on potential outcomes at $W_i = c(X_i)$

Assumption 10.2.1 (Continuity of $\mu(X_i, W_i, d)$ at $W_i = c$)**.** *If W_i has a known cutoff at c, we assume that $E[Y_i(1)|X_i, W_i]$ and $E[Y_i(0)|X_i, W_i]$ are continuous around $W_i = c$. Put if differently,*

$$E[Y_i(d)|X_i, W_i = c^+] = E[Y_i(d)|X_i, W_i = c^-] \textit{ for each } d \in \{0, 1\}$$

The above condition also implies that $(\epsilon_i(1), \epsilon_i(0))$ also has a continuous distribution around $W_i = c(X_i)$. If this condition is not satisfied, this implies that there could be more than the W_i variable that jumps around the cutoff, which is what you do not want in an RDD regression.

There is another version of the RDD which is more general than the sharp RDD. A **fuzzy RDD** setup exploits the jump in the probability of being treated at $W_i = c$. This jump is not necessarily from 0 to 1. Formally, we say we look for discontinuities of

$$\Pr(D_i|X_i, W_i)$$

at $W_i = c(X_i)$.

10.3 Identification in RDD

We start with a sharp RDD case. Assume that $c(X_i)$ is a constant c. We can write

$$E[Y_i|X_i, W_i] = E[Y_i(0)|X_i, W_i] + E[(Y_i(1) - Y_i(0))1(W_i \geq c(X_i))|X_i, W_i]$$

And we can separate the above into two. One with $W_i = c^+$ and the other with $W_i = c^-$.

$$\begin{aligned} E[Y_i|X_i, W_i = c^+] &= E[Y_i(0)|X_i, W_i = c^+] + E[(Y_i(1) - Y_i(0))|X_i, W_i = c^+] \\ E[Y_i|X_i, W_i = c^-] &= E[Y_i(0)|X_i, W_i = c^-] \end{aligned}$$

This is because $1(W_i \geq c(X_i)) = 1$ if $W_i = c^+$ and 0 otherwise. Therefore,

$$\begin{aligned} E[Y_i|X_i, W_i = c^+] - E[Y_i|X_i, W_i = c^-] &= E[Y_i(0)|X_i, W_i = c^+] - E[Y_i(0)|X_i, W_i = c^-] \\ &\quad + E[(Y_i(1) - Y_i(0))|X_i, W_i = c^+] \\ &= E[(Y_i(1) - Y_i(0))|X_i, W_i = c^+] \end{aligned}$$

where the first line on the right hand side vanishes due to the continuity assumptions. Reusing the continuity argument that $E[Y_i(d)|X_i, W_i = c^+] = E[Y_i(d)|X_i, W_i = c^-]$ for each $d \in \{0, 1\}$, we can write

$$E[(Y_i(1) - Y_i(0))|X_i, W_i = c^+] = E[(Y_i(1) - Y_i(0))|X_i, W_i = c]$$

Therefore, we have backed out

$$E[(Y_i(1) - Y_i(0))|X_i, W_i = c] = E[Y_i|X_i, W_i = c^+] - E[Y_i|X_i, W_i = c^-]$$

which is the average treatment effect at $W_i = c$ and a given X_i. This is similar to the LATE in the sense if we let $Z_i = 1(W_i \geq c)$, we can get back what is effectively equivalent to the LATE estimator (with the difference in the probability of treatment exactly at 1). Also, we are looking at the treatment effect on the compliers, albeit a very narrow group of people.

For the fuzzy RDD, we can write as follows

$$E[Y_i|X_i, W_i] = E[Y_i(0)|X_i, W_i] + E[(Y_i(1) - Y_i(0)) \Pr(D_i = 1|X_i, W_i)|X_i, W_i]$$

So we can do the similar approach as in sharp RDD and get

$$E[Y_i|X_i = x, W_i = c^+] = E[Y_i(0)|X_i = x, W_i = c^+] + \Pr(D_i = 1|X_i = x, W_i = c^+)E[Y_i(1) - Y_i(0)|X_i = x, W_i = c^+]$$
$$E[Y_i|X_i = x, W_i = c^-] = E[Y_i(0)|X_i = x, W_i = c^-] + \Pr(D_i = 1|X_i = x, W_i = c^-)E[Y_i(1) - Y_i(0)|X_i = x, W_i = c^-]$$

So using the continuity assumption, the difference in the left hand side can be written as

$$\begin{aligned} E[Y_i|X_i = x, W_i = c^+] - E[Y_i|X_i = x, W_i = c^-] = [&\Pr(D_i = 1|X_i = x, W_i = c^+) - \Pr(D_i = 1|X_i = x, W_i = c^-)] \\ &\times E[Y_i(1) - Y_i(0)|X_i, W_i = c] \end{aligned}$$

Thus, we also identify an average treatment effect at $W_i = c$ and $X_i = x$, which is now divided by something that is not necessarily 1.

$$E[Y_i(1) - Y_i(0)|X_i = x, W_i = c] = \frac{E[Y_i|X_i = x, W_i = c^+] - E[Y_i|X_i = x, W_i = c^-]}{\Pr(D_i = 1|X_i = x, W_i = c^+) - \Pr(D_i = 1|X_i = x, W_i = c^-)}$$

10.4 Implementation of RDD

So what can we do in terms of implementing the RDD in practice? One way to implement RDD in practice is to make use of nonparametric regression. The bandwidth, however, should be bounded at c. If you are doing a nonparametric regression from the left of the bandwidth, you should apply this on $[c - \epsilon, c)$. For the right hand side, $[c, c + \epsilon]$. If the domain reaches beyond c from either side, there is a huge risk of misfitting the data. While you can do local constant regression in theory, it comes at a huge cost. With this, you can only pin down the level of Y_i from either side. You lose out the rate in which outcome changes within both sides of the boundary. Therefore, if carrying out nonparametric regression, you should do at least a local linear regression.

Since we are doing a nonparametric regression, we also run into similar problems as in the typical nonparametric regressions. If you have many other variables to consider, you risk running into the curse of dimensionality. There is also an issue of selecting an optimal bandwidth. We still minimize the AMISE, but at $W_i = c$. $MSE(h)$ is defined as

$$E[(\hat{\mu}(x, c, 0) - \mu(x, c, 0))^2 + (\hat{\mu}(x, c, 1) - \mu(x, c, 1))^2]$$

Calonico, Cattaneo, and Tituinik (2014) suggests the use of local quadratic estimator with the optimal bandwidth selected by minimizing above $MSE(h)$ which is typically known as IK bandwidth (Imbens, Kalyanaraman, 2012). Also note that the same trade-off between variance and bias is still in play.

There is also a parametric way to implement this as well. One way is to run a separate regression for both left and right of the cutoff. That is, run

$$\begin{aligned} Y_i &= X_i\beta^- + (c - w_i)X_i\gamma^- + \epsilon_i^- \text{ for } w_i < c \\ Y_i &= X_i\beta^+ + (w_i - c)X_i\gamma^+ + \epsilon_i^+ \text{ for } w_i \geq c \end{aligned}$$

This setup allows us to fit a different trend on the left and the right of the cutoff c. There is also a way in which we make use of both sides of the cutoff. Specifically, we run

$$Y_i = X_i\beta + X_i \cdot 1(W_i \geq c)\gamma + X_i \cdot (c - W_i)\delta + X_i \cdot (c - W_i) \cdot 1(W_i \geq c)\mu + \epsilon_i$$

Here, what happens is that

- $W_i \geq c$: $Y_i = X_i(\beta + \gamma) + X_i \cdot (c - W_i)(\delta + \mu)$
- $W_i < c$: $Y_i = X_i(\beta) + X_i \cdot (c - W_i)(\delta)$

To make this simple, suppose that X_i is just a vector of 1's. Then, on the (Y_i, W_i) plane, the jump of the outcome at $W_i = c$ would be captured by the γ parameter. Using δ, μ, we can allow the slopes to differ on either side of the equation.

There are some graphical tests that should be carried out. First, we plot Y_i as a function of W_i for a given X_i and see if there is a mean shift. Also, we plot the propensity score $\Pr(D_i = 1|X_i, W_i)$ and look for a shift. One thing that should not shift is the distribution of X_i's. So plot X_i's as function of W_i and confirm that there is no shift. Another thing to worry regarding W_i is the possibility of manipulating or gaming. Maybe it might be possible to get into the treatment when you should not be treated (e.g. Ost et al. (2018) checks for this). Therefore, use a McCrary test to confirm that the distribution of W_i do not change at $W_i = c$.

Last but not least, what if we do not know the exact cutoff? That is actually not much to worry about, since we can figure out what they might be based on the distribution of W_i. Chay, McEwan, and Urquiola (2005) conducts and RDD in the context where the exact cutoffs are unknown using visual evidence. Also, since we lose a lot of data in selecting the optimal bandwidth around the cutoff, RDD is usually poor in terms of external validity.

Chapter 11

Introduction to Machine Learning

11.1 Machine Learning

Machine learning in the context of econometrics is interested in model selection for prediction - finding the parsimonious model in the context of a very large dataset. It ranges from regularized models such as LASSO, Ridge regression, and elastic nets and methods that are much more involved - random forests and neural networks. All of them differ in how they fine-tune the model.

11.2 Model Selection

Often, we are interested in controlling a large set of covariates (e.g. propensity score). We can generalize this to the structure where structure refers to specification as well as parameter values. In determining the structure, we face a trade-off: We are tempted to increase the number of covariates since it allows us to fit the data better. However, in doing so, we risk overfitting - the out-sample prediction can perform poorly in terms of variance. Therefore, some method to penalize complexity is required.

We start with a set $\mathcal{M}$ of candidate models. Then for a model $M \in \mathcal{M}$, we define a penalty parameter $p(M)$ that increases in complexity. Two classical information criteria include

- Akaike's Information Criterion: Choose M satisfying

$$\max_{M} \left(\max_{\theta} \sum_{i=1}^{n} \log l_i(\theta; M) - p(M) \right)$$

- Bayesian Information Criterion: Choose M satisfying

$$\max_M \left(\max_\theta \sum_{i=1}^n \log l_i(\theta; M) - p(M) \frac{\log n}{2} \right)$$

The subtle difference between the two is that BIC tends to be harsher for complexities. This can be seen since for $n > 8$, the penalty term is larger. Moreover, the derivation of BIC is based on a Bayesian probabilistic framework - If a selection of candidate models includes a true model for the dataset, then the probability that BIC selects the correct model approaches to 1 as training sample size increases (From Hastie, Tibshirani, Friedman ”The Element of Statistical Learning”, 2016). Therefore, BIC performs better for selecting the model within the given sample. For out-sample performance, AIC performs better. Both of them, however, can be impractical to calculate when we work with large models. This can be common especially if we allow for multiple levels of interaction between the covariates. In other words, if p is a number of potential covariates, and n is the sample size, we run into the case where $p >> n$.

One way out is to penalize complexities with the convex norm penalties. Define

$$\|\theta\|_m = \sqrt[m]{\sum_{k=1}^p |\theta_k|^m}$$

and for $m = 0$, we have $\|\theta\|_p = p$, penalties for the two information criteria above. Depending on the selection of m, we get different estimators that incorporates different shrinkage mechanisms - Ridge for $m = 2$ and LASSO for $m = 1$.

The general idea of shrinkage is as follows: We know that OLS is BLUE - It has the least variance among the linear, unbiased estimators. But what if we are willing to relax the ’unbiased’ condition? Then can we find some estimators that have lower variance and even a lower mean squared error? James, Stein (1961) shows that it is possible and generalizable.

Ridge and LASSO builds upon this idea, but uses different regularization methods. Consider the setup $Y_i = X_i\theta_0 + \epsilon_i$. Ridge regression minimizes

$$\sum_{i=1}^n (Y_i - X_i\theta_0)^2 + \lambda \sum_{k=1}^p \theta_k^2 \quad \left(\sum_{k=1}^p \theta_k^2 = \|\theta\|_2^2 \right)$$

and yields

$$\hat{\theta} = (X'X + \lambda I_p)^{-1} X'y$$

Note that as λ increases, the $(X'X + \lambda I_p)^{-1}$ term decreases, which drives the coefficients close to 0 and reduce overfitting. With LASSO, the minimization is

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\theta_0)^2 + \lambda\sum_{k=1}^{p}|\theta_k|$$

and the first order condition can be written as

$$\frac{2}{n}\sum_{i} X_{ik}(y_i - X_i\theta) = \lambda s_k, \;\; s_k = \begin{cases} 1 & (\hat{\theta}_k > 0) \\ -1 & (\hat{\theta}_k < 0) \\ [-1,1] & (\hat{\theta}_k = 0) \end{cases}$$

Where λs_k is not necessarily 0, consistent with the idea that we are open to allowing for some bias in order to find an estimator that reduces MSE.

So what is the difference between LASSO and Ridge? Ridge regression drives your coefficients towards 0 but not equal to it. So when it comes to model selection in terms of ruling out irrelevant variables, LASSO does better. This idea is captured by the figure below. As far as computation is concerned, Ridge has a quicker computation speed and works relatively better with multicollinearity. So the choice of the estimation method depends on your goal of minimizing MSE as well as computation power (but with modern computational power, LASSO is preferred to Ridge).

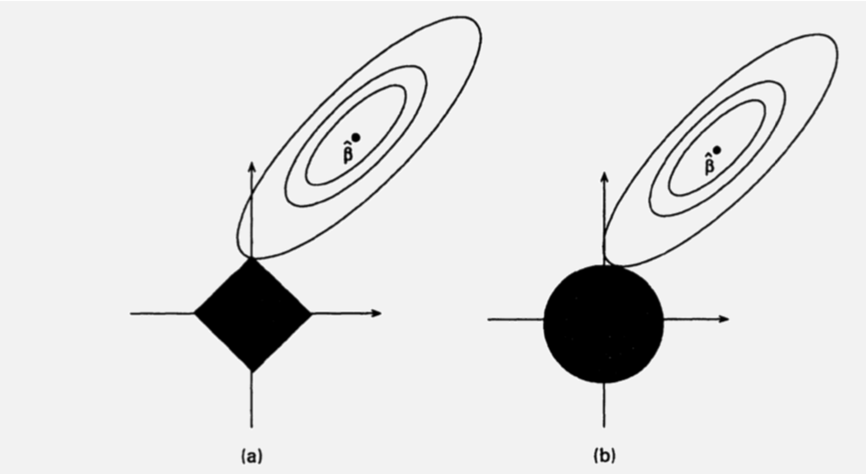


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

Figure 11.1: Ridge vs LASSO

Comment 11.2.1 (Can you do both at once?). *With **Elastic Net Regression**, you can. The idea is to use a penalty term that is a convex combination of both l_1 and l_2 penalty term. Specifically,*

$$\sum_{i=1}^{n}(Y_i - X_i\theta_0)^2 + \lambda \sum_{k=1}^{n}\left[\alpha\|\theta\|_1^2 + (1-\alpha)\|\theta\|_2^2\right]$$

If $\lambda = 0$ we have an OLS. For $\lambda \neq 0$, if $\alpha = 1$, we have LASSO, and if $\alpha = 0$ we get Ridge.

Comment 11.2.2 (Least Angle Regression (LARS)). *The algorithm is as follows*

1. *Start with a constant model: For all non-constant covariates, coefficients are 0 and $\lambda = \infty$.*
2. *Then, find the covariate that has the largest covariance with the residual.*
3. *Change the coefficient for that variable in the same direction as its covariance with the residual ($Y_i - \hat{Y}_i$) until you find a different covariate that has as much covariance with the residuals. λ decrease and θ is updated along the way.*
4. *Change the coefficient for the two variables until you find a third variable with the same covariance with the residual.*
5. *Stop when all predictors are in the model*

When the covariate has the largest covariance with the residual, its angle is smallest. Hence the name LARS. The following figure captures how it works.

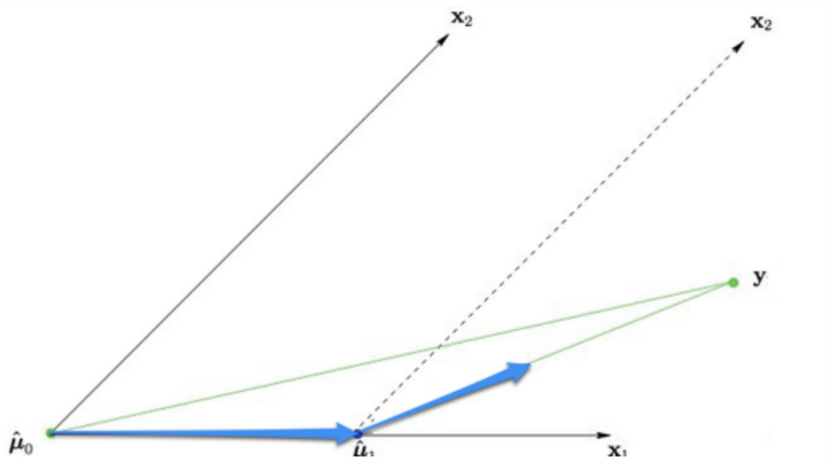


Figure 11.2: LARS with two variables

So what is the optimal λ? It depends on your goal. If your goal is to minimize the prediction error, you can use the following cross-validation methods to come up with one

- Leave-one-out: Leave one observation out at a time
- Training vs Test sample: You split the sample into two, build a model given some λ and gain prediction error from the test sample. Then, select λ minimizing the prediction error there.
- K-fold cross validation: For $K-1$ blocks, estimate the model given some λ and compute the error on the remaining block. Get this for K blocks and compute the average error. Select λ minimizing the error.

11.3 Some Primer on the Consistency of LASSO

We want to identify what the true model is in a setting with potentially an infinite number of covariates. To do that, we first set a sparsity assumption, where the number of nonzero coefficients is not too large. As a thought experiment, we increase the number of regressors as $n \to \infty$ in order to approximate the true model. To satisfy the assumption, we require that the number of nonzero regressors to rise at a slower rate than n. Specifically, $s_n = o\left(\sqrt{\frac{n}{\log p}}\right)$.

If we let $\lambda_n = O\left(\sqrt{\frac{\log p}{n}}\right)$ and have $s_n = o\left(\sqrt{\frac{n}{\log p}}\right)$, then we can show that the prediction error goes to 0 and the LASSO estimator is consistent. Or

$$\frac{1}{n}\sum_i (X_i\hat{\theta}_n - X_i\hat{\theta}_{0n})^2 = O_p(\lambda_n^2 s_n)$$

and

$$\|\hat{\theta}_n - \hat{\theta}_{0n}\|_1 = o_p(\lambda_n s_n) = o_p(1)$$

This comes at the cost of a slightly slower rate of convergence, which is tolerable as long as it is faster than $n^{-1/4}$.

11.4 Random Forests

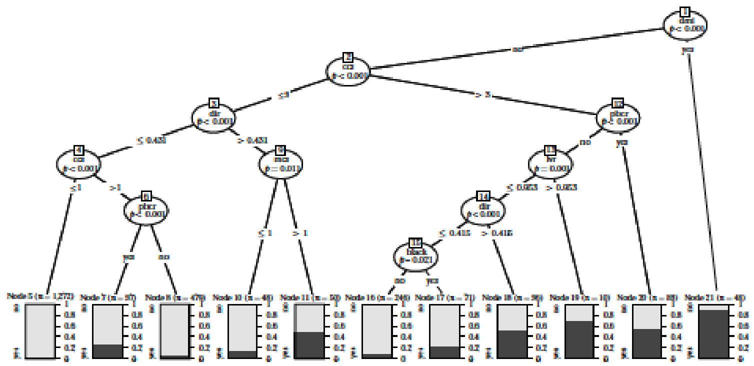
Random forests use a series of decision trees - a stack of two-way splits with several number of depths - to solve a classification problem. The splits are based on covariate k being above or below a certain threshold. The process stops at some depth and predict with the mean of y at the terminal node x belongs to. The covariate and the threshold combination that determine the split are selected so as to have the largest separation possible. This is commonly referred to as a 'greedy' way of creating random forests.

The detailed way of constructing random forests are as follows. Randomness is injected in several levels: one to select a random sample to train each tree on. Second level of randomness is involved in selecting variable candidates to determine splits for the trees. At the end, we get a prediction for any x, average of the predictors of the trees. We can also estimate the prediction error at any observation x_i and gain expected prediction error overall. We can also identify which covariates contributed the most the the prediction using the variance-importance plots. This is where the random forests gain the advantage of interpretability - explaining as to why the results of the prediction came out to be as it is (not just spitting out the number).

There are some trade-offs. We want some level of depths in order to reduce the variance in the prediction and get a better fitting. However, too much depth will lead to overfitting. It is usually very difficult to find the happy medium.

Below is the example of the decision brought from Varian (2014). A random forests would be obtained by multiple trees, each with different sample (selected through bootstrap) and different variable and threshold combination.

Home Mortgage Disclosure Act (HMDA) Data Tree



Notes: Figure 5 shows a conditional tree estimated using the **R** package **party**. The black bars indicate the fraction of each group who were denied mortgages. The most important determinant of this is the variable "dmi," or "denied mortgage insurance." Other variables are: "dir," debt payments to total income ratio; "hir," housing expenses to income ratio; "lvr," ratio of size of loan to assessed value of property; "ccs," consumer credit score; "mcs," mortgage credit score; "pbcr," public bad credit record; "dmi," denied mortgage insurance; "self," self-employed; "single," applicant is single; "uria," 1989 Massachusetts unemployment rate applicant's industry; "condominium," unit is condominium; "black," race of applicant black; and "deny," mortgage application denied.

Figure 11.3: From Varian (2014), Figure 5

11.5 Double Machine Learning De-biasing Technics

Often we are interested in testing for regarding the value of the parameter of interest. For instance, the average treatment effect under the conditional independence assumption. Specifically, we assume $(\epsilon_i(1), \epsilon_i(0)) \perp\!\!\!\perp u_i | X_i$ where X_i contains a large set of covariates. What we may do is to estimate the propensity score using machine learning techniques such as LASSO. Then, estimate the average treatment effect with the formula of our choice. Ideally, we want to conduct some standard tests.

Here is the problem - what we have for the propensity score is an estimate with some error. So depending on the size of this error, the estimate of the average treatment effect could be affected. Moreover, the propensity score is just a means towards our ultimate goal of obtaining the average treatment effect - it is merely a nuisance parameter. Since what we are doing involves errors at two different stages, we need a double machine learning debiasing technics. Specifically, we decouple the propensity score and the treatment effect by applying Neyman Orthogonalization. In principle, what this method allows us to do is to do a two-step estimation by getting the estimate for the propensity score and then estimate the ATE without having to worry too much about the contaminated standard errors originating from the fact that we are not using the exact the propensity score.

In our case, what we can do is to

1. Split data into K folds
2. Use LASSO to estimate $\hat{P}_k = \Pr(D_i = 1|X_i)$ in all folds except the kth one. k can be from 1 to K.
3. If we are using inverse probability weighting method, for instance, we can get

$$\widehat{ATE} = \sum_{k=1}^{K} \left[\frac{\sum_{i\in k} \frac{Y_i D_i}{\hat{P}_k(X_i)}}{\sum_{i\in k} \frac{D_i}{\hat{P}_k(X_i)}} - \frac{\sum_{i\in k} \frac{Y_i(1-D_i)}{1-\hat{P}_k(X_i)}}{\sum_{i\in k} \frac{(1-D_i)}{1-\hat{P}_k(X_i)}} \right]$$

and use standard asymptotics to conduct tests.