

## SW Ch.4: Simple linear regression

Ryu Myeonggyu

April 29, 2025

# Introduction to Linear Regression

- ▶ What is regression?
- ▶ Why do we need regression?

## Joint Density and Marginal Density

- ▶ Let  $X, Y$  be random variables with joint density  $p_{XY}(x, y)$
- ▶ The marginal density  $p_X(x)$  is

$$p_X(x) = \int p_{XY}(x, y) dy$$

which describes the rate at which probability mass is distributed.

- ▶ The conditional density of  $Y$  given  $X$

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

- ▶ relative likelihood of obtaining  $Y = y$  given the information  $X = x$ .

# The Conditional Expectation Function

- ▶ The expectation of  $X$

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x p_X(x) dx$$

- ▶ The conditional expectation of  $Y$  given  $X$  is a function of the random variable  $X$  defined by

$$\mathbb{E}[Y|X = x] := \int_{-\infty}^{\infty} y \cdot p_{Y|X}(y|x) dy$$

# The Conditional Expectation Function

- ▶ We want to explain  $Y$  in terms of  $X$ .
- ▶ e.g. Does investment in education ( $X$ ) increase student's test score ( $Y$ )? Given today's stock price ( $X$ ), what will be tomorrow's ( $Y$ )?
- ▶ We want a function of  $X$ , say  $f(X)$ , explaining the  $Y$  variable.
- ▶ Of course, there is always an error, say  $u$ :

$$Y = f(X) + u = f(X) + (Y - f(X))$$

- ▶ The best function  $f$  will be such that minimizing the error  $u$

# The Conditional Expectation Function

- ▶ **Fact** The conditional expectation function minimizes the mean-squared error

$$MSE = \mathbb{E}[u^2] = \mathbb{E}[(Y - f(X))^2]$$

- ▶ Thus we want to know the conditional expectation function.

## Linear Conditional Expectation

- ▶ In general, it is almost impossible to find a function  $f$  using finitely many data points.
- ▶ We assume the conditional expectation function is **linear** in  $X$ . (life is easier with some assumptions)

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

- ▶ This is called the population regression function.
- ▶ So we can write

$$Y = \mathbb{E}[Y|X] + u = \beta_0 + \beta_1 X + u$$

- ▶ Here we only need to know 2 numbers! (we call them **parameters**)

# Estimating the Coefficients

- ▶ How can we estimate  $\beta_0$  and  $\beta_1$ ?
- ▶ Suppose we are given data  $\{(Y_i, X_i) | i = 1, \dots, n\}$
- ▶ The linear regression model

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{part of } Y \text{ explained by } X} + \underbrace{u_i}_{\text{error}}$$

- ▶ Choose the best  $\beta_0$  and  $\beta_1$ .
- ▶ One criterion: Minimize the unexplained part!

## Least Squares Estimation

- ▶ The ordinary least squares (OLS) estimator  $(\hat{\beta}_0, \hat{\beta}_1)$  of  $(\beta_0, \beta_1)$  is defined as the solution of the minimization problem below

$$\min_{b_0, b_1} \sum_{i=1}^n u_i^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

- ▶ The First-order conditions (FOC) are

$$(\partial b_0) \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$(\partial b_1) \sum_{i=1}^n X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

# Least Squares Estimation

- ▶ Solving this, we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$$

- ▶ Now we can write

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_i}_{\text{predicted value} = \hat{Y}_i} + \underbrace{\hat{u}_i}_{\text{residual}} \end{aligned}$$

## Some Properties

- ▶ 1.  $\sum_{i=1}^n \hat{u}_i = 0$
- ▶ 2.  $\sum_{i=1}^n X_i \hat{u}_i = 0$
- ▶ 3.  $\bar{Y}_i = \bar{\hat{Y}}_i$
- ▶ 4.  $\sum_{i=1}^n \hat{Y}_i \hat{u}_i = 0$
- ▶ Proof)

## Goodness-of-fit

- ▶ The decomposition of  $Y_i$

$$Y_i = \hat{Y}_i + \hat{u}_i$$

- ▶ Taking the average on both sides,

$$\bar{Y}_i = \bar{\hat{Y}}_i + \bar{\hat{u}}_i$$

- ▶ Subtract and square-sum

$$\sum_{i=1}^n (Y_i - \bar{Y}_i)^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}}_i)^2 + \sum_{i=1}^n (\hat{u}_i - \bar{\hat{u}}_i)^2$$

(the freshman's dream come true! do you know why?)

## Goodness-of-fit

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2}_{TSS} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}}_i)^2}_{ESS} + \underbrace{\sum_{i=1}^n (\hat{u}_i - \bar{\hat{u}}_i)^2}_{RSS}$$

- ▶ The coefficient of determination (almost never called like this) or the regression  $R^2$  is

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- ▶ SER (almost never used in practice) and RMSE

$$SER = \sqrt{\frac{1}{n-2} SSR}, \quad RMSE = \sqrt{\frac{1}{n} SSR}$$

## Interpretation of OLS coefficient

- ▶ Under the assumptions (A1)-(A3), the OLS coefficient has a causal interpretation. (Otherwise it should be interpreted as correlation)
- ▶ (A1) (conditional mean independence)

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

- ▶ (A2) (iid)  $(X_i, Y_i)$  are independent and identically distributed.  
*Note.* It does not mean  $X_i$  and  $Y_i$  are independent!
- ▶ (A3) Large outliers are unlikely. (technically needed)

$$\mathbb{E}[X_i^4], \mathbb{E}[Y_i^4] < \infty$$

## The meaning of CMI

- ▶ Remark that

$$Y_i = \beta_0 + \beta_1 X_i + u_i \cdots (*)$$

- ▶ We wish that  $\beta_0 + \beta_1 X_i = \mathbb{E}[Y_i | X_i]$
- ▶ Taking  $\mathbb{E}[\cdot | X_i]$  on (\*), we have

$$\mathbb{E}[Y_i | X_i] = \beta_0 + \beta_1 X_i + \mathbb{E}[u_i | X_i]$$

- ▶ If  $\mathbb{E}[u_i | X_i] \neq 0$ , it means your linearity assumption is wrong! (model misspecification)
- ▶ The error term contains all the variables other than  $X$  that explains  $Y$ .
- ▶ If  $\mathbb{E}[u_i | X_i] \neq 0$ , it also means your error term contains some important factors that you omitted.
- ▶ later in Potential Outcomes Framework in detail

# The sampling distribution of OLS

- ▶ Recall:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- ▶ Expectation

- ▶ Variance

## The Asymptotic distribution

- ▶ We say that a sequence of random variables  $(X_n)$  converges in probability to a random variable  $X$  if for every  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$$

and in this case, we write  $X_n \xrightarrow{P} X$ .

- ▶ An estimator  $\theta_n$  of a parameter  $\theta$  is consistent if

$$\theta_n \xrightarrow{P} \theta$$

- ▶ **Weak Law of Large Numbers (WLLN):** If  $(X_i)$  is an iid sequence of random variables such that  $\mathbb{E}[X_i] = \mu < \infty$ , then

$$\bar{X}_n \xrightarrow{P} \mu$$

## The Asymptotic distribution

- ▶ A random variable  $X$  is said to be normally distributed if its density is

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

for some  $\mu$  and  $\sigma > 0$ .

- ▶ In this case we write  $X \sim N(\mu, \sigma^2)$ .
- ▶ Suppose  $X_n$  has distribution function  $F_n$  and  $X$  has a distribution function  $F$ . If  $F_n(x) \rightarrow F(x)$  for all  $x$  at which  $F$  is continuous, then we say  $X_n$  converges in distribution to  $X$ .
- ▶ **Central Limit Theorem** (Lindeberg-Levy CLT) If  $(X_n)$  is an iid sequence of random variables with  $\mathbb{E}[X_i] = \mu$ ,  $\text{Var}(X_i) = \sigma^2 < \infty$ , then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

# The Asymptotic distribution

- ▶ **Theorem.** Under (A1) – (A3), the OLS estimator  $\hat{\beta}_1$  is consistent and asymptotically normal.
- ▶ Consistency
- ▶ Asymptotic normality

## Exercise 1

- ▶ Suppose  $Y_i = \beta + u_i$  (intercept-only model).
- ▶ (a) Find the OLS estimator of  $\beta$ .
- ▶ (b) Find the  $R^2$  of this regression.

## Exercise 2

- ▶ Suppose  $Y_i = \beta_0 + \beta_1 X_i + u_i$
- ▶ Let  $\rho$  be the sample correlation between  $X$  and  $Y$ .
- ▶ (a) Write down the formula for  $\rho$ .
- ▶ (b) Show that

$$R^2 = \rho^2.$$

- ▶ (c) Show that  $\hat{\beta}_1 = \rho \frac{s_y}{s_x}$  where  $s$  is the sample standard deviation.

## Exercise 3

- ▶ (Method of Moments) We say  $X$  and  $u$  are orthogonal if  $\mathbb{E}[Xu] = 0$ . (here  $\mathbb{E}[u] = 0$ )
- ▶ **The Analogy Principle.** Since the population mean ( $\mathbb{E}$ ) is not feasible, one way to estimate it is by replacing the unknown expectation with the sample average ( $\frac{1}{n} \sum_{i=1}^n$ )
- ▶ Using the analogy principle, show that the orthogonality condition and mean-zero error condition characterize the OLS estimator  $(\hat{\beta}_0, \hat{\beta}_1)$ :

$$\mathbb{E}[Xu] = 0$$

$$\mathbb{E}[u] = 0$$

## Exercise 4

- ▶ Suppose you want to analyze the effect of average temperature on average weekly earnings (AWE) using the regression model

$$AWE_i = \beta_0 + \beta_1 \text{Temperature}_i + u_i$$

Your friend Rachel is an American and decides to analyze the effect of temperature measured in Fahrenheit, while you did the same analysis in Celsius. ( $F^\circ = 32 + \frac{9}{5}C^\circ$ )

- ▶ If everything else is the same in Rachel's analysis, how will the following quantities differ?
  - ▶ (a)  $\hat{\beta}_0$
  - ▶ (b)  $\hat{\beta}_1$
  - ▶ (c)  $R^2$

# SW6. Multiple Linear Regression

## Part 1

Ryu Myeonggyu

May 4, 2025

## Motivating Example

- ▶ Consider a relationship between income ( $Y$ ) and years of education ( $X$ ).
- ▶ Collecting data  $\{(Y_i, X_i)\}$ , you estimate the linear regression.
- ▶ Do we fully capture the impact of  $X$  on  $Y$ ?
- ▶ e.g. Consider the ability ( $Z$ ).
- ▶ Higher ability → Higher education
- ▶ Higher ability → Higher income

## Omitted Variable Bias

- ▶ Here  $X$  and  $Z$  are correlated and  $Z$  and  $Y$  are correlated.
- ▶ The true model is

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + u$$

but you omit  $Z$  in your model

$$Y = \beta_0 + \beta_1 X + e$$

so that

$$e = \beta_2 Z + u$$

## Omitted Variable Bias

- ▶ In this case the probability limit of  $\hat{\beta}_1$  is

$$p \lim \hat{\beta}_1 = \frac{Cov(X, Y)}{Var(X)} = \frac{Cov(X, \beta_0 + \beta_1 X + \beta_2 Z + u)}{Var(X)}$$

$$= \beta_1 + \beta_2 \frac{Cov(X, Z)}{Var(X)} = \beta_1 + \frac{Cov(X, u)}{Var(X)}$$

assuming  $Cov(X, e) = \mathbb{E}[Xe] = 0$  in the true model.

- ▶ This is called omitted variable bias. (OVB)

# Multiple Linear Regression

- ▶ The solution to omitted variable bias is not to omit the variable.
- ▶ Our regression model becomes

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- ▶ How do we estimate this?
- ▶ The same principle applies: the Least Squares!

# Least Squares Estimator

- ▶ Solve the minimization problem

$$\min_{b_0, b_1, b_2} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i - b_2 Z_i)^2$$

- ▶ FOC

## Least Square Estimator

- ▶ Solving this system of equations can be insanely difficult.
- ▶ What if the number of regressors is extremely large? (say 1,000)
- ▶ We will introduce a nicer approach using vectors and matrices.
- ▶ To this end, we need some mathematical background, the linear algebra.

# Vector Space

- ▶ For a natural number  $m$ , we define

$$\mathbb{R}^m = \{(x_1, \dots, x_m)' | x_j \text{'s are real numbers}\}$$

the Euclidean  $m$ -dimensional space.

- ▶ A set  $V$  is called a vector space (over  $\mathbb{R}$ )<sup>1</sup> if it is equipped with addition (+) and scalar multiplication ( $\cdot$ ) such that
  - ▶  $(v + w) + u = v + (w + u)$  for all  $v, w, u \in V$
  - ▶  $v + w = w + v$  for all  $v, w \in V$
  - ▶ There exists  $0 \in V$  such that  $0 + v = v + 0 = v$
  - ▶ For each  $v \in V$ , there exists  $(-v) \in V$  such that  $v + (-v) = 0$
  - ▶ For  $1 \in \mathbb{R}$ , we have  $1 \cdot v = v$
  - ▶ For all  $a, b \in \mathbb{R}$  and  $v \in V$ ,  $(ab) \cdot v = a \cdot (b \cdot v)$
  - ▶ For all  $v, w \in V$  and  $c \in \mathbb{R}$ ,  $c \cdot (v + w) = c \cdot v + c \cdot w$
  - ▶ For all  $a, b \in \mathbb{R}$  and  $v \in V$ ,  $(a + b) \cdot v = a \cdot v + b \cdot v$

---

<sup>1</sup>From now on, every vector space is over  $\mathbb{R}$  unless otherwise stated.

# Linear Algebra

- ▶ Define the addition and scalar multiplication on  $\mathbb{R}^m$  by

$$(x_1, \dots, x_m)' + (y_1, \dots, y_m)' = (x_1 + y_1, \dots, x_m + y_m)'$$

and

$$c(x_1, \dots, x_m)' = (cx_1, \dots, cx_m)'$$

- ▶ Under these operations,  $(\mathbb{R}^m, +, \cdot)$  is a vector space over  $\mathbb{R}$ .  
(exercise)
- ▶ We call  $m$  the dimension of  $\mathbb{R}^m$ .

# Matrix

- ▶ A matrix is a rectangular array of numbers.
- ▶ A matrix with  $n$  rows and  $k$  columns is called a  $n \times k$  matrix.

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ & \vdots & \\ a_{n1} & \cdots & a_{nk} \end{pmatrix} = (a_{ij})_{n \times k}$$

- ▶ A matrix all of whose entries equal to zero is called a zero matrix and denoted by  $O$ .
- ▶ A matrix is called a square matrix if number of rows=number of columns.
- ▶  $M_{n,k}$  denotes the set of  $n \times k$  matrices.
- ▶ If  $n = k$ , we may write  $M_n$  instead of  $M_{n,n}$ .

# Linear Algebra

- ▶ Addition and multiplication of matrices
- ▶ Addition: For  $A, B \in M_{n,k}$ , we define  $A + B$  by

$$A + B = (a_{ij} + b_{ij})_{n \times k}$$

the component-wise addition

- ▶ scalar multiplication

$$cA = (ca_{ij})_{nk}$$

- ▶ matrix multiplication: For  $A \in M_{n,k}$  and  $B \in M_{k,\ell}$ ,

$$AB = \left( \sum_{t=1}^k a_{it} b_{tj} \right)_{n \times \ell}$$

# Linear Algebra

- ▶ For  $A = (a_{ij})_{n \times k} \in M_{n \times k}$ , the transpose of  $A$  is a  $k \times n$  matrix whose  $(i, j)$ -component is  $a_{ji}$ . We denote the transpose of  $A$  by  $A^T$  or  $A'$ .
- ▶ i.e.

$$A^T = (a_{ji})_{k \times n} \in M_{n \times k}$$

- ▶ In econometrics, it is customary to use  $A'$  instead of  $A^T$ .

## Example

- $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, B = \begin{pmatrix} 0 & 2 & 1 \\ -1 & 1 & -2 \end{pmatrix}, C = \begin{pmatrix} -3 & 1 \\ 0 & 2 \\ -1 & 0 \end{pmatrix}$
- $A + B =$
- $AB =$
- $A' =$
- $A + C =$
- $AC =$
- $C' =$

# Useful Properties of Matrix Operations

- ▶  $A + B = B + A$
- ▶  $(A + B) + C = A + (B + C)$
- ▶  $(A + B)' = A' + B'$
- ▶  $A(BC) = (AB)C$
- ▶  $A(B + C) = AB + AC$
- ▶  $(AB)' = B'A'$

# Matrix multiplication

- ▶ **Remark.** In general, we cannot say

$$AB = BA$$

even when the multiplication is defined!

- ▶ The following square matrix  $I$  is called the identity matrix

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- ▶ That is,  $I_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

## Matrix multiplication

- ▶ You can directly check that for any matrix  $A$  and  $B$  (with adequate dimension)

$$IA = A$$

and

$$BI = B$$

- ▶ For a square matrix  $A \in M_n$ , if there exists a matrix  $B$  such that

$$AB = BA = I$$

we say  $A$  is invertible (or non-singular).

- ▶ In this case we call  $B$  the inverse of  $A$  and denote it by  $A^{-1}$ .

## Inverse matrix

- ▶ For example, let  $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$
- ▶  $A^{-1} =$

## Inverse matrix

- ▶ In general, it is not easy to find an inverse of a matrix.
- ▶ We first check if the matrix has an inverse.
- ▶ Some useful concepts regarding this: rank and determinant
- ▶ In the next lecture, we will learn the meaning of matrix equation and why finding an inverse is important.

# Exercises

- ▶ Stock and Watson, Exercise 6.6, 6.9, 6.11,

## Exercises

- ▶ Let  $V$  be a vector space. Prove **the cancellation law** below using the axioms of vector space.
- ▶ Let  $u, v, w \in V$ . If

$$u + v = w + v$$

then  $u = w$ .

## Exercises

- ▶ Let  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$
- ▶ Show that  $B = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$  is the inverse of  $A$ .

## Exercises

- ▶ A matrix  $D$  is called diagonal if all of its off-diagonal components are zero.
- ▶ For a diagonal matrix  $D$ , when is  $D$  invertible? Under such condition, find its inverse.

## Exercises

- ▶ (a) Determine whether the set of square matrices  $M_n$  is a vector space with addition and scalar multiplication.
- ▶ (b) For  $A \in M_n$ , suppose  $A$  is invertible. How many inverse matrices can  $A$  have?

## Exercises

- ▶ (a) A matrix  $A \in M_n$  is called *symmetric* if  $A = A'$  and *anti-symmetric* if  $A = -A'$ . Give some examples of symmetric and anti-symmetric matrices.
- ▶ (b) Determine if the following matrices are symmetric or anti-symmetric

$$(i) \frac{A + A'}{2} \quad (ii) \frac{A - A'}{2}$$

- ▶ (c)<sup>†</sup> Show that any square matrix  $A$  can be written as a sum of a symmetric matrix and an anti-symmetric matrix. i.e. there exist a symmetric matrix  $B$  and an anti-symmetric matrix  $C$  such that

$$A = B + C$$

- ▶ (d)<sup>†</sup> Do  $B$  and  $C$  in (c) exist uniquely?

# SW6. Multiple Linear Regression

## Part 1

Ryu Myeonggyu

May 8, 2025

## Motivating Example

- ▶ Consider a relationship between income ( $Y$ ) and years of education ( $X$ ).
- ▶ Collecting data  $\{(Y_i, X_i)\}$ , you estimate the linear regression.
- ▶ Do we fully capture the impact of  $X$  on  $Y$ ?
- ▶ e.g. Consider the ability ( $Z$ ).
- ▶ Higher ability → Higher education
- ▶ Higher ability → Higher income
- ▶ Is higher income attributable solely to the education?

## Omitted Variable Bias

- ▶ Here  $X$  and  $Z$  are correlated and  $Z$  and  $Y$  are correlated.
- ▶ The true model is

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + u$$

but you omit  $Z$  in your model

$$Y = \beta_0 + \beta_1 X + e$$

so that

$$e = \beta_2 Z + u$$

## Omitted Variable Bias

- ▶ In this case the probability limit of  $\hat{\beta}_1$  is

$$p \lim \hat{\beta}_1 = \frac{Cov(X, Y)}{Var(X)} = \frac{Cov(X, \beta_0 + \beta_1 X + \beta_2 Z + u)}{Var(X)}$$

$$= \beta_1 + \beta_2 \frac{Cov(X, Z)}{Var(X)} = \beta_1 + \frac{Cov(X, e)}{Var(X)}$$

assuming  $Cov(X, u) = \mathbb{E}[Xu] = 0$  in the true model.

- ▶ This is called omitted variable bias. (OVB)
- ▶ Remember that in our model,  $\mathbb{E}[Xe] \neq 0$ .

# Multiple Linear Regression

- ▶ The solution to omitted variable bias is not to omit the variable.
- ▶ Our regression model becomes

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- ▶ How do we estimate this?
- ▶ The same principle applies: the Least Squares!

# Least Squares Estimator

- ▶ Solve the minimization problem

$$\min_{b_0, b_1, b_2} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i - b_2 Z_i)^2$$

- ▶ FOC

## Least Squares Estimator

- ▶ Solving this system of equations by hand with summation and fractions becomes increasingly difficult when you have a lot of regressors. (say 1,000 or more)
- ▶ We will introduce a nicer approach using vectors and matrices.
- ▶ To this end, we need some mathematical background, the linear algebra.
- ▶ Warning: This can be really complicated.

# Vector Space

- ▶ For a natural number  $m$ , we define

$$\mathbb{R}^m = \{(x_1, \dots, x_m)' | x_j \text{'s are real numbers}\}$$

the Euclidean  $m$ -dimensional space.

- ▶ A set  $V$  is called a vector space (over  $\mathbb{R}$ )<sup>1</sup> if it is equipped with addition (+) and scalar multiplication ( $\cdot$ ) such that
  - ▶  $(v + w) + u = v + (w + u)$  for all  $v, w, u \in V$
  - ▶  $v + w = w + v$  for all  $v, w \in V$
  - ▶ There exists  $0 \in V$  such that  $0 + v = v + 0 = v$
  - ▶ For each  $v \in V$ , there exists  $(-v) \in V$  such that  $v + (-v) = 0$
  - ▶ For  $1 \in \mathbb{R}$ , we have  $1 \cdot v = v$
  - ▶ For all  $a, b \in \mathbb{R}$  and  $v \in V$ ,  $(ab) \cdot v = a \cdot (b \cdot v)$
  - ▶ For all  $v, w \in V$  and  $c \in \mathbb{R}$ ,  $c \cdot (v + w) = c \cdot v + c \cdot w$
  - ▶ For all  $a, b \in \mathbb{R}$  and  $v \in V$ ,  $(a + b) \cdot v = a \cdot v + b \cdot v$

---

<sup>1</sup>From now on, every vector space is over  $\mathbb{R}$  unless otherwise stated.

# Linear Algebra

- ▶ Define the addition and scalar multiplication on  $\mathbb{R}^m$  by

$$(x_1, \dots, x_m)' + (y_1, \dots, y_m)' = (x_1 + y_1, \dots, x_m + y_m)'$$

and

$$c(x_1, \dots, x_m)' = (cx_1, \dots, cx_m)'$$

- ▶ Under this operations,  $(\mathbb{R}^m, +, \cdot)$  is a vector space over  $\mathbb{R}$ .  
(exercise)
- ▶ We call  $m$  the dimension of  $\mathbb{R}^m$ .
- ▶ Let  $W$  be a subset of  $V$ . If  $W$  is a vector space under addition and scalar multiplication inherited from  $V$ , then we say  $W$  is a subspace of  $V$ .
- ▶ e.g.  $V = \mathbb{R}^2$ ,  $W = \{(x, y) | y = 3x\}$

# Matrix

- ▶ A matrix is a rectangular array of numbers.
- ▶ A matrix with  $n$  rows and  $k$  columns is called a  $n \times k$  matrix.

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ & \vdots & \\ a_{n1} & \cdots & a_{nk} \end{pmatrix} = (a_{ij})_{n \times k}$$

- ▶ A matrix all of whose entries equal to zero is called a zero matrix and denoted by  $O$ .
- ▶ A matrix is called a square matrix if number of rows=number of columns.
- ▶  $M_{n,k}$  denotes the set of  $n \times k$  matrices. (or  $M_{n \times k}$  synonymously)
- ▶ If  $n = k$ , we may write  $M_n$  instead of  $M_{n,n}$ .

# Linear Algebra

- ▶ Addition and multiplication of matrices
- ▶ Addition: For  $A, B \in M_{n,k}$ , we define  $A + B$  by

$$A + B = (a_{ij} + b_{ij})_{n \times k}$$

the component-wise addition

- ▶ scalar multiplication

$$cA = (ca_{ij})_{n \times k}$$

- ▶ matrix multiplication: For  $A \in M_{n,k}$  and  $B \in M_{k,\ell}$ ,

$$AB = \left( \sum_{t=1}^k a_{it} b_{tj} \right)_{n \times \ell}$$

# Linear Algebra

- ▶ For  $A = (a_{ij})_{n \times k} \in M_{n \times k}$ , the transpose of  $A$  is a  $k \times n$  matrix whose  $(i, j)$ -component is  $a_{ji}$ . We denote the transpose of  $A$  by  $A^T$  or  $A'$ .
- ▶ i.e.

$$A^T = (a_{ji})_{k \times n} \in M_{k \times n}$$

- ▶ In econometrics, it is customary to use  $A'$  instead of  $A^T$ .

## Example

- $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, B = \begin{pmatrix} 0 & 2 & 1 \\ -1 & 1 & -2 \end{pmatrix}, C = \begin{pmatrix} -3 & 1 \\ 0 & 2 \\ -1 & 0 \end{pmatrix}$
- $A + B =$
- $AB =$
- $A' =$
- $A + C =$
- $AC =$
- $C' =$

# Useful Properties of Matrix Operations

- ▶ Whenever the operations are well-defined,
- ▶  $A + B = B + A$
- ▶  $(A + B) + C = A + (B + C)$
- ▶  $(A + B)' = A' + B'$
- ▶  $A(BC) = (AB)C$
- ▶  $A(B + C) = AB + AC$
- ▶  $(AB)' = B'A'$

# Matrix multiplication

- ▶ **Remark.** In general, we cannot say

$$AB = BA$$

even when the multiplication is defined!

- ▶ The following square matrix  $I$  is called the identity matrix

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- ▶ That is,  $I_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

## Matrix multiplication

- ▶ You can directly check that for any matrix  $A$  and  $B$  (with adequate dimension)

$$IA = A$$

and

$$BI = B$$

- ▶ For a square matrix  $A \in M_n$ , if there exists a matrix  $B$  such that

$$AB = BA = I$$

we say  $A$  is invertible (or non-singular).

- ▶ In this case we call  $B$  the inverse of  $A$  and denote it by  $A^{-1}$ .

## Inverse matrix

- ▶ For example, let  $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$
- ▶  $A^{-1} =$

## Inverse matrix

- ▶ In general, it is not easy to find an inverse of a matrix.
- ▶ We first check if the matrix has an inverse.
- ▶ Some useful concepts regarding this: rank and determinant
- ▶ In the next lecture, we will learn the meaning of matrix equation and why finding an inverse is important.

# Exercises

- ▶ Stock and Watson, Exercise 6.6, 6.9, 6.11,

## Exercises

- ▶ Let  $V$  be a vector space. Prove the following using the axioms of vector spaces.
- ▶ (a) (**Cancellation law**) Let  $u, v, w \in V$ . If

$$u + v = w + v$$

then  $u = w$ .

- ▶ (b) Show that a vector space has a unique zero vector. i.e. if  $0$  and  $\tilde{0}$  are two zero vectors, then  $0 = \tilde{0}$ .
- ▶ (c) Show that for all  $v \in V$ , we have

$$0 \cdot v = 0.$$

(Notice that the  $0$  on the left hand side is a scalar, while  $0$  on the right hand side is the zero vector)

## Exercises

- ▶ Let  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$
- ▶ Show that  $B = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$  is the inverse of  $A$ .

## Exercises

- ▶ A matrix  $D$  is called diagonal if all of its off-diagonal components are zero.
- ▶ i.e.

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}$$

- ▶ For a diagonal matrix  $D$ , when is  $D$  invertible? Under such condition, find its inverse.

## Exercises

- ▶ (a) Determine whether the set of square matrices  $M_n$  is a vector space with addition and scalar multiplication.
- ▶ (b) For  $A \in M_n$ , suppose  $A$  is invertible. How many inverse matrices can  $A$  have?

## Exercises

- ▶ (a) A matrix  $A \in M_n$  is called *symmetric* if  $A = A'$  and *anti-symmetric* if  $A = -A'$ . Give some examples of symmetric and anti-symmetric matrices.
- ▶ (b) Determine if the following matrices are symmetric or anti-symmetric

$$(i) \frac{A + A'}{2} \quad (ii) \frac{A - A'}{2}$$

- ▶ (c)<sup>†</sup> Show that any square matrix  $A$  can be written as a sum of a symmetric matrix and an anti-symmetric matrix. i.e. there exist a symmetric matrix  $B$  and an anti-symmetric matrix  $C$  such that

$$A = B + C$$

- ▶ (d)<sup>†</sup> Do  $B$  and  $C$  in (c) exist uniquely?

# SW6. Multiple Linear Regression

## Part 2

Ryu Myeonggyu

May 18, 2025

## Inverse of a matrix

- ▶ Recall) For  $A \in M_n$ , if there exists a matrix  $B \in M_n$  such that

$$AB = BA = I$$

then we say  $A$  is an invertible matrix and  $B$  is the inverse of  $A$ .

- ▶ Why is finding an inverse so important?

# System of Linear Equations

- ▶ Consider a system of linear equations

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n = b_n \end{cases}$$

- ▶ Define

$$A = (a_{ij}) = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

$$x = (x_1, \dots, x_n)' \text{ and } b = (b_1, \dots, b_n)'$$

## System of Linear Equations

- ▶ Then the above system of linear equations can be written in matrix form by

$$Ax = b$$

- ▶ If we want to find the solution  $x$  and if we know  $A^{-1}$ , then we must have

$$x = A^{-1}b$$

by multiplying  $A^{-1}$  to both sides.

- ▶ Hence, finding an inverse  $\iff$  solving a system of linear equations!
- ▶ From now on, we consider a matrix  $A \in M_n$  with

$$A = (a_{ij}) = (A_1 \quad \cdots \quad A_n)$$

- ▶ That is,  $i$ -th column of  $A$  will be denoted by  $A_i$ .

# Linear Map

- ▶ Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . We say  $f$  is a linear map if for any  $a \in \mathbb{R}$  and  $x, y \in \mathbb{R}^n$  we have

$$f(x + y) = f(x) + f(y)$$

and

$$f(ax) = af(x)$$

- ▶ Note that the matrix multiplication is a linear map:

$$L_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

with  $x \mapsto Ax$  is linear.

- ▶ Also it is useful to remember that

$$Ax = A_1x_1 + \cdots + A_nx_n$$

## Invertibility

- ▶ We know that a function is invertible if and only if it is 1-1 and onto.
- ▶ Recall: A function  $f$  is 1-1 if  $f(x) = f(y)$  implies  $x = y$ .
- ▶ A function  $f$  is onto if for every  $y$ , there exists  $x$  such that  $y = f(x)$ .
- ▶ Since  $L_A$  is also a function, we know that  $L_A$  is invertible  
 $\iff L_A$  is 1-1 and onto.

## Invertible Matrix

- ▶ Let  $V$  be a vector space. For  $v_1, \dots, v_n \in V$ , we say that  $S := \{v_1, \dots, v_n\}$  is linearly independent if the following equation

$$c_1 v_1 + \cdots + c_n v_n = 0$$

has the unique solution  $c_1 = \cdots = c_n = 0 \in \mathbb{R}$ .

- ▶ If  $S$  is not linearly independent, we say  $S$  is linearly dependent.
- ▶ Also we define the span of  $S$  by

$$\text{span}(S) = \{c_1 v_1 + \cdots + c_n v_n \mid c_1, \dots, c_n \in \mathbb{R}\}$$

- ▶ Each element in the  $\text{span}(S)$  is called a linear combination of (elements of)  $S$ .

## Invertible Matrix

- ▶  $S$  is called a basis of  $V$  if  $S$  is linearly independent and  $\text{span}(S) = V$ .
- ▶ The number of elements of basis is called the dimension of  $V$  and we denote it by  $\dim V$ .
- ▶ e.g. Let  $V = \mathbb{R}^3$ .
- ▶  $S_1 = \{(1, 0, 0)', (0, 1, 0)'\}$
- ▶  $S_2 = \{(1, 0, 0)', (0, 1, 0)', (0, 0, 1)', (0, 1, 1)'\}$
- ▶  $S_3 = \{(1, 0, 0)', (0, 1, 0)', (0, 0, 1)'\}$
- ▶ In general for  $V = \mathbb{R}^n$ , the vectors  $\{e_1, \dots, e_n\}$  is a basis for  $V$  where  $e_j$  is defined by

$$e_j = (0, 0, \dots, 0, \underbrace{1}_{j\text{-th coordinate}}, 0, \dots, 0)'$$

## Basis of a Vector Space

- ▶ **Proposition.** Let  $V = \mathbb{R}^n$ . Suppose  $S \subset V$  has  $n$  elements. Then  $S$  is linearly independent if and only if  $\text{span}(S) = \mathbb{R}^n$ .
- ▶ **Proof)** First note that  $E := \{e_1, \dots, e_n\}$  is a basis for  $\mathbb{R}^n$ .  
 $(\Rightarrow)$  Suppose  $S = \{v_1, \dots, v_n\}$  is linearly independent. If  $\text{span}(S) \neq \mathbb{R}^n$ , then there exists  $w \in \mathbb{R}^n$  such that  $w \notin \text{span}(S)$ . On the other hand, since  $E$  spans  $\mathbb{R}^n$ , for each  $v_j$  we can write

$$v_j = \sum_{i=1}^n a_{ij} e_i$$

Since  $v_j$ 's are not zero vectors, in particular  $v_1$  is nonzero, so at least one  $a_{i1}$  is not zero. WLOG suppose  $a_{11} \neq 0$ . Then  $e_1 = \frac{1}{a_{11}}(v_1 - \sum_{i=2}^n a_{i1} e_i)$  so that  $v_1, e_2, \dots, e_n$  spans  $\mathbb{R}^n$ . Repeating this, we have  $\{v_1, \dots, v_n\}$  spans  $\mathbb{R}^n$  therefore  $w \in \text{span}(\{v_1, \dots, v_n\})$ . This is a contradiction.

## Proof of Proposition

- ▶ ( $\Leftarrow$ ) Now suppose  $\text{span}(S) = \mathbb{R}^n$ . Suppose not,  $S$  is linearly dependent. Then there is a linear relation

$$c_1 v_1 + \cdots + c_n v_n = 0$$

in which some  $c_k$  is not zero. WLOG suppose  $c_n \neq 0$ . Then  $v_n = -\frac{1}{c_n}(c_1 v_1 + \cdots + c_{n-1} v_{n-1}) \in \text{span}(\{v_1, \dots, v_{n-1}\})$ . Therefore,  $\mathbb{R}^n = \text{span}(\{v_1, \dots, v_{n-1}\})$ . Repeat this until we have  $\mathbb{R}^n = \text{span}(\{v_1, \dots, v_m\})$  and  $\{v_1, \dots, v_m\}$  is linearly independent. In particular, we may write

$$e_j = \sum_{i=1}^m b_{ij} v_i$$

## Proof of Proposition

- ▶ Let  $u = (u_1, \dots, u_n)' = u_1 e_1 + \dots + u_n e_n \in \mathbb{R}^n$ . Then

$$u = \sum_{j=1}^n u_j e_j = \sum_{j=1}^n u_j \sum_{i=1}^m b_{ij} v_i = \sum_{i,j} (u_j b_{ij}) v_i$$

Suppose  $u = 0$ . Then we need to have  $\sum_j u_j b_{ij} = 0$  for all  $i = 1, \dots, m$ . i.e.

$$\begin{cases} b_{11}u_1 + \dots + b_{1n}u_n = 0 \\ \dots \\ b_{m1}u_1 + \dots + b_{mn}u_n = 0 \end{cases}$$

That is, the number of equations =  $m < n$  = the number of unknowns. Hence we have a nontrivial solution  $(u_1, \dots, u_n)'$  which is a contradiction.

## Extension to a basis

- ▶ (Basis Extension Theorem) Let  $S$  be a linearly independent subset of  $V$ . Then there exists a basis  $\mathcal{B}$  of  $V$  such that  $S \subset \mathcal{B}$ .
- ▶ **Lemma.** Let  $T$  be a linearly independent subset of  $V$ . Then  $T' = T \cup \{v\}$  is linearly independent if and only if  $v \notin \text{span}(T)$ .
- ▶ Proof of Lemma)

## Invertible Matrix

- ▶ **Theorem.** (a) A linear map  $L_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is 1-1 if and only if  $Ax = 0$  has only a trivial solution  $x = 0$ .
- ▶ (b) A linear map  $L_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is onto if and only if the column vectors of  $A$  spans  $\mathbb{R}^n$ .
- ▶ (c)  $L_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is 1-1 if and only if it is onto.

# Proof

## Invertibility

- ▶ For a linear map  $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , we define

$$\ker L := \{x \in \mathbb{R}^n | Lx = 0\}$$

the kernel of  $L$ . We say  $L$  has a trivial kernel if  $\ker L = \{0\}$ .

- ▶ By the previous theorem, we know that  $L_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is invertible if and only if  $L_A$  has trivial kernel.
- ▶ We also define for a linear map  $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ,

$$imL_A := \{y \in \mathbb{R}^m | y = L_A(x) \text{ for some } x \in \mathbb{R}^n\}$$

which is called the image of  $L_A$ .

- ▶ It is straightforward that  $\ker L_A$  is a subspace of  $\mathbb{R}^n$  and  $imL_A$  is a subspace of  $\mathbb{R}^m$  (exercise)

## Inverse matrix

- ▶ We need to know whether the inverse map of  $L_A$  is also linear.
  - ▶ **Proposition.** The inverse of a linear map is also linear.
  - ▶ Proof)
- 
- ▶ We showed that a square matrix has an inverse matrix if and only if its column vectors are linearly independent. (or span  $\mathbb{R}^n$ )<sup>1</sup>

---

<sup>1</sup>In fact, we need to show more to claim this, but this is not a linear algebra course so we skip it.

# Rank

- ▶ We can generalize this concept of linear independence to non-square matrices.
- ▶ For  $A \in M_{n,k}$ , the maximum number of linearly independent column vectors of  $A$  is called the rank of  $A$ .
- ▶ If  $\text{rank}(A) = \min(n, k)$ , then we say  $A$  is of full rank.

# Rank

- ▶ Note also that one can define rank by the maximum number of linearly independent row vectors of  $A$ .
- ▶ **Fact.** The two definitions are equivalent.
- ▶ Now we can also say that a square matrix  $A$  is invertible if and only if it has full rank.

# The dimension theorem

- ▶ Let  $A \in M_{m,n}$ . Then

$$n = \dim \ker A + \text{rank}(A)$$

- ▶ This theorem is called the dimension theorem or the rank-nullity theorem.
- ▶ This is a fundamental theorem of linear algebra.

## Proof of Dimension Theorem

- ▶ Proof) Since  $\ker L_A$  is a subspace of  $\mathbb{R}^n$ , it has a basis, say  $\{v_1, \dots, v_k\}$ . Using the lemma on page 12, we can extend this to a basis of  $\mathbb{R}^n$ , say  $\{v_1, \dots, v_k, v_{k+1}, \dots, v_n\}$ . Let  $w_j = L_A v_j$  for  $j > k$ . Then we claim that  $\{w_{k+1}, \dots, w_n\}$  is a basis for  $\text{im } L_A$  which completes the proof.
- ▶ Proof of Claim) (linear independence)

(spans  $\text{im } L_A$ )

## Example

- ▶ Let  $A = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \end{pmatrix}$
- ▶  $\text{rank}(A)$
- ▶  $\ker A$
- ▶  $\text{im}A$
- ▶ basis of  $\ker A$
- ▶ basis of  $\text{im}A$
- ▶ Dimension theorem

## Example

- ▶ Consider a system of linear equations

$$\begin{cases} 2x + y = 0 \\ -x + y = 1 \\ x + 3y = -1 \end{cases}$$

- ▶ In matrix form, we can write this as

- ▶  $\text{rank}(A)$
- ▶  $\ker A$
- ▶  $\text{im}A$
- ▶ basis of  $\ker A$
- ▶ basis of  $\text{im}A$
- ▶ Dimension theorem

# Summary

- ▶ So far, we have learned...
- ▶ basis and dimension of a vector space
- ▶ matrix as a linear map
- ▶ kernel and image of a matrix
- ▶ invertibility conditions of a matrix
- ▶ rank and dimension theorem
- ▶ Next lecture we will learn one more invertibility condition:  
determinant.
- ▶ Then we will cover **projection theory** which is our final destination.

## Exercises

- ▶ Let  $L : V \rightarrow W$  be a linear map where  $V, W$  are vector spaces.
- ▶ (a) Show that  $L(0) = 0$ .
- ▶ (b) Suppose  $V = W = \mathbb{R}^2$ . Determine whether the function

$$f \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + 2y \\ 3x + 4y \end{pmatrix}$$

is a linear map. If so, find a matrix  $A$  such that  $f(x) = Ax$ .

- ▶ (c) Repeat (b) for the function

$$g \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 + 2x + y \\ -x + 2y \end{pmatrix}$$

## Exercises

- ▶ Let  $A, B \in M_n$ . Show that  $AB$  is invertible if and only if  $A$  and  $B$  are invertible.

**Hint** To show the if part, you can explicitly find the inverse of  $AB$ .  
To show the only if part, suppose to the contrary and the nontrivial kernel condition.

## Exercises

- ▶ Determine whether the following vectors form a basis of  $\mathbb{R}^3$ .
- ▶ (a)  $v_1 = (1, 0, 0)', v_2 = (0, 1, 0)', v_3 = (0, 0, 1)'$
- ▶ (b)  $w_1 = (0, 1, 1)', w_2 = (1, 0, 1)', w_3 = (1, 1, 0)'$
- ▶ (c)  $u_1 = (1, 1, 2)', u_2 = (0, 3, 2)'$
- ▶ (d)  $z_1 = (1, 0, 2)', z_2 = (3, 2, 3)', z_3 = (1, 0, 5)', z_4 = (3, 3, 1)'$
- ▶ (e)  $x_1 = (1, 1, 2)', x_2 = (0, -1, 0)', x_3 = (3, -2, 6)'$

## Exercises

- ▶ Find a basis of the following vector spaces and determine their dimensions.
- ▶ (a)

$$V_1 = \{(x, y) | x, y \in \mathbb{R}\}$$

- ▶ (b)

$$V_2 = \{(x, y, z) | x + y + z = 1\}$$

- ▶ (c)

$$V_3 = \{(x, y) | y = 3x\}$$

- ▶ (d)

$$V_4 = \{(x, y, z) | x + y = 1\}$$

## Exercises

- ▶ Determine whether the column vectors of the following matrices are linearly independent. Find the rank of the following matrices.
- ▶ (a)

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

- ▶ (b)

$$B = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 2 & -1 \end{pmatrix}$$

- ▶ (c)

$$C = \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}$$

- ▶ (d)

$$D = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & -1 \end{pmatrix}$$

## Exercises

- (a) Let  $V$  be a vector space and  $U \subset V$ . Show that  $U$  is a subspace of  $V$  if and only if it is closed under addition and scalar multiplication. i.e. for every  $u, v \in U$  and  $c \in \mathbb{R}$ , we have

$$u + v \in U$$

and

$$cu \in U$$

- (b) Let  $L_A : V \rightarrow W$  be a linear map where  $V, W$  are vector spaces. Show that the kernel and image are subspaces of  $V$  and  $W$ , respectively.

## Exercises

- ▶ Let  $A \in M_n$  be invertible and  $b \in \mathbb{R}^n$ . Explain why

$$Ax = b$$

has a unique solution using the following concepts:  $\text{im } A$ ,  $\ker A$

**Hint** First show existence and then show uniqueness.

# SW6. Multiple Linear Regression

## Part 3

Ryu Myeonggyu

May 19, 2025

# Review

- ▶ Let  $A \in M_n$ . The following are equivalent.
  - ▶  $A$  is invertible.
  - ▶  $A$  has trivial kernel.
  - ▶  $\text{im } A = \mathbb{R}^n$
  - ▶ The column vectors  $A_1, \dots, A_n$  are linearly independent.
  - ▶ The column vectors  $A_1, \dots, A_n$  span  $\mathbb{R}^n$ .
  - ▶ Equation  $Ax = 0$  has trivial solution only.
  - ▶ Equation  $Ax = b$  has a unique solution.
  - ▶  $A$  has full rank.

# Determinant

- ▶ We will add another invertibility condition: the determinant.
- ▶ There are many ways to define the determinant.
- ▶ We will use the easiest way.
- ▶ Let

$$A = (a_{ij}) \in M_n.$$

# Determinant

- ▶ Let  $A_{ij}$  be a matrix obtained from  $A$  by deleting the  $i$ -th row and  $j$ -th column.
- ▶ The determinant of  $A$  is given by

$$\det A := \sum_{k=1}^n (-1)^{i+k} a_{ik} \det A_{ik} = \sum_{k=1}^n (-1)^{k+j} a_{kj} \det A_{kj}$$

for  $n \geq 2$ .

- ▶ For  $n = 1$ , we define

$$\det A = a$$

for  $1 \times 1$ -matrix  $A = (a)$ .

- ▶ Remark that the choice of row or column along which you expand is independent of the final result.

**Tip** Choose row or column that contains the most zeros.

# Determinant

- ▶ e.g.



$$A = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix}$$

- ▶  $\det A =$

## Properties of Determinant

- ▶ Let  $A = (a_{ij}) = (A_1, \dots, A_i, \dots, A_j, \dots, A_n)$
- ▶ Let  $B = (A_1, \dots, A_j, \dots, A_i, \dots, A_n)$
- ▶ (a) Then  $\det B = -\det A$ .
- ▶ (b) In particular, if  $A_i = A_j$ , then

$$\det A = 0$$

- ▶ (c) If  $v$  is an  $n$ -vector and  
 $C = (A_1, \dots, aA_i + bv, \dots, A_j, \dots, A_n)$ , then  
 $\det C = a\det A + b\det(A_1, \dots, v, \dots, A_j, \dots, A_n).$
- ▶ (d) For each  $n$ ,

$$\det I_n = 1$$

# Properties of Determinant

**Proof** (a) Use induction on the size of the matrix.

(b) Obvious from (a).

(c),(d) Direct calculation.

**Fact** The properties (a)-(d) actually characterize the determinant function.

# Determinant and Invertibility

- ▶ **Theorem.** For  $A, B \in M_n$ ,

$$\det AB = \det A \det B.$$

- ▶ **Theorem.** Let  $A = (A_1, \dots, A_n)$ .  $A_1, \dots, A_n$  are linearly independent if and only if  $\det A \neq 0$ .

Proof

## Determinant

- ▶ Hence we can add one more invertibility condition.
- ▶  $A$  is invertible  $\iff \det A \neq 0$ .
- ▶ Determinant has more profound meaning especially in geometry.
- ▶ We will not cover this and move on to the projection theory.

## Eigenvalue and Eigenvector

- ▶ Let  $A \in M_n$ .
- ▶ Suppose there exist a nonzero vector  $v \in \mathbb{R}^n$  and  $\lambda \in \mathbb{R}$  such that

$$Av = \lambda v$$

- ▶ Then we say  $v$  is an eigenvector if  $A$  and  $\lambda$  is an eigenvalue of  $A$ .
- ▶ e.g.  $A = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$

# Eigenvalue and Eigenvector

- ▶ How can we find eigenvalues?
- ▶  $A\mathbf{v} = \lambda\mathbf{v} \iff A\mathbf{v} = \lambda I\mathbf{v} \iff (A - \lambda I)\mathbf{v} = 0 \iff \det(A - \lambda I) = 0$
- ▶ e.g.

$$A = \begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix}$$

- ▶ We say  $\phi_A(t) = \det(A - tI)$  is the characteristic polynomial of  $A$ .
- ▶ Eigenvalues are the roots of the characteristic polynomial.

# Eigenvalue and Eigenvector

- ▶ Let  $\lambda_1, \dots, \lambda_n$  be the eigenvalues of  $A$ .
- ▶ Then,

$$\lambda_1 + \dots + \lambda_n = \text{tr}(A)$$

and

$$\lambda_1 \cdots \lambda_n = \det(A)$$

proof

## Eigenvalue and Eigenvector

- ▶ Using eigenvalues and eigenvectors, we can write

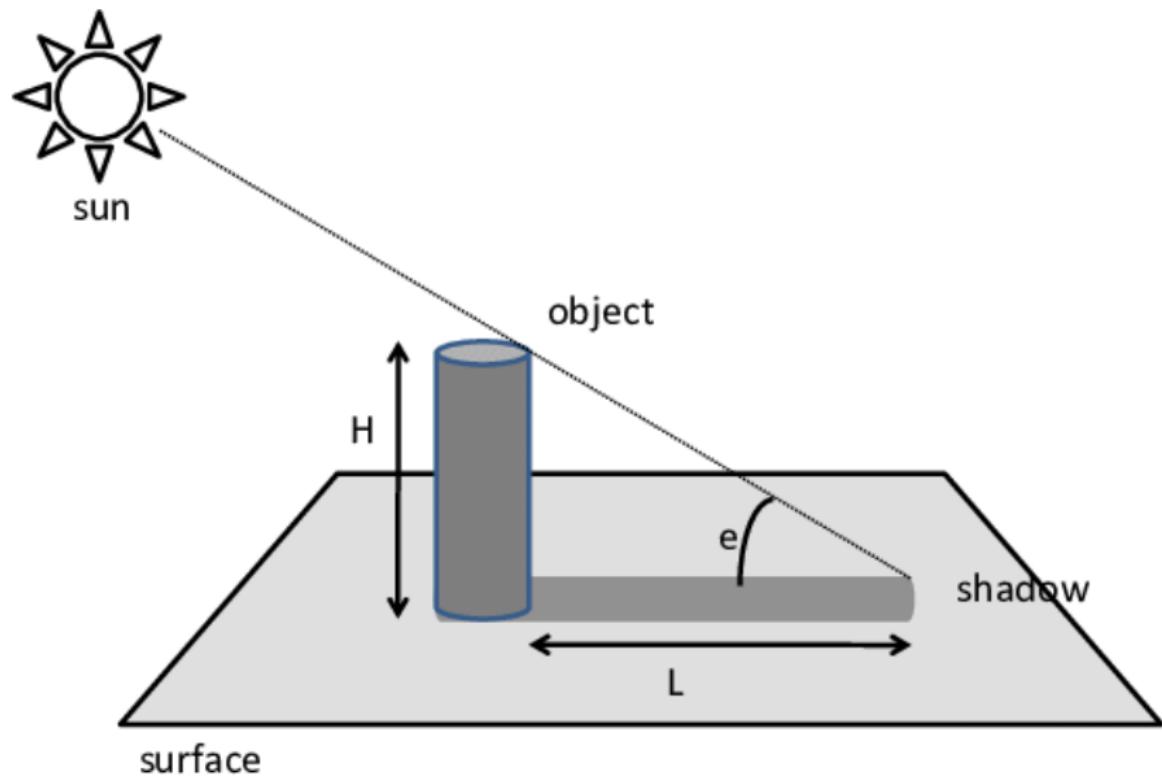
$$A = Q\Lambda Q^{-1}$$

where  $\Lambda$  is a diagonal matrix consisting of eigenvalues of  $A$  and  $P$  is a matrix whose column vectors are eigenvectors of  $A$ , if  $Q$  is invertible. In this case, we say  $A$  is diagonalizable.

- ▶ There are some cases where  $Q$  is invertible, for example when the eigenvalues of  $A$  are distinct.
- ▶ Moreover it is known that if  $A$  is symmetric,  $Q$  satisfies  $Q' = Q^{-1}$ . In this case we say  $Q$  is an orthogonal matrix and  $A$  is orthogonally diagonalizable.
- ▶ It is known that  $A$  is orthogonally diagonalizable if and only if  $A$  is symmetric.

## Inner Product

- ▶ Now, we will start the projection theory which is our final destination.



# Inner Product

- ▶ To understand the projection which is basically a geometric concept, we need to establish a geometric structure on  $\mathbb{R}^n$ .<sup>1</sup>
- ▶ This will be given by the inner product on  $\mathbb{R}^n$ .
- ▶ Let  $x, y \in \mathbb{R}^n$ . The (standard) inner product<sup>2</sup> of  $x = (x_1, \dots, x_n)'$  and  $y = (y_1, \dots, y_n)'$  is

$$\langle x, y \rangle := \sum_{i=1}^n x_i y_i$$

- ▶ e.g.  $x = (1, 2, 3)'$  and  $y = (0, -1, 1)'$

---

<sup>1</sup>Plato, “Let no one untrained in geometry enter.”

<sup>2</sup>or dot product

# Inner Product

- ▶ The followings hold for  $\langle ., . \rangle$
- ▶  $\langle x, x \rangle \geq 0$  for all  $x \in \mathbb{R}^n$ .
- ▶  $\langle x, x \rangle = 0 \iff x = 0$
- ▶  $\langle x, y + cz \rangle = \langle x, y \rangle + c\langle x, z \rangle$
- ▶  $\langle x, y \rangle = \langle y, x \rangle$
- ▶  $\langle x, y \rangle = x'y$

# Inner Product

- ▶ Observe that from the inner product we define the norm

$$\|x\| := \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^n x_i^2}$$

- ▶ This allows us to measure the distance between two points

$$\|x - y\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- ▶ Also we can measure the angle between two vectors

$$\cos \theta = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$$

- ▶ e.g. ( $n=2$ )  $x = (1, 0)'$  and  $y = (1, \sqrt{3})'$

# Inner Product

- ▶ Recall that  $\theta = 90^\circ \iff \cos \theta = 0$ <sup>3</sup>
- ▶ Hence we say two vectors  $x$  and  $y$  are perpendicular (or orthogonal) to each other if  $\langle x, y \rangle = 0$ .
- ▶ In this case we write  $x \perp y$ .

---

<sup>3</sup>here we have  $0^\circ \leq \theta \leq 180^\circ$

# Geometry in $\mathbb{R}^n$

- ▶ **Theorem (Pythagoras).** Let  $x, y \in \mathbb{R}^n$ . If  $x \perp y$ , then

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2$$

Proof

## Transpose

► **Theorem.** Let  $x, y \in \mathbb{R}^n$  and  $A \in M_n$ . Then

$$\langle x, Ay \rangle = \langle A'x, y \rangle$$

**Proof** We prove by direct calculation to see what is happening.  
(shorter proof is exercise)

$$\begin{aligned}\langle x, Ay \rangle &= \langle x, A_1y_1 + \cdots + A_ny_n \rangle \\&= y_1\langle x, A_1 \rangle + \cdots + y_n\langle x, A_n \rangle \\&= y_1(x_1a_{11} + \cdots + x_na_{n1}) + \cdots + y_n(x_1a_{1n} + \cdots + x_na_{nn}) \\&= x_1(y_1a_{11} + \cdots + y_na_{1n}) + \cdots + x_n(y_1a_{n1} + \cdots + y_na_{nn}) \\&= x_1\langle y, A'_1 \rangle + \cdots + x_n\langle y, A'_n \rangle \\&= \langle y, A'_1x_1 \rangle + \cdots + \langle y, A'_nx_n \rangle \\&= \langle y, A'x \rangle = \langle A'x, y \rangle\end{aligned}$$

# Projection

- ▶ We say a square matrix  $P$  is idempotent if  $P^2 = P$
- ▶ An idempotent matrix is called a projection matrix.
- ▶ Moreover if  $P$  is symmetric, i.e.  $P = P'$ , then we say  $P$  is an orthogonal projection.
- ▶ Why is  $P$  called a *projection*?

# Projection

- ▶  $P$  projects vectors of  $\mathbb{R}^n$  to the image of  $P$ .
- ▶  $P$  acts like an identity on the vectors in its image. (this is why  $P$  is called a *projection*)

$$Pw = P(Pv) = P^2v = Pv = w$$

for any  $w \in \text{im } P$ .

- ▶ This allows us to decompose any vector  $v \in \mathbb{R}^n$  by

$$v = \underbrace{Pv}_{\text{part of } v \text{ explained by } P} + \underbrace{(v - Pv)}_{\text{part of } v \text{ unexplained by } P}$$

- ▶ We define  $M = I - P$  so that  $v - Pv = Mv$ .

# Projection

- ▶ Thus

$$v = Pv + Mv$$

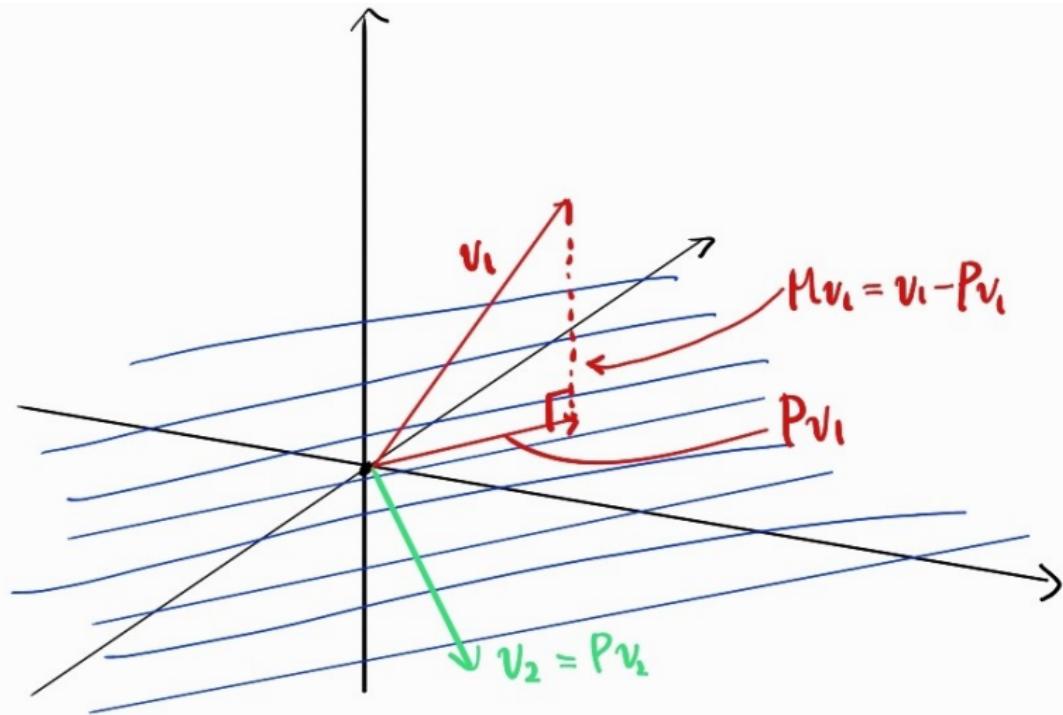
- ▶ Moreover if  $P$  is symmetric, then  $Pv \perp Mv$ .

$$\begin{aligned} (\because) \langle Pv, Mv \rangle &= \langle Pv, v - Pv \rangle \\ &= \langle Pv, v \rangle - \langle Pv, Pv \rangle \\ &= \langle Pv, v \rangle - \langle P'Pv, v \rangle \\ &= \langle Pv, v \rangle - \langle P^2v, v \rangle \\ &= \langle Pv, v \rangle - \langle Pv, v \rangle = 0 \end{aligned}$$

- ▶ In fact, a stronger result holds:  $Pv \perp Mv_0$  for any  $v, v_0 \in \mathbb{R}^n$ .
- ▶ Thus by the Pythagorean theorem, we have

$$\|v\|^2 = \|Pv\|^2 + \|Mv\|^2$$

# Projection



## Closest Vector

- ▶ From this we also have the following theorem.
- ▶ For an orthogonal projection matrix  $P$  and a vector  $v \in \mathbb{R}^n$ , the closest vector  $w$  from  $\text{im } P$  to  $v$  is the projected vector  $Pv$  of  $v$  which is unique.
- ▶ That is,

$$Pv = \arg \min_{w \in \text{im } P} \|v - w\|$$

**Proof** Since  $w \in \text{im } P$ ,  $w = Pv_0$  for some  $v_0 \in \mathbb{R}^n$ . Observe that

$$\begin{aligned}\|v - w\|^2 &= \|v - Pv + Pv - w\|^2 \\&= \|Mv + P(v - v_0)\|^2 \\&= \|Mv\|^2 + \|P(v - v_0)\|^2 \geq \|Mv\|^2 = \|v - Pv\|^2\end{aligned}$$

Note that the equality holds if and only if  $w = Pv$ , which shows the uniqueness.

# Eigenspace Decomposition

- ▶ Let  $A$  be a square matrix with eigenvalue  $\lambda$ .
- ▶ The set of eigenvectors corresponding to  $\lambda$

$$E_\lambda := \{v \in \mathbb{R}^n | Av = \lambda v\}$$

is a subspace of  $\mathbb{R}^n$ , which we call an eigenspace.

- ▶ For a projection  $P$ ,  $P$  has eigenvalues 0 and 1.

$$(\cdot) \lambda v = Pv = P^2 v = P(Pv) = P(\lambda v) = \lambda(Pv) = \lambda^2 v$$

$$\Rightarrow \lambda = \lambda^2$$

$$\therefore \lambda = 1 \text{ or } \lambda = 0.$$

# Eigenspace Decomposition

- ▶ Observe that

$$v = \underbrace{Pv}_{\in \text{im } P} + \underbrace{(v - Pv)}_{\in \ker P}$$

- ▶ Since  $P$  is idempotent,

$$\text{im } P = E_1 \text{ and } \ker P = E_0$$

- ▶ Because of the dimension theorem, we see that  $P$  is diagonalizable.

# Projection and Eigenspace decomposition

- ▶ Moreover if  $P$  is symmetric, then the image and kernel of  $P$  are orthogonal.
- ▶ That is, for each  $w \in \text{im } P$  and  $v \in \ker P$ , we have  $v \perp w$ .
- ▶ This makes  $P$  an orthogonally diagonalizable matrix.

$$P = Q\Lambda Q'$$

where  $\Lambda = \text{diag}(\underbrace{1, 1, \dots, 1}_{\# \text{ of } 1's = \dim \text{im } P}, \underbrace{0, \dots, 0}_{\# \text{ of } 0's = \dim \ker P})$  and

$$Q = \begin{pmatrix} v_1 & \cdots & v_k & v_{k+1} & \cdots & v_n \\ \underbrace{\phantom{v_1}}_{\text{eigenvectors for } \lambda=1} & & \underbrace{\phantom{v_k}}_{\text{eigenvectors for } \lambda=0} & \underbrace{\phantom{v_{k+1}}}_{\text{eigenvectors for } \lambda=0} & & \end{pmatrix}$$

## Exercise

- ▶ Let  $A \in M_n$  and  $r \in \mathbb{R}$ . Prove that

$$\det(rA) = r^n \det(A)$$

- ▶ Let  $D = \text{diag}(d_1, \dots, d_n)$  be a diagonal matrix. What is  $\det D$ ? Find eigenvalues of  $D$ .

## Exercise

- ▶ Provide a one-line proof of the theorem that

$$\langle x, Ay \rangle = \langle A'x, y \rangle$$

**Hint** Use transpose.

## Exercise

- ▶ Calculate the determinant of the following matrices and find eigenvalues.
- ▶ (a)  $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$
- ▶ (b)  $B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}$
- ▶ (c)  $C = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$

## Exercise

- ▶ Calculate the following for  $x = (1, 2, -1)'$  and  $y = (-1, 0, 1)'$ .
- ▶ (a)  $\langle x, y \rangle$
- ▶ (b)  $\|x\|$  and  $\|y\|$
- ▶ (c)  $\cos \theta$
- ▶ (d)  $\det A$  where  $A = (x, y, 2x + y)$

## Exercise

- ▶ For  $x, y \in \mathbb{R}^n$ , prove the following.
- ▶ (a) (Parallelogram identity)

$$\|x - y\|^2 + \|x + y\|^2 = 2(\|x\|^2 + \|y\|^2)$$

- ▶ (b) (Triangle inequality)

$$\|x + y\| \leq \|x\| + \|y\|$$

- ▶ (c) (Cauchy-Schwarz inequality)

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$$

- ▶ (d) (Non-degeneracy) If  $\langle x, z \rangle = 0$  for every  $z \in \mathbb{R}^n$ , then  $x = 0$ .

## Exercise

- ▶ Let  $v = (1, 2)'$ .
- ▶ (a) Construct an orthogonal projection matrix  $P$  such that  $\text{im}P = \text{span}(v)$ .
- ▶ (b) Verify that  $P^2 = P$  and  $P = P'$ .

## Exercise

- ▶ Let  $w = (1, 0, 0)'$  and  $v = (2, 1, -1)'$ .
- ▶ (a) Find the closest vector in  $\text{span}(w)$  to  $v$ .
- ▶ (b) What is the orthogonal projection matrix  $P$  to  $\text{span}(w)$ ?
- ▶ (c) What is the rank of  $P$ ?

## Exercise

- ▶ Let  $P \in M_n$  be a projection matrix.
- ▶ (a) Show that either  $\det(P) = 1$  or  $\det(P) = 0$ .
- ▶ (b) Show that if  $\det(P) = 1$ , then  $P = I$ .

**Hint** Use the fact that  $P$  is diagonalizable.

- ▶ (c) Explain the situation in (b) geometrically.

# SW6. Multiple Linear Regression: Part 4

Ryu Myeonggyu

May 23, 2025

# Multiple Linear Regression

- ▶ We finally begin the multiple linear regression. This will compensate the hard work in linear algebra.
- ▶ Suppose we have a sample  $\{(Y_i, X_{i1}, \dots, X_{ik})\}$ .
- ▶ We assume  $n > k + 1$ .
- ▶ The regression is<sup>1</sup>

$$Y_1 = \beta_0 + X_{11}\beta_1 + \dots + X_{1k}\beta_k + u_1$$

$$Y_2 = \beta_0 + X_{21}\beta_1 + \dots + X_{2k}\beta_k + u_2$$

$$\vdots$$

$$Y_n = \beta_0 + X_{n1}\beta_1 + \dots + X_{nk}\beta_k + u_n$$

---

<sup>1</sup>Is it customary to include constant as regressor, so we include it explicitly following the notation of Stock and Watson.

# Multiple Linear Regression

- With vectors and matrices, we can write this as

$$Y = X\beta + u$$

i.e.

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}_{n \times 1} = \underbrace{\begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}_{n \times (k+1)}}_{(k+1) \times 1} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}_{n \times 1}$$

- We assume  $X$  has full rank.
- We will denote the transposed  $i$ -th row vector of  $X$  by  $X_i$ , so that

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

## Least Squares

- ▶ Now the least squares estimator is the solution to the minimization problem

$$\min_{\beta} \sum_{i=1}^n u_i^2 = u' u = \|u\|^2$$

i.e.

$$\min_{\beta} (Y - X\beta)'(Y - X\beta) = \|Y - X\beta\|^2$$

- ▶ Note that  $X\beta$  is in the image of  $X$ .

# Least Squares

- ▶ After the hard training, we see that the above problem reads:  
“What is the closest vector in the image of  $X$  to  $Y$ ? ”
- ▶ We know that the answer is  $PY$  where  $P$  is the orthogonal projection matrix to the image of  $X$ .
- ▶ What is the orthogonal projection matrix  $P$ ?
- ▶ That is, we want to find a square matrix  $P$  such that
  - (1)  $P' = P$  and  $P^2 = P$
  - (2)  $Pv = v \iff v \in \text{im}X$

# Orthogonal Projection

- ▶ We claim that  $P_X = X(X'X)^{-1}X'$  is the answer.
- ▶ First we need to show that  $X'X \in M_{k+1}$  is invertible.
- ▶ Note that if  $v \in \ker X'X$

$$0 = \langle v, X'Xv \rangle = \langle Xv, Xv \rangle = \|Xv\|^2$$

so that  $Xv = 0$ .  $\therefore v \in \ker X$ .

- ▶ Since  $X$  has full rank,  $X$  has a trivial kernel by the dimension theorem<sup>2</sup>, which implies  $v = 0$ .
- ▶ Hence  $\ker X'X$  is also trivial.

---

<sup>2</sup>here be aware that  $n > k + 1$ .

# Orthogonal Projection

- ▶ Thus we have

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

- ▶ Now we claim that  $\hat{\beta}$  is unique. i.e. if  $X\alpha = P_X Y$ , then  $\alpha = \hat{\beta}$ .
- ▶ Since  $X$  has trivial kernel, we see that

$$X\alpha = X\hat{\beta} \Rightarrow X(\alpha - \hat{\beta}) = 0 \Rightarrow \alpha = \hat{\beta}.$$

- ▶ Hence by the uniqueness of  $\hat{\beta}$ , we have

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

as our OLS estimator of  $\beta$ .

## Orthogonal Projection

- ▶ Observe that  $\hat{\beta}$  is given by the equation

$$(X'X)\hat{\beta} = X'Y$$

- ▶ The above equation is called the **normal equation**.
- ▶ The above equation is equivalent to

$$X'(Y - X\hat{\beta}) = 0$$

- ▶ That is, for each column  $X^j$  of  $X$ , we have

$$(X^j)'(Y - X\hat{\beta}) = \langle X^j, Y - X\hat{\beta} \rangle = 0 \Rightarrow X^j \perp (Y - X\hat{\beta})$$

- ▶ The orthogonality condition characterizes the OLS estimator of  $\beta$ .
- ▶ Isn't it remarkable that we solved the optimization problem without using differentiation?

## The OLS estimator

- We remember that we also have the decomposition

$$Y = P_X Y + Y - P_X Y = X \hat{\beta} + \underbrace{(I - P_X)}_{=:M_X} Y$$

- The  $M_X$  matrix is called the annihilator matrix of  $X$ .
- Here we see that

$$\hat{u} = M_X Y$$

- As usual we write

$$X \hat{\beta} = \hat{Y}$$

so that

$$Y = \hat{Y} + \hat{u}$$

## Properties of OLS residual

- ▶ The followings are true.
- ▶ (a)  $\sum_{i=1}^n \hat{u}_i = 0$
- ▶ (b)  $X' \hat{u} = 0$
- ▶ (c)  $\sum_{i=1}^n \hat{Y}_i = \sum_{i=1}^n Y_i$

## Measures of Fit

- ▶ As usual,

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$RSS = \sum_{i=1}^n \hat{u}_i^2 = \|\hat{u}\|^2$$

and

$$R^2 = 1 - \frac{RSS}{TSS}$$

## adjusted $R^2$

- ▶ We see that  $R^2$  always increases as the number of regressors increases. (proof is left as an exercise)
- ▶ One can include infinitely many meaningless regressors to make  $R^2$  bigger.
- ▶ Correction to this: the adjusted  $R^2$  ( $= \bar{R}^2$  or adj. $R^2$ )

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

- ▶ Compensate the decrease of  $RSS$  by  $\frac{1}{n-k-1}$ .

# The Least Square Assumptions for Multiple Linear Regression

- ▶ Now we establish the multiple linear regression model.
- ▶ (A1) The conditional distribution of  $u_i$  given  $X_{1i}, \dots, X_{ki}$  has mean zero

$$\mathbb{E}[u_i | X_{1i}, \dots, X_{ki}] = 0$$

- ▶ (A2)  $(Y_i, X_{1i}, \dots, X_{ki})$  are i.i.d.
- ▶ (A3) Large outliers are unlikely.

$$\mathbb{E}[X_{ji}^4], \mathbb{E}[Y_i^4] < \infty$$

- ▶ (A4) No perfect multicollinearity. i.e.  $X$  has full rank.
- ▶ Under the assumptions (A1)-(A4), the OLS coefficient  $\hat{\beta}$  has a causal interpretation.

## The Interpretation of OLS coefficient

- ▶ Under the assumptions (A1)-(A4),  $\hat{\beta}_1$  is interpreted as
- ▶ “the expected change in  $Y_i$  caused by one unit change in  $X_1$ , holding other variables  $X_2, \dots, X_k$  constant.”
- ▶ In multiple regression, the interpretation of each coefficient is isolated.
- ▶ This means we are pretending that all other variables do not change when we consider how one variable affects the outcome.
- ▶ “ceteris paribus”

## Sampling Distribution of the OLS estimator

- ▶ Under the assumptions (A1)-(A4), we will derive the sampling distribution of  $\hat{\beta}$  given  $X$ .
- ▶ Before, we need to know the mean and variance of random vectors.
- ▶ Random vectors are a vector of random variables.
- ▶ Let  $Z = (z_1, \dots, z_k)'$  be a random vector.
- ▶ The expectation or the mean vector of  $Z$  is

$$\mathbb{E}[Z] = \begin{pmatrix} \mathbb{E}[z_1] \\ \vdots \\ \mathbb{E}[z_k] \end{pmatrix}$$

- ▶ The (co)variance matrix of  $Z$  whose mean vector is  $\mu_Z$  is

$$Var(Z) = \mathbb{E}[(Z - \mu_Z)(Z - \mu_Z)'] = (Cov(z_i, z_j))_{ij}$$

## Sampling Distribution of the OLS estimator

- ▶ **Lemma.** Let  $Z$  be a vector of random variables with mean vector  $\mu_Z$ . For a nonrandom matrix  $A$  and a nonrandom vector  $b$ , we have

$$\mathbb{E}[AZ + b] = A\mathbb{E}[Z] + b = A\mu_Z + b$$

and

$$\text{Var}(AZ + b) = A\text{Var}(Z)A'$$

- ▶ It is useful to notice that

$$\hat{\beta} = \beta + (X'X)^{-1}X'u$$

Proof

# Sampling Distribution of the OLS estimator

- ▶ **Theorem.** The followings hold under the assumptions (A1)-(A4).
  - ▶ (a) (unbiasedness)

$$\mathbb{E}[\hat{\beta}|X] = \beta$$

- ▶ (b) (conditional variance)

$$Var(\hat{\beta}|X) = (X'X)^{-1}X'Var(u|X)X(X'X)^{-1}$$

- ▶ Under homoskedasticity, i.e.  $Var(u|X) = \sigma^2 I$ , (b) reduces to

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

(exercise)

# Proof

## Gauss-Markov theorem

- ▶ A classical justification of the use of OLS.
- ▶ Assume (A1)-(A4) and homoskedasticity

$$\text{Var}(u|X) = \sigma^2 I.$$

- ▶ Then the OLS estimator  $\hat{\beta}$  is the best linear unbiased estimator. (**BLUE**)
- ▶ “Best” means the smallest variance.
- ▶ “Linear” means the estimator is linear in  $Y$ .
- ▶ “Unbiased” means

$$\mathbb{E}[\hat{\beta}] = \beta.$$

# Multicollinearity

- ▶ Perfect multicollinearity means the  $X$  matrix is not of full rank.
- ▶ In this case, the OLS formula becomes useless.
- ▶ Mathematically, there exist infinitely many solutions to the least squares problem.
- ▶ Hence we need to avoid the perfect multicollinearity.
- ▶ Perfect multicollinearity often occurs when you include dummy variables.
- ▶ e.g.

$$\text{Income}_i = \beta_0 + \beta_1 \cdot \text{Female}_i + \beta_2 \cdot \text{Male}_i + u_i$$

## Imperfect Multicollinearity

- ▶ Imperfect multicollinearity means that two or more of the regressors are highly correlated.
- ▶ Imperfect multicollinearity **does not pose any problems** for the theory of the OLS estimators.
- ▶ The OLS estimator of the coefficient will have a larger variance, just as your data is small.<sup>3</sup>
- ▶ In general, when multiple regressors are imperfectly multicollinear, the coefficients on one or more of these regressors will be imprecisely estimated in that they will have a large sampling variance.
- ▶ Imperfect multicollinearity is not necessarily an error but rather just a feature of OLS, your data, and the question you are trying to answer.
- ▶ In most cases modern econometricians do not care much about imperfect multicollinearity.

---

<sup>3</sup><https://www.econlib.org/archives/2005/09/multicollinearit.html>



## Control Variables

- ▶ A control variable (or covariate) is not the object of interest in the study.
- ▶ It is a regressor included to hold constant factors that, if neglected, could lead the estimated causal effect of interest to suffer from omitted variable bias.
- ▶ Control variables are needed for technical reasons to avoid the omitted variable bias, but we are not interested in those variables.
- ▶ make the variables of interest no longer correlated with the error term, once the control variables are held constant.

## Control Variables and Least Squares Assumptions

- ▶ Since we are not interested in the coefficients of control variables, the least squares assumptions are relaxed.
- ▶ Our regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

where  $\beta_1, \dots, \beta_k$  are causal effects and  $W$ 's are control variables.

- ▶ (A1)  $u_i$  has a conditional mean that does not depend on the  $X$ 's given the  $W$ 's

$$\mathbb{E}[u_i | X_{1i}, \dots, X_{ki}, W_{1i}, \dots, W_{ri}] = 0$$

- ▶ (A2)  $(X_{1i}, \dots, X_{ki}, W_{1i}, \dots, W_{ri}, Y_i)$  are i.i.d.
- ▶ (A3) Large outliers are unlikely:  $X_{1i}, \dots, X_{ki}, W_{1i}, \dots, W_{ri}, Y_i$  has nonzero finite 4th moment.
- ▶ (A4) No perfect multicollinearity.

## Frisch-Waugh-Lovell theorem

- ▶ Since we are not interested in the coefficients of  $W$ 's, we don't need to estimate  $(\beta_{k+1}, \dots, \beta_{k+r}) =: \gamma'$ .
- ▶ Our regression model is

$$Y = X\hat{\beta} + W\hat{\gamma} + \hat{u}$$

- ▶ The Frisch-Waugh-Lovell (FWL) theorem states that you can estimate  $\beta$  by the following regressions:

Step 1 Regress  $X$  on  $W$  and get residual  $\hat{e}$ .

$$X = W\hat{\alpha} + \hat{e}$$

## FWL theorem

Step 2 Regress  $Y$  on  $W$  and get residual  $\tilde{e}$ .

$$Y = W\tilde{\alpha} + \tilde{e}$$

Step 3 Regress  $\tilde{e}$  on  $\hat{e}$  and this gives you  $\hat{\beta}$  and  $\hat{u}$ .

$$\tilde{e} = \hat{e}\hat{\beta} + \hat{u}$$

- ▶ This is called the residual regression.

Proof On step 1, we see that

$$X = P_W X + M_W X$$

On step 2, similarly

$$Y = P_W Y + M_W Y$$

## FWL theorem

- ▶ On step 3, we are regressing  $M_W Y$  on  $M_W X$ .
- ▶ On the other hand, multiplying  $M_W$  to our original regression, we have

$$M_W Y = M_W X \hat{\beta} + M_W W \hat{\gamma} + M_W \hat{u}$$

where  $M_W W = W - P_W W = W - W = 0$  so that

$$M_W Y = M_W X \hat{\beta} + M_W \hat{u}$$

## FWL theorem

- ▶ Also recall that  $W'\hat{u} = 0$ , so that

$$P_W \hat{u} = 0$$

- ▶ This implies that  $M_W \hat{u} = \hat{u}$ , which further implies

$$M_W Y = M_W X \hat{\beta} + \hat{u}$$

Here observe that

$$(M_W X)' \hat{u} = X' M_W \hat{u} = X' \hat{u} = 0$$

hence our  $\hat{\beta}$  is indeed the OLS coefficient from the regression of  $M_W Y$  on  $M_W X$ . This completes the proof.

## Remarks on FWL theorem

- ▶ From the OLS formula, we have

$$\hat{\beta} = (X' M_W X)^{-1} (X' M_W Y)$$

- ▶ This means your OLS coefficients of  $X$  **partials out** the effects of  $W$  on  $X$  and  $Y$ .
- ▶ By multiplying  $M_W$ , you eliminate the part of  $X$  and  $Y$  explained by  $W$ , and then you obtain your OLS coefficients.
- ▶ From this it becomes clear that the multiple linear regression coefficients measure the expected change in  $Y$  by  $X$ , **holding other factors constant**.
- ▶ In fact, FWL theorem can be applied to various situations, not just to suppress the control variables. Examples of its powerful applications are in the exercises.

## Next lecture

- ▶ In the next lecture, we will cover the theory of standard errors.
- ▶ And then we study the statistical hypothesis testing.
- ▶ This will require some understanding in asymptotics and (multivariate) normal distribution.
- ▶ I will try to explain the spirit and not to use too much mathematics.

## Exercise

- ▶ Using the matrix version of OLS formula, show that it is equivalent to the OLS formula

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

when  $k = 1$ .

## Exercise

- ▶ Let  $\mathbb{1} = (1, 1, \dots, 1)' \in \mathbb{R}^n$ . Consider the usual multiple linear regression

$$Y = X\beta + u.$$

- ▶ (a) For  $Y$ , calculate  $P_{\mathbb{1}} Y$ . What is  $M_{\mathbb{1}} Y$ ?
- ▶ (b) Calculate

$$\|M_{\mathbb{1}} Y\|^2.$$

What is its name?

- ▶ (c) Calculate

$$1 - (Y' M_{\mathbb{1}} Y)^{-1} (Y' M_X Y)$$

What is this number called?

- ▶ (d) Interpret the meaning of the number calculated at (c).

**Hint** It is equal to  $(Y' M_{\mathbb{1}} Y)^{-1} (Y' M_{\mathbb{1}} Y - Y' M_X Y)$

## Exercise

- ▶ (a) Prove that RSS is non-decreasing as the number of regressors increases.
- ▶ (b) Prove that under homoskedasticity, the conditional variance of  $\hat{\beta}$  is given by

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

- ▶ (c) If  $X$  has full rank, what is the trace of the projection matrix  $P_X$ ?
- ▶ (d) Prove that  $M_X$  is symmetric and idempotent.
- ▶ (e) Consider the simple linear regression. Instead of  $Y_i$  and  $X_i$ , you estimate the slope coefficient  $\beta_1$  using the demeaned variables  $\tilde{Y}_i = Y_i - \bar{Y}$  and  $\tilde{X}_i = X_i - \bar{X}$ . Will your  $\hat{\beta}_1$  be different? What about the  $R^2$ ?

**Hint** Previous exercise and FWL theorem.

## Exercise

- ▶ Download the dataset Birthweight and Smoking from  
[https://www.princeton.edu/~mwatson/Stock-Watson\\_3u/Students/](https://www.princeton.edu/~mwatson/Stock-Watson_3u/Students/)  
Stock-Watson-EmpiricalExercises-DataSets.htm  
(6th from the top)
- ▶ You should solve the problems below with R **without using any regression packages.**
- ▶ (a) Using the dataset, replicate the table on the next page.  
You can use chatGPT, which is strongly not recommended.
- ▶ (b) Briefly interpret the result.
- ▶ (c) Read the description file and explain why you cannot include tripred1 as regressor in column (3).
- ▶ (d) In column (2), calculate the coefficient of smoker using the residual regression.

## Exercise

	(1)	(2)	(3)
	birthweight	birthweight	birthweight
smoker	-253.2284	-217.5801	-228.8476
alcohol		-30.49129	-15.09998
nprevist		34.06991	
tripre0			-697.9687
tripre2			-100.8373
tripre3			-136.9553
_cons	3432.06	3051.249	3454.549
<i>N</i>	3000	3000	3000
<i>R</i> <sup>2</sup>	0.029	0.073	0.046
adj. <i>R</i> <sup>2</sup>	0.028	0.072	0.045

# SW7. Hypothesis Tests and Confidence Intervals

Ryu Myeonggyu

June 13, 2025

# Standard Error

- ▶ Standard error is a measure of uncertainty.
- ▶ Consider the multiple linear regression

$$Y = X\beta + u$$

Recall

$$\hat{\beta} = (X'X)^{-1}(X'Y) = \beta + (X'X)^{-1}(X'u)$$

and

$$V_{\hat{\beta}} := Var(\hat{\beta}|X) = (X'X)^{-1}(X'Var(u|X)X)(X'X)^{-1}$$

## Standard Error

- ▶ Under homoskedasticity,

$$\text{Var}(u|X) = \sigma^2 I$$

so that

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

- ▶ Since  $\sigma^2$  is unknown, we estimate it by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2$$

or its bias-corrected version

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n \hat{u}_i^2$$

## Heteroskedasticity-robust Standard Error

- ▶ Under heteroskedasticity,

$$\text{Var}(u|X) = \mathbb{E}[uu'|X]$$

is a diagonal matrix under iid assumption.

- ▶ Thus

$$X' \text{Var}(u|X) X = \sum_{i=1}^n X_i X'_i \mathbb{E}[u_i^2 | X]$$

so that

$$V_{\hat{\beta}} = (X'X)^{-1} \left( \sum_{i=1}^n X_i X'_i \mathbb{E}[u_i^2 | X] \right) (X'X)^{-1}$$

## Heteroskedasticity-robust Standard Error

- ▶ Hence its estimator

$$\hat{V}_{\beta}^{HC0} := (X'X)^{-1} \left( \sum_{i=1}^n X_i X_i' \hat{\epsilon}_i^2 \right) (X'X)^{-1}$$

is called the heteroskedasticity-robust standard error (HC0).

- ▶ Another version is HC1, which accounts for finite-sample bias by degrees of freedom correction

$$\hat{V}_{\beta}^{HC1} := (X'X)^{-1} \left( \frac{n}{n-k} \sum_{i=1}^n X_i X_i' \hat{\epsilon}_i^2 \right) (X'X)^{-1}$$

## Clustered Standard Error

- ▶ Recall that we used the iid assumption ( $\text{Var}(u|X)$  is diagonal matrix) to derive robust standard errors.
- ▶ In empirical research, there are many cases where this assumption is not plausible.

e.g. Duflo, Dupas,

Kremer (2011) <https://www.aeaweb.org/articles?id=10.1257/aer.101>

- ▶ Investigate the impact of tracking on educational attainment in a randomized experiment in Kenya.

$$\text{TestScore}_{ig} = -0.071 + 0.138 \cdot \text{Tracking}_{ig} + u_{ig}$$

where  $i = \text{student}$ ,  $g = \text{school}$ .

- ▶ Does iid assumption make sense for two students from the same school?

## Clustered Standard Error

- ▶ Suppose we have a sample  $\{(Y_{ig}, X_{ig})\}$ .
- ▶  $g = 1, \dots, G$  indexes the cluster and  $i = 1, \dots, n_g$  indexes the number of observations in cluster  $g$ .
- ▶ Total number of observations =  $n = \sum_{g=1}^G n_g$ .
- ▶ Observations are not independent within cluster, but independent across clusters.
- ▶

$$Y_{ig} = X'_{ig}\beta + u_{ig}, \mathbb{E}[u_{ig} | \mathbf{X}_g] = 0$$

or in the cluster notation,

$$\mathbf{Y}_g = \mathbf{X}_g\beta + \mathbf{u}_g, \mathbb{E}[\mathbf{u}_g | \mathbf{X}_g] = 0.$$

## Clustered Standard Error

- ▶ Note that

$$\begin{aligned}\hat{\beta} &= \left( \sum_{g=1}^G \sum_{i=1}^{n_g} X_{ig} X'_{ig} \right)^{-1} \left( \sum_{g=1}^G \sum_{i=1}^{n_g} X_{ig} Y_{ig} \right) \\ &= \left( \sum_{g=1}^G \mathbf{X}_g \mathbf{X}'_g \right)^{-1} \left( \sum_{g=1}^G \mathbf{X}_g \mathbf{Y}_g \right) \\ &= (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{Y})\end{aligned}$$

- ▶ From this we see that  $\mathbb{E}[\hat{\beta} | \mathbf{X}] = \beta$ .

## Clustered Standard Error

- ▶ The covariance matrix is

$$Var(\hat{\beta}|X) = (\mathbf{X}'\mathbf{X})^{-1} \left( \sum_{g=1}^G \mathbf{X}_g \mathbb{E}[\mathbf{u}_g \mathbf{u}_g' | \mathbf{X}_g] \mathbf{X}_g' \right) (\mathbf{X}'\mathbf{X})^{-1}$$

- ▶ Under the clustered setting, the error covariance matrix is not diagonal, but it is block diagonal. i.e.

$$Var(u|\mathbf{X}) = \begin{pmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_G \end{pmatrix}$$

where  $\Sigma_g = Var(\mathbf{u}_g | \mathbf{X}_g)$  is not a diagonal matrix in general.

- ▶ If  $\Sigma_g$  is diagonal for all  $g = 1, \dots, G$ , then this reduces to the heteroskedasticity setup.

## Clustered Standard Error

- ▶ The clustered covariance matrix estimator (Arellano, 1987)

$$\hat{V}_{\hat{\beta}}^{\text{cluster}} := a_n (\mathbf{X}' \mathbf{X})^{-1} \left( \sum_{i=1}^G \mathbf{X}'_g \hat{u}_g \hat{u}'_g \mathbf{X}_g \right) (\mathbf{X}' \mathbf{X})^{-1}$$

where  $a_n = \frac{n-1}{n-k} \frac{G}{G-1}$  is a finite-sample adjustment, where  $a_n \rightarrow 1$  as  $G \rightarrow \infty$  under some regular conditions.

- ▶ This allows a robust estimation relaxing the iid assumption.

## Standard Error

- ▶ Let  $\hat{\beta}_j$  be the  $j$ -th element of  $\hat{\beta}$ .
- ▶ Having estimated the covariance matrix by  $\hat{V}_{\hat{\beta}}$ , the standard error of  $\hat{\beta}_j$  is given by

$$s(\hat{\beta}_j) := \sqrt{(\hat{V}_{\hat{\beta}})_{jj}}$$

- ▶ It is an estimate of the standard deviation of the distribution of  $\hat{\beta}_j$ .
- ▶ But why do we have to care about the standard errors?
- ▶ This is because we are doing an inference using a sample.
- ▶ There is always a possibility that we are wrong.
- ▶ Instead of not doing anything, we report the probability that we might be wrong.

## Standard Error

- ▶ How can we calculate such probability?
- ▶ Notice that we did not make any assumptions on the distributions of  $X_i$  or  $Y_i$ , so we cannot calculate anything.
- ▶ But using asymptotics, we can find the well-known distribution if the sample is large enough.
- ▶ First we will study the process of statistical decision making.

# Hypothesis test

- ▶ We begin with a motivating example.
- ▶ Consider a coin toss.
- ▶ You want to know whether the coin is fair or not. (i.e. the probabilities for head and tail are equal)
- ▶ How can we test if the coin is fair?

## Hypothesis test

- ▶ To test this, you actually toss the coin.
- ▶ After tossing  $n$  times, define

$$X_i = \begin{cases} 1 & \text{if } i\text{-th coin is head} \\ 0 & \text{if } i\text{-th coin is tail} \end{cases}$$

- ▶ If the coin was a fair coin, then we know that

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = 0) = \frac{1}{2}$$

- ▶ Note that

$$Y = \sum_{i=1}^n X_i = \text{number of heads}$$

## Hypothesis test

Fact  $Y \sim \text{Binomial}(n, \frac{1}{2})$

- ▶ From this distribution, we can actually calculate the probability that you get your data.
- ▶ For example, suppose you tossed 10 times but you have only 2 heads.
- ▶ Then  $Y \sim \text{Binomial}(10, \frac{1}{2})$  so

$$\mathbb{P}(Y \leq 2) = \sum_{k=1}^2 \binom{10}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{10-k} = \frac{55}{1024} \approx 0.054$$

- ▶ That is, by tossing a fair coin 10 times, the probability that the number of heads are no more than two is about 5.4%.

## Hypothesis test

- ▶ This is quite small probability to justify the belief that your coin is fair.
- ▶ This is the statistical decision making rule.
- ▶ Based on a certain hypothesis, calculate the likelihood using the data.
- ▶ This hypothesis is called the **null hypothesis**. ( $H_0$ ) The alternative hypothesis ( $H_1$ ) is what is believed to be true instead of  $H_0$ .
- ▶ In statistical decision making, there are **two** sources of errors.
  1. Reject the null hypothesis if the null hypothesis is correct.
  2. Accept the null hypothesis if the null hypothesis is wrong.
- ▶ The first kind is called the **Type I error** and the latter is called the **Type II error**.

# Hypothesis test

- e.g. Consider a convict in a trial.
- ▶ The null hypothesis is that he is innocent.
  - ▶ The alternative hypothesis is that he is guilty.
  - ▶ Type I error =
  - ▶ Type II error =
  - ▶ In most cases, Type I error is considered to be more important than Type II error.
  - ▶ We call the probability of making a Type I error the **size** of the test.
  - ▶ Conversely, we call the (maximal) probability of rejecting the false null the **power** of the test.
  - ▶ In general, there is a trade-off between the size and power.

# Hypothesis test

- ▶ The procedure of hypothesis test:

- Step 1 Set the null hypothesis and the alternative hypothesis.
- Step 2 Set the significance level(=maximum probability of Type I error.)
- Step 3 Construct a test statistic.
- Step 4 Derive the distribution of the test statistic under the null hypothesis.
- Step 5 Considering the significance level, set the critical region of the test.
- Step 6 Using the data, calculate the value of test statistic and decide whether to reject or accept the null hypothesis.

# Well-known distributions

- ▶ To control for the size of the test, we need to know some distributions.
- ▶ Here are some well-known distributions which are frequently used in hypothesis tests.

**Normal** A random variable  $X$  is normal if its density is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

cf. Multivariate normal

$$p(x) = \frac{1}{(2\pi)^{k/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)' \Sigma^{-1} (x - \mu)\right)$$

## Well-known distributions

- $\chi^2$  If  $X = Z_1^2 + \cdots + Z_k^2$  where  $Z_i$ 's are iid standard normal, then we say  $X$  follows  $\chi^2$ -distribution of degrees of freedom  $k$  and write  $X \sim \chi_k^2$
- $F$  If  $X_1 \sim \chi_t^2$  and  $X_2 \sim \chi_s^2$  are independent, we say

$$X = \frac{X_1/t}{X_2/s}$$

follows an  $F$ -distribution and write  $X \sim F(t, s)$ .

- $t$  If  $Z \sim N(0, 1)$  and  $Y \sim \chi_k^2$  are independent, we say

$$X = \frac{Z}{\sqrt{Y/k}}$$

follows (Student's) t-distribution of degrees of freedom  $k$ .

## Well-known distributions

- ▶ In case where  $X$  is discrete,

**Bernoulli** If

$$X = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1 - p \end{cases}$$

then we say  $X \sim \text{Bernoulli}(p)$ .

**Binomial** If  $X = X_1 + \cdots + X_n$  where  $X_i \stackrel{iid}{\sim} \text{Bernoulli}(p)$  then we say

$$X \sim \text{Binomial}(n, p).$$

**Poisson** If  $X$  has a density

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for  $x = 0, 1, 2, \dots$  and  $\lambda > 0$  then we say  $X \sim \text{Poisson}(\lambda)$ .

## Example

- ▶ We test if the coin is fair. We denote by  $p$  the probability of head.
- ▶  $H_0$ : the coin is fair. vs.  $H_1$ : the coin is unfair.  
i.e.  $H_0: p = 0.5$  vs  $H_1: p \neq 0.5$  (two-sided test)
- ▶ Define  $X_i = 1$  if  $i$ -th coin toss was head, otherwise 0.
- ▶ Under the null,  $X_i \stackrel{iid}{\sim} Bernoulli(0.5)$  and

$$Y := \sum_{i=1}^n X_i \sim Binomial(n, 0.5)$$

- ▶ Suppose  $n = 10$  and you want your test size < 5%.

## Example

- ▶ Recall that under the null,

$$\mathbb{P}(Y \leq 1) = \binom{10}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^9 = \frac{10}{1024} \approx 0.01$$

- ▶ By symmetry,

$$\mathbb{P}(Y = 1 \text{ or } Y = 10) \approx 0.02 < 5\%$$

- ▶ If you have three heads, you think this could happen by luck even if  $p = 0.5$  is true.
- ▶ However, if you have only one head or one tail, you think this is too unlikely to happen if  $p = 0.5$  was true.
- ▶ Your standard of “too unlikely” was set as “less than 5% of chance”.

## *t*-test

- ▶ *t*-test is the most frequently used test.
- ▶ In the simple regression analysis, the null hypothesis that we care the most is

$$\beta_1 = 0$$

i.e.  $X$  has no effect on  $Y$ .

- ▶ Under the null, the  $t$ -statistic

$$t := \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \stackrel{A}{\sim} N(0, 1)$$

where  $s(\hat{\beta}_1)$  is the standard error.

## Asymptotic Standard Error

- Standard error of  $\hat{\beta}_1$  comes from the sampling distribution.

Recall

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

- 

$$\sqrt{n}(\hat{\beta}_1 - \beta) \xrightarrow{d} N(0, V_\beta)$$

where

$$V_\beta = \frac{Var((X_i - \mu_X)u_i)}{Var(X_i)^2} = \frac{\mathbb{E}[(X_i - \mu_X)^2 u_i^2]}{\mathbb{E}[(X_i - \mu_X)^2]^2}$$

so that

$$\hat{\beta}_1 \xrightarrow{A} N(\beta_1, V_\beta/n)$$

## Asymptotic Standard Error

- ▶ Under the null hypothesis that  $\beta_1 = 0$ ,

$$\hat{\beta}_1 \xrightarrow{A} N(0, V_\beta/n)$$

- ▶ Since we cannot compute  $\mathbb{E}$ , we replace it by  $\frac{1}{n} \sum$  to have an sample analogue estimator

$$\hat{V}_\beta^{HC0} := \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- ▶ Remark that

$$\hat{V}_\beta^{HC0} \xrightarrow{P} V_\beta$$

and

$$\hat{V}_\beta^{HC0} = n \cdot \hat{V}_{\hat{\beta}}^{HC0}.$$

## Asymptotic Standard Error

- ▶ Thus under the null, for large  $n$ ,

$$t := \frac{\hat{\beta}_1}{\sqrt{\hat{V}_{\hat{\beta}}^{HC0}}} \stackrel{A}{\sim} N(0, 1)$$

- ▶ Since we know the normal distribution, we can use  $t$ -statistic for hypothesis testing.
- ▶ e.g.

## Confidence Interval

- ▶ Now using this asymptotic distribution, the critical region for two-sided  $t$ -test is

$$\{|t| \geq z_{\alpha/2}\}$$

where  $z_{\alpha/2}$  is such that  $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ .

- ▶ The  $(1 - \alpha)$ -confidence region for  $\hat{\beta}_1$  is

$$[\hat{\beta}_1 - z_{\alpha/2} s(\hat{\beta}_1), \hat{\beta}_1 + z_{\alpha/2} s(\hat{\beta}_1)]$$

e.g. If  $\alpha = 5\%$ , then

$$[\hat{\beta}_1 - 1.96 s(\hat{\beta}_1), \hat{\beta}_1 + 1.96 s(\hat{\beta}_1)]$$

as usual.

## p-value

- ▶ p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.
- ▶ In our  $t$ -test situation, the (asymptotic) p-value under the null is

$$\mathbb{P}(|t| > |t(X)|)$$

with  $t \stackrel{A}{\sim} N(0, 1)$  and  $t(X)$  is the observed value of  $t$ -statistic.

## Multiple Linear Regression

- ▶ In multiple linear regression case, similar result holds but everything holds in vectors and matrices.
- ▶ Recall that

$$\hat{\beta} = \beta + (X'X)^{-1}(X'Y)$$

- ▶ Thus by CLT,

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

where

$$V_\beta = \mathbb{E}[X_i X'_i]^{-1} \mathbb{E}[X_i X'_i e_i^2] \mathbb{E}[X_i X'_i]^{-1} = Q_{XX}^{-1} \Omega Q_{XX}^{-1}$$

# Multiple Linear Regression

- ▶ Under some regularity conditions, heteroskedasticity robust standard error can be estimated by

$$\hat{V}_{\beta}^{HC0} := \left( \frac{1}{n} X' X \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n X_i X_i' e_i^2 \right) \left( \frac{1}{n} X' X \right)^{-1}$$

- ▶ Remark that

$$\hat{V}_{\beta}^{HC0} \xrightarrow{P} V_{\beta}$$

and

$$\hat{V}_{\beta}^{HC0} = n \cdot \hat{V}_{\hat{\beta}}^{HC0}.$$

- ▶ This justifies the use of  $t$ -test for regression analysis.

## Test of Joint Hypothesis

- ▶ We can use  $t$ -test to test if  $\beta_j = c$  for scalar.
- ▶ Can we do similar test for vectors or joint hypotheses?

e.g.  $\beta_1 = \beta_2 = 0$

- ▶ In general, this does not give the intended test size.
- ▶ For simplicity, assume  $t_1$  and  $t_2$  are independent, where

$$t_j = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)}$$

- ▶ Then the test size is

$$\begin{aligned}\mathbb{P}(\text{Reject } H_0 | H_0 \text{ is true}) &= 1 - \mathbb{P}(|t_1| < 1.96, |t_2| < 1.96) \\ &= 1 - \mathbb{P}(|t_1| < 1.96)(\mathbb{P}|t_2| < 1.96) \\ &= 1 - 0.95^2 = 0.0975 > 0.05\end{aligned}$$

## Test of Joint Hypothesis

- ▶ So the actual test size is bigger than 5%.
- ▶ We use  $F$ -test to avoid this size distortion.
- ▶ Suppose our null hypothesis is

$$R\beta = r$$

where  $R$  is a  $q \times k$  matrix,  $\beta$  is  $k \times 1$  vector and  $r$  is  $q \times 1$  vector.

- ▶ Assume  $R$  has full rank.

## F test

- ▶ Calculate the restricted OLS residual under the null<sup>1</sup>

$$\tilde{u}_i = Y_i - X\tilde{\beta}$$

and the unrestricted OLS residual  $\hat{u}_i$ .

- ▶ Calculate

$$F = \frac{n - k}{q} \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}$$

where

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2$$

and

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \tilde{u}_i^2$$

---

<sup>1</sup>Technically, CLS (constrained least squares) estimator is used

## F test

- ▶ Then

$$F \sim F(q, n - k)$$

- ▶ Also it is known that

$$q \cdot F \stackrel{A}{\sim} \chi_q^2$$

- ▶ Recall that  $\hat{\sigma}^2 = \text{RSS}$  in the unrestricted model and  $\tilde{\sigma}^2 = \text{RSS}$  in the restricted model.
- ▶ Hence we may write  $F$  as

$$F = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n - k)}$$

## F test

- ▶ In particular, consider the case

$$\beta = 0$$

- ▶ Then the F-statistic is

$$F = \frac{TSS - RSS}{RSS} \frac{k - 1}{n - k}$$

- ▶ Plugging in  $R^2 = 1 - \frac{RSS}{TSS}$ , we have

$$F = \frac{R^2}{1 - R^2} \frac{k - 1}{n - k}$$

## The meaning of $F$ test

- ▶ Intuitively,  $F$  test compares the explanatory power between the restricted and the unrestricted models.
- ▶ If the restriction was not true, then the explanatory power under the restricted model would have largely dropped, making  $F$  statistic big.
- ▶ If the restriction was true, then the explanatory power would not differ by big, so  $F$  statistic will be small.

# Standard Reporting Style

Linear regression

Number of obs	=	3,000
F(3, 2996)	=	59.48
Prob > F	=	0.0000
R-squared	=	0.0729
Root MSE	=	570.47

birthweight	Robust					
	Coefficient	std. err.	t	P> t	[95% conf. interval]	
smoker	-217.5801	26.10764	-8.33	0.000	-268.7708	-166.3894
alcohol	-30.49129	72.59671	-0.42	0.675	-172.8357	111.8531
nprevist	34.06991	3.608326	9.44	0.000	26.99487	41.14496
_cons	3051.249	43.71445	69.80	0.000	2965.535	3136.962

## Next Lecture

- ▶ We will skip Chapter 8 and 9 of Stock and Watson.
- ▶ In the next lecture, we will cover the Panel data regression.

## Exercise

- ▶ Using the same dataset `birthweight_smoking`, replicate the multiple linear regression table on p.38. In particular, report
  - (i) homoskedastic standard error
  - (ii) heteroskedasticity robust standard error (HC0)
  - (iii) F-statistic
- ▶ Do not use regression package. (`readxl` only)

# Exercise

## Stock and Watson, Ch5: 5.15

- 5.15** A researcher has two independent samples of observations on  $(Y_i, X_i)$ . To be specific, suppose  $Y_i$  denotes earnings,  $X_i$  denotes years of schooling, and the independent samples are for men and women. Write the regression for men as  $Y_{m,i} = \beta_{m,0} + \beta_{m,1}X_{m,i} + u_{m,i}$  and the regression for women as  $Y_{w,i} = \beta_{w,0} + \beta_{w,1}X_{w,i} + u_{w,i}$ . Let  $\hat{\beta}_{m,1}$  denote the OLS estimator constructed using the sample of men,  $\hat{\beta}_{w,1}$  denote the OLS estimator constructed from the sample of women, and  $SE(\hat{\beta}_{m,1})$  and  $SE(\hat{\beta}_{w,1})$  denote the corresponding standard errors. Show that the standard error of  $\hat{\beta}_{m,1} - \hat{\beta}_{w,1}$  is given by  $SE(\hat{\beta}_{m,1} - \hat{\beta}_{w,1}) = \sqrt{[SE(\hat{\beta}_{m,1})]^2 + [SE(\hat{\beta}_{w,1})]^2}$ .

## Exercise

Suppose that you estimated the following regression

$$\log(wage) = \beta_0 + \beta_1 \cdot education + \beta_2 \cdot experience + \beta_3 \cdot \frac{experience^2}{100} \\ + \beta_4 \cdot Northeast + \beta_5 \cdot South + \beta_6 \cdot West \\ + \beta_7 \cdot married + \beta_8 \cdot widowed or divorced \\ + \beta_{10} \cdot separated + u$$

and the point estimates and the robust standard errors (HC0) are

$\log(wage)$	$\hat{\beta}$	$s(\hat{\beta})$
Education	0.0883	0.0029
Experience	0.0279	0.0028
Experience <sup>2</sup> /100	-0.0363	0.0055
Northeast	0.0616	0.0360
South	-0.0679	0.0296
West	0.0195	0.0282
Married	0.1778	0.0250
Widowed or Divorced	0.0857	0.0443
Separated	0.0167	0.0525
Intercept	1.1929	0.0501

## Exercise

- (a) Calculate the expected log(wage) for a person with 12 years of education and 10 years work experience who lives in the West and has never married.
- (b) Find the 95% asymptotic confidence interval (CI) for  $\beta_1$ , the returns to education. Briefly interpret the CI.
- (c) Test the null hypothesis that marital status does not affect wage. Write down the null hypothesis in terms of the parameters in the above equation. What is the number of restrictions and what is the appropriate test statistic?

## Exercise

- ▶ You have two independent samples  $(Y_{1i}, X_{1i})$  and  $(Y_{2i}, X_{2i})$  both with sample sizes  $n$  which satisfy  $Y_1 = X'_1\beta_1 + e_1$  and  $Y_2 = X_2\beta_2 + e_2$ , where  $\mathbb{E}[e_1|X_1] = 0$  and  $\mathbb{E}[e_2|X_2] = 0$ . Let  $\hat{\beta}_1$  and  $\hat{\beta}_2$  be the OLS estimators of  $\beta_1 \in \mathbb{R}^1$  and  $\beta_2 \in \mathbb{R}^1$ .
- (a) Find the asymptotic distribution of

$$\sqrt{n} \left( (\hat{\beta}_2 - \hat{\beta}_1) - (\beta_2 - \beta_1) \right)$$

as  $n \rightarrow \infty$ .

- (b) Find an appropriate test statistic for  $H_0 : \beta_2 = \beta_1$ .
- (c) Find the asymptotic distribution of this statistic under the null hypothesis.

# SW10. Regression with Panel Data

Ryu Myeonggyu

June 18, 2025

## Panel Data

- ▶ Until now, the data we covered has only one time period.
- ▶ This kind of data is called the **cross-sectional data**.
- ▶ Now we will consider a data with multiple time periods.
- ▶ **Panel data** (also called **longitudinal data**) refers to data for  $N$  different entities observed at  $T$  different time periods.
- ▶ Typically, we denote it by

$$\{(Y_{it}, X_{it}) | i = 1, \dots, N, t = 1, \dots, T\}$$

- ▶ Assume  $X_{it}$  is  $k \times 1$  vector.

## Panel Data

- ▶ A panel that has some missing data for at least one time period for at least one entity is called an **unbalanced panel**.
- ▶ Otherwise, it is called a **balanced panel**.
- ▶ We assume our data is a balanced panel unless otherwise stated.
- ▶ Obviously Panel data has much richer structure than a cross-sectional data.
- ▶ You can always run an OLS regression with panel data.
- ▶ This is called the Pooled OLS (POLS) estimation.
- ▶ The question is, how can we exploit this panel structure?

## Panel Data

- ▶ A motivating example: return to education
- ▶  $Y = \log(\text{income})$ ,  $X = \text{years of education}$
- ▶ Does education increase income?
- ▶ We can answer this question by regressing  $Y$  on  $X$ .
- ▶ An OVB:  $Z = \text{intelligence}$
- ▶  $Z$  causes an OVB because  $Z$  is not only (positively) correlated with  $X$  but also (positively) correlated with  $Y$ .
- ▶ Without including  $Z$  into our regression, the regression coefficients will be biased.

## Panel Data

- ▶ However, in most cases, we do not know how intelligent the workers are.
- ▶ Thus the omitted variable bias will not be solved by including  $Z$ .
- ▶ Suppose that we have a panel data.
- ▶ We believe our true model will be

$$Y_{it} = X'_{it}\beta + Z'_{it}\gamma + u_{it}$$

- ▶ Notice that by definition, the intelligence of a worker is time-invariant. i.e.

$$Z_{it} = Z_i$$

## Panel Data

- ▶ So our regression model becomes

$$Y_{it} = X'_{it}\beta + Z'_i\gamma + u_{it}$$

or

$$Y_{it} = X'_{it}\beta + e_i + u_{it}$$

and this model is called the **one-way error component model.**

- ▶ At time period  $t - 1$ ,

$$Y_{i,t-1} = X'_{i,t-1}\beta + Z'_i\gamma + u_{i,t-1}$$

- ▶ Subtracting the two, we have

$$\begin{aligned} Y_{it} - Y_{i,t-1} &= (X_{it} - X_{i,t-1})'\beta + (Z_i - Z_i)'\gamma + u_{it} - u_{i,t-1} \\ &= (X_{it} - X_{i,t-1})'\beta + (u_{it} - u_{i,t-1}) \end{aligned}$$

## First-Difference Estimator

- ▶ Define the first-differencing operator  $\Delta$  by  
$$\Delta X_{it} = X_{it} - X_{i,t-1}$$
, we may write this as

$$\Delta Y_{it} = \Delta X'_{it} \beta + \Delta u_{it}$$

- ▶ The above approach is called the **first-difference** (FD) estimation.
- ▶ Notice that this model does not include the  $Z_{it}$  term.
- ▶ Another way to get rid of  $Z_{it}$  is by taking the time mean.
- ▶ Taking the time mean  $\frac{1}{T} \sum_{t=1}^T$ , we have

$$\begin{aligned}\frac{1}{T} \sum_{t=1}^T Y_{it} &= \frac{1}{T} \sum_{t=1}^T X'_{it} \beta + \frac{1}{T} \sum_{t=1}^T Z'_i \gamma + \frac{1}{T} \sum_{t=1}^T u_{it} \\ &= \frac{1}{T} \sum_{t=1}^T X'_{it} \beta + Z'_i \gamma + \frac{1}{T} \sum_{t=1}^T u_{it}\end{aligned}$$

## Fixed Effects

- ▶ Define the time-demeaning operator  $\ddot{X}_{it} := X_{it} - \frac{1}{T} \sum_{t=1}^T X_{it}$ .
- ▶ Subtracting the time mean from the original model, we have

$$\ddot{Y}_{it} = \ddot{X}'_{it}\beta + \ddot{Z}'_i\gamma + \ddot{u}_{it} = \ddot{X}'_{it}\beta + \ddot{u}_{it}$$

- ▶ This model also has no unobservable  $Z_i$  term.
- ▶ This model is called the **fixed effects** (FE) model.
- ▶ Either using FD or FE model, we can estimate  $\beta$  using the least squares estimator.
- ▶ This shows how powerful the panel data structure is.

## LSDV

- ▶ Another way to estimate the fixed effects model is to include individual dummy variables.

$$Y_{it} = X'_{it}\beta + \sum_{i'=1}^N \mathbb{1}_{\{i=i'\}} \delta_{i'} + u_{it}$$

where  $\mathbb{1}$  is the indicator function.

- ▶ This model is called the **least squares dummy variables (LSDV)** model.

**Remark** This model is ill-posed if  $X_{it}$  contains an intercept term.  
(why?)

- ▶ If your  $X$  contains an intercept, your model should change to

$$Y_{it} = X'_{it}\beta + \sum_{i=1}^{N-1} \mathbb{1}_{\{i=i'\}} \delta_{i'} + u_{it}$$

## FE and LSDV

- ▶ Using OLS, we can estimate the same  $\beta$  of FE model.

**proof** Define

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1T} \\ Y_{21} \\ \vdots \\ Y_{2T} \\ \vdots \\ Y_{NT} \end{pmatrix}, X = \begin{pmatrix} X_{11}^{(1)} & \cdots & X_{11}^{(k)} \\ \vdots & \ddots & \vdots \\ X_{1T}^{(1)} & \cdots & X_{1T}^{(k)} \\ X_{21}^{(1)} & \cdots & X_{21}^{(k)} \\ \vdots & \ddots & \vdots \\ X_{2T}^{(1)} & \cdots & X_{2T}^{(k)} \\ \vdots & \ddots & \vdots \\ X_{NT}^{(1)} & \cdots & X_{NT}^{(k)} \end{pmatrix}, u = \begin{pmatrix} u_{11} \\ \vdots \\ u_{1T} \\ u_{21} \\ \vdots \\ u_{2T} \\ \vdots \\ u_{NT} \end{pmatrix}$$

## FE and LSDV

► and

$$\mathbb{1} := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \delta = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_N \end{pmatrix}$$

► so that

$$Y = X\beta + \mathbb{1}\delta + u$$

## FE and LSDV

- ▶ Consider the annihilator matrix

$$M_{\mathbb{1}} = I - \mathbb{1}(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}'$$

which equals

$$M_{\mathbb{1}} = I_{NT} - \frac{1}{T} \mathbb{1} \mathbb{1}' = \begin{pmatrix} 1 - \frac{1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} \\ \vdots & \vdots & \ddots & \vdots \\ 1 - \frac{1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} \\ -\frac{1}{T} & 1 - \frac{1}{T} & \cdots & -\frac{1}{T} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{T} & 1 - \frac{1}{T} & \cdots & -\frac{1}{T} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{T} & -\frac{1}{T} & \cdots & 1 - \frac{1}{T} \end{pmatrix}$$

## FE and LSDV

- ▶ Multiplying  $M_1$  to both sides, we have

$$\ddot{Y} = \ddot{X}\beta + \ddot{u}$$

which is the fixed effects regression.

- ▶ This is a typical (and powerful) application of FWL theorem.
- ▶ This is the reason why FE model got its name: the dummy variables absorb the effect fixed to individuals.
- ▶ This also shows that FE model removes effects from all time-invariant sources, including the effect from intelligence, gender, intercept, region of birth, etc.

## Quasi-demeaning framework

- ▶ We present a unified framework for the estimators above: the ANOVA framework.
- ▶ Let us denote the time mean by  $\dot{Y}_i$  and the grand-mean

$$\frac{1}{N} \sum_{i=1}^N \frac{1}{T} \sum_{t=1}^T Y_{it} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T Y_{it}$$

by the bar notation  $\bar{Y}$ .

- ▶ We may subtract  $\theta \cdot \dot{Y}_i$  from  $Y_{it}$

$$Y_{it} - \theta \dot{Y}_i$$

for  $\theta \in [0, 1]$ . (this is called the **quasi-demeaning**)

- ▶ This has a total sample mean

$$\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \theta \dot{Y}_i) = (1 - \theta) \bar{Y}$$

## Quasi-demeaning framework

- ▶ Now taking the variance, we see that

$$\begin{aligned}& \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T [Y_{it} - \theta \dot{Y}_i - (1-\theta) \bar{Y}]^2 \\&= \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T [(Y_{it} - \dot{Y}_i) - (1-\theta)(\dot{Y}_i - \bar{Y})]^2 \\&= \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \dot{Y}_i)^2 + (1-\theta)^2 \frac{1}{N} \sum_{i=1}^N (\dot{Y}_i - \bar{Y})^2\end{aligned}$$

where the cross-term vanishes by taking the time-mean.

## Quasi-demeaning framework

- ▶ The first term is each observation's variation from its time-mean, which is called the **within variation**.
- ▶ The second term is the variation of each individual's time mean from the grand mean, which is called the **between variation**.
- ▶ On the other hand, the quasi-demeaning means

$$Y_{it} - \theta \dot{Y}_i = (X_{it} - \theta \dot{X}_i)' \beta + (Z_i - \theta \dot{Z}_i)' \gamma + (u_{it} - \theta \dot{u}_i)$$

- ▶ The above model is equivalent to the pooled OLS if we let  $\theta = 0$ , and if we let  $\theta = 1$ , the above model equals the FE model.

## Analysis of Variance

- ▶ But if we let  $\theta = 1$ , the total variation becomes

$$\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \bar{Y}_i)^2$$

which is the within-variation.

- ▶ Hence the estimation depends solely on the within variation of the data.
- ▶ That's why we call the time-demeaning the "within-transformation."
- ▶ On the other hand, the pooled OLS model depends on within variation and between variation with equal weights.

## Between Effects Estimator

- ▶ One may use the time-averaged data for estimation.
- ▶ The between effects (BE) estimator is estimated from

$$\dot{Y}_i = \dot{X}'_i \beta + e_i + \dot{u}_i$$

- ▶ Treats as a cross-sectional data and depends on between-variations only.
- ▶ Rarely used as major specification.
- ▶ Sometimes used to detect measurement errors.

## Random Effects

- ▶ In some situation we can try efficient estimation.
- ▶ Suppose the  $Z_i$  (or  $e_i$ ) term is a **random error**

$$Y_{it} = X'_{it}\beta + e_i + u_{it}$$

- ▶ In particular,  $e_i$  and the  $X_{it}$  are uncorrelated.

$$\text{Cov}(X_{it}, e_i) = 0$$

- ▶ Under this situation, we can do the random effects estimation, where we set

$$\theta_{RE} = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_e^2}}$$

where  $\sigma_u^2$  and  $\sigma_e^2$  are homoskedastic variances of  $u_{it}$  and  $e_i$ , respectively.

## Comments on Random Effects Model

- ▶ GLS is used to estimate the RE model.
- ▶ It is known that random effects estimator is the most efficient estimator under the assumptions above.
- ▶ However, the random effects estimator depends quite strong assumptions which are usually not plausible in social sciences.
- ▶ For instance, the assumption that  $\text{Cov}(e_i, X_{it}) = 0$  implies in our unobserved intelligence situation that the intelligence and workers' education is not correlated, which does not make sense.
- ▶ In practice, random effects model is rarely used.

## Hausman Test

- ▶ There is a statistical test called Hausman test that tests fixed effects vs. random effects model.
- ▶ The test statistic is

$$H := (\hat{\beta}_{RE} - \hat{\beta}_{FE})'(\hat{V}_{\hat{\beta}, FE} - \hat{V}_{\hat{\beta}, RE})^{-1}(\hat{\beta}_{RE} - \hat{\beta}_{FE})$$

- ▶  $H_0$ : RE is efficient and consistent. FE is inefficient but consistent.  
vs  $H_1$ : RE is inconsistent and FE is consistent.
- ▶ The test statistic measures how different RE and FE estimates are.
- ▶  $H$  is asymptotically  $\chi^2$ -distributed with degrees of freedom  $k - 1$ .

## Two-way Error Components Model

- ▶ Back to the original example, now you want to consider the time-fixed effects.
- ▶ e.g. COVID-19, global recession, etc.
- ▶ Now your model is

$$Y_{it} = X'_{it}\beta + e_i + v_t + u_{it}$$

- ▶ You can estimate this **two-way error components model** using the within-transformation

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

or dummy variables regression.

- ▶ Practically, we usually do time-demeaning and include time dummies.
- ▶ This estimator is called the **two-way fixed effects (TWFE)** estimator.

## Standard Errors for Fixed Effects Regression

- ▶ In general, for individual  $i$ , the errors  $u_{it}$  and  $u_{is}$  are not independent.
- ▶ We say the errors are **autocorrelated** or **serially correlated** if

$$\mathbb{E}[u_{it} u_{is}] \neq 0$$

- ▶ In this case, the usual iid standard error cannot be applied.
- ▶ Instead, we need to use heteroskedasticity-autocorrelation-robust (HAR) standard error.
- ▶ A common way is to use standard errors clustered at individual level.
- ▶ Clustered standard errors allow for heteroskedasticity and for arbitrary autocorrelation within an entity but treat the errors as uncorrelated across entities.

# A Standard Reporting Style

**TABLE 10.1** Regression Analysis of the Effect of Drunk Driving Laws on Traffic Deaths

Dependent variable: traffic fatality rate (deaths per 10,000).

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36 (0.05) [0.26, 0.46]	-0.66 (0.29) [-1.23, -0.09]	-0.64 (0.26) [-1.35, 0.07]	-0.45 (0.30) [-1.04, 0.14]	-0.69 (0.35) [-1.38, 0.00]	-0.46 (0.31) [-1.07, 0.15]	-0.93 (0.34) [-1.60, -0.26]
Drinking age 18		0.10		0.03 (0.07) [-0.11, 0.17]	-0.01 (0.08) [-0.17, 0.15]	0.04 (0.10) [-0.16, 0.24]	
Drinking age 19				-0.02 (0.05) [-0.12, 0.08]	-0.08 (0.07) [-0.21, 0.06]	-0.07 (0.10) [-0.26, 0.13]	
Drinking age 20				0.03 (0.05) [-0.07, 0.13]	-0.10 (0.06) [-0.21, 0.01]	-0.11 (0.13) [-0.36, 0.14]	
Drinking age					0.00 (0.02) [-0.05, 0.04]		
Mandatory jail or community service?				0.04 (0.10) [-0.17, 0.25]	0.09 (0.11) [-0.14, 0.31]	0.04 (0.10) [-0.17, 0.25]	0.09 (0.16) [-0.24, 0.42]
Average vehicle miles per driver				0.008 (0.007)	0.017 (0.011)	0.009 (0.007)	0.124 (0.049)
Unemployment rate				-0.063 (0.013)		-0.063 (0.013)	-0.091 (0.021)
Real income per capita (logarithm)				1.82 (0.64)		1.79 (0.64)	1.00 (0.68)
Years	1982–88	1982–88	1982–88	1982–88	1982–88	1982–88	1982 & 1988 only
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	yes	yes	yes	yes	yes	yes
<i>F</i> -Statistics and <i>p</i> -Values Testing Exclusion of Groups of Variables							
Time effects = 0		4.22 (0.002)	10.12 (<0.001)	3.48 (0.006)	10.28 (<0.001)	3749 (<0.001)	
Drinking age coefficients = 0			0.35 (0.786)	1.41 (0.253)		0.42 (0.738)	
Unemployment rate, income per capita = 0				29.62 (<0.001)	31.96 (<0.001)	25.20 (<0.001)	
<i>R</i> <sup>2</sup>	0.091	0.889	0.891	0.926	0.893	0.926	0.899

These regressions were estimated using panel data for 48 U.S. states. Regressions (1) through (6) use data for all years 1982 to 1988, and regression (7) uses data from 1982 and 1988 only. The data set is described in Appendix 10.1. Standard errors are given in parentheses under the coefficients, 95% confidence intervals are given in square brackets under the coefficients, and *p*-values are given in parentheses under the *F*-statistics.

## Exercise

- (a) Suppose you have a balanced panel of  $N$  individuals and  $T$  time periods. Your regression model contains an intercept.
- ▶ In the one-way error component model, what is the maximum number of dummy variables for individuals you can include? What about the two-way error components model? Explain why.
- (b) In the one-way error component model, prove that FE and FD estimates are equal if  $T = 2$ .

## Exercise

- ▶ Download the dataset named Income and Democracy from [https://www.princeton.edu/~mwatson/Stock-Watson\\_3u/Students/Stock-Watson-EmpiricalExercises-DataSets.htm](https://www.princeton.edu/~mwatson/Stock-Watson_3u/Students/Stock-Watson-EmpiricalExercises-DataSets.htm).
- ▶ Do Exercise E10.2 on p.386 of Stock and Watson textbook.  
(except (b))

# Exercise

- ▶ Stock and Watson, Ch 10.
- ▶ Exercise: 10.7

# Resampling methods

## & SW12. Instrumental Variables Regression: Part 1

Ryu Myeonggyu

June 21, 2025

## Resampling methods

- ▶ We learned exact inference and asymptotic inference.
- ▶ Resampling methods: simulation-based inference.
- ▶ In particular, we will study bootstrapping.

## Ideal Inference

- ▶ Suppose we have iid observations  $\{(Y_i, X_i)\}$  from joint cdf  $F$ .
- ▶ Sampling distribution of a statistic  $T_n$  is

$$G_n(u, F) = \mathbb{P}(T_n \leq u | F)$$

- ▶  $G_n$  depends on  $F$  and sample size  $n$ .
- ▶ If we know  $G_n$  directly, then we can calculate exact size.
- ▶ Since  $G_n$  is not known in general, use asymptotics

$$G(u, F) = \lim G_n(u, F)$$

## Alternative Approach

- ▶ An alternative approach is to approximate with

$$G_n^*(u, F) := G_n(u, F_n)$$

where  $F_n$  is the empirical distribution function (edf).

- ▶  $(Y_i^*, X_i^*)$ : a random sample from edf  $F_n$  is called resample or bootstrap sample.
- ▶ A statistic  $T_n^*$  based on bootstrap sample  $(Y_i^*, X_i^*)$  is called a bootstrap statistic.

# Empirical Distribution Function

- ▶ For a random sample  $(X_1, Y_1), \dots, (X_n, Y_n) \sim F$ , the empirical distribution function is

$$F_n(x, y) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{[X_i \leq x]} \mathbb{1}_{[Y_i \leq y]}$$

Thm (Glivenko-Cantelli)

$$\|F_n - F\|_\infty \xrightarrow{a.s.} 0$$

where  $\|\cdot\|_\infty$  is the (essential) supremum norm.

- ▶ The convergence is ‘uniform’, not ‘pointwise.’
- ▶ This theorem justifies the use of Bootstrap method.

# Nonparametric Bootstrap

- ▶ The key idea of bootstrap is “resampling.”
- ▶ You draw your sample from your sample data! (with replacement)
- ▶ There are  $\binom{2n - 1}{n}$  possible outcomes.
- ▶ For a practical  $n$ , having all feasible outcomes will not be feasible.
- ▶ We generate  $B$  bootstrap samples  $\{(Y_i, X_i)_{i=1}^n\}_{b=1}^B$  instead of all possible outcomes.

# Nonparametric Bootstrap

- ▶ Let  $\theta$  be the parameter of interest.
- ▶ The estimator for  $\theta$  is  $\hat{\theta}$ .
- ▶ The variance of  $\hat{\theta}$  is

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

- ▶ Its bootstrap estimate is

$$V^*(\hat{\theta}) = \mathbb{E}^*[(\hat{\theta}^* - \mathbb{E}^*[\hat{\theta}^*])^2] = \frac{1}{B} \sum_{i=1}^B \left( \hat{\theta}_b^* - \frac{1}{B} \left( \sum_{b=1}^B \hat{\theta}_b^* \right) \right)^2$$

# Nonparametric Bootstrap

- ▶ Ordered bootstrap distribution of  $\hat{\theta}^*$

$$\theta_{(1)}^*, \theta_{(2)}^*, \dots, \theta_{(B)}^*$$

- ▶ Efron's 95% percentile interval when  $B = 1,000$ :

$$[\theta_{(26)}^*, \theta_{(975)}^*]$$

- ▶ Commonly used when it is difficult to calculate standard error
- ▶ May perform poorly when the sampling distribution of  $\hat{\theta}$  is biased and asymmetric.

## Bootstrapping t-statistic

- ▶ We may bootstrap t-statistic directly, rather than  $\hat{\theta}$ .
- ▶ For each bootstrap sample, calculate the bootstrap t-statistic

$$T_b^* = \frac{\hat{\beta}_b^* - \hat{\beta}_b}{s(\hat{\beta}_b)}$$

- ▶ Get the bootstrap distribution of  $T_b^*$ .
- ▶ Do not use bootstrap standard error when constructing t-statistic.
- ▶  $T^*$  is centered around  $\hat{\theta}$ , not  $\theta$ .

## Bootstrap Percentile-t Intervals

- ▶ Equal-tailed vs. symmetric percentile
- ▶ Equal-tailed: Same probability on both tails

$$1 - \alpha = \mathbb{P}(q_n(\alpha/2) \leq t_n \leq q_n(1 - \alpha/2))$$

where  $q_n(\alpha)$  is the  $1 - \alpha$ -th quantile of the sampling distribution of  $t_n$ .

- ▶ Since  $q_n(\alpha/2), q_n(1 - \alpha/2)$  are unknown, replace them with bootstrap estimates.

e.g. If  $\alpha = 0.05$  and  $B = 1,000$ , then

$$q_n^*(0.025) = \left( \frac{\hat{\beta}_b^* - \hat{\beta}_b}{s(\hat{\beta}_b)} \right)_{(26)}$$

and

$$q_n^*(0.975) = \left( \frac{\hat{\beta}_b^* - \hat{\beta}_b}{s(\hat{\beta}_b)} \right)_{(975)}$$

## Symmetric percentile t-interval

- ▶ Symmetric percentile: Same distance around the center

$$1 - \alpha = \mathbb{P} \left( \left| \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \right| \leq q_n(1 - \alpha) \right)$$

where  $q_n(\alpha)$  is the  $1 - \alpha$ -th quantile of the sampling distribution of  $|t_n|$ .

- ▶ Replace the unknown  $q_n(1 - \alpha)$  with bootstrap estimate.  
e.g. If  $\alpha = 0.05$  and  $B = 1,000$ , then

$$q_n^*(0.95) = \left| \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \right|_{(950)}$$

## Summary

- ▶ Neither large-sample asymptotics (CLT) nor bootstrap provides the exact sampling distribution of a test statistic.
- ▶ Nevertheless, the bootstrap is useful when standard errors are infeasible or hard to calculate.
- ▶ Furthermore, by bootstrapping asymptotically pivotal statistics, e.g.,  $t_n$ , we can get more accurate approximation to the sampling distribution of the test statistic. This is called asymptotic refinements or higher-order improvements of the bootstrap.

# Endogeneity

- ▶ Consider a linear model

$$Y_i = X'_i \beta + u_i$$

- ▶ This model is called a structural equation.
- ▶ Endogeneity refers to a situation where

$$\mathbb{E}[Xu] \neq 0$$

- ▶ In this case we say  $X$  is an endogenous regressor.
- ▶ If  $X$  is not endogenous, we say  $X$  is exogenous.

## Endogeneity

- ▶ There can be various sources of endogeneity.

e.g. (Measurement error)

Suppose the true relationship is

$$Y = Z'\beta + e, \quad \mathbb{E}[e|Z] = 0$$

- ▶ But  $Z$  is unobserved but  $X = Z + u$  is observed where  $u$  is a measurement error independent of  $e$  and  $Z$ .
- ▶ Then

$$Y = Z'\beta + e = (X - u)'\beta + e = X'\beta + v$$

where  $v = -u\beta + e$ .

- ▶ In this case we have

$$\mathbb{E}[Xv] = \mathbb{E}[X'(-u\beta + e)] = -\beta\mathbb{E}[u^2] \neq 0$$

# Endogeneity

e.g. Choice variables

Suppose

$$\log(wage) = \beta_0 + \beta_1 \cdot edu + X'\gamma + u$$

where  $X$  is a vector of observable characteristics.

- ▶ Ability or motivation is unobserved but affects wage.
- ▶ Induces an omitted variable bias.

# Instrumental Variables

- ▶ Consider the structural equation

$$Y_i = X'_i \beta + u_i$$

- ▶ Suppose we have  $k_1$  exogenous variables and  $k_2$  endogenous variables with  $k = k_1 + k_2$ .
- ▶  $X = (X_1, X_2)$  with  $X_1$  being exogenous and  $X_2$  being endogenous. i.e.  $\mathbb{E}[X'_1 u] = 0, \mathbb{E}[X'_2 u] \neq 0$ .

**def** An  $\ell \times 1$  random variable  $Z$  is called an instrumental variable if it satisfies

- (1) (exogeneity)  $\mathbb{E}[Z' u] = 0$
- (2) (no redundancy)  $\mathbb{E}[ZZ'] > 0$
- (3) (relevance)  $\text{rank}(\mathbb{E}[ZX']) = k$

## Instruments

- ▶ Partition  $Z = (Z_1, Z_2) = (X_1, Z_2)$  where  $Z_2 \in \mathbb{R}^{\ell_2}$ ,  
 $\ell = \ell_1 + \ell_2$ .
- ▶  $Z_1 = X_1$  is called the included exogenous variable.
- ▶  $Z_2$  is called the excluded exogenous variable.
- ▶ If  $\ell_2 = k_2$  then we say the model is just-identified. (or exactly identified)
- ▶ If  $\ell_2 > k_2$  then we say the model is over-identified.
- ▶ If  $\ell_2 < k_2$  then we say the model is under-identified.

## IV estimator

- ▶ If the instrument  $Z$  satisfies the conditions of instrument relevance and exogeneity, the coefficient can be estimated using an IV estimator called two stage least squares (2SLS).
- ▶ The first-stage relationship between  $X$  and  $Z$  is

$$X_2 = \Gamma' Z + u = \Gamma'_1 Z_1 + \Gamma'_2 Z_2 + e$$

with  $\mathbb{E}[Zu'_2] = 0$ .

- ▶ Often called the reduced-form equation for  $X$ .

## IV estimator

- ▶ Substituting for  $X_2$  in the structural equation,

$$Y = X'_1\beta_1 + X'_2\beta_2 + u = Z'_1\beta_1 + (\Gamma'_{12}Z_1 + \Gamma'_{22}Z_2 + e')\beta_2 + u$$

so that

- ▶

$$Y = Z'_1(\underbrace{\beta_1 + \Gamma_1\beta_2}_{=: \lambda_1}) + Z'_2(\underbrace{\Gamma_2\beta_2}_{=: \lambda_2}) + \underbrace{e'\beta_2 + u}_{=: \nu}$$

- ▶ Hence we write

$$Y = Z'\lambda + \nu$$

## IV estimator

- ▶ Since

$$\begin{cases} \lambda_1 = \beta_1 + \Gamma_1 \beta_2 \\ \lambda_2 = \Gamma_2 \beta_2 \end{cases}$$

we can write this as

$$\lambda = \underbrace{\begin{pmatrix} I_{k_1} & \Gamma_1 \\ O_{\ell_2 \times k_1} & \Gamma_2 \end{pmatrix}}_{= \bar{\Gamma}} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

- ▶ Once we identify  $\lambda$ , then we can recover the information about  $\beta$  from the above relation.
- ▶ Here we need some “invertibility conditions.”

## IV estimator

- ▶ In the just-identified case,  $\bar{\Gamma}$  is a square matrix.
- ▶ Hence in this case if  $\bar{\Gamma}$  is invertible, then we have

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

- ▶ In the over-identified case, we need  $\bar{\Gamma}$  to have full rank.
- ▶ Once this is satisfied, we have

$$\beta = (\bar{\Gamma}'\bar{\Gamma})^{-1}(\bar{\Gamma}'\lambda)$$

as our least squares solution. (Recall the normal equation)

- ▶ Once we do this, we estimate by analogy principle.

## Exercise

- Before conducting the simulation, you are required to set the seed number to 1234. Consider the linear regression model below:

$$Y_i = X_i\beta + e_i$$

Suppose the true  $\beta$  is 1.

- (a) Generate a random sample of size 500 by
- $X_i \stackrel{iid}{\sim} N(1, 1)$
  - $e_i \stackrel{iid}{\sim} N(0, 1)$
  - $Y_i = X_i\beta + e_i.$

## Exercise

- (b) Using samples in (a), conduct the nonparametric bootstrap with  $B = 1,000$ .
- i. Calculate bootstrap OLS estimators  $\{\hat{\beta}_b^*\}_{b=1}^B$  and the bootstrap standard error  $\sqrt{V^*}$ .
  - ii. Calculate the bootstrap t-statistics  $\{T_b^*\}$  and the equal-tailed percentile-t interval with  $\alpha=0.05$  for  $\beta$ . Any standard error must be calculated as a form of HC0. You must derive the t-statistics at the same step at (b) i.

## Exercise

- ▶ Consider the measurement error example. True relationship is

$$Y = Z\beta + e, \quad \mathbb{E}[e|Z] = 0$$

but you observe  $X = Z + u$  instead of  $Z$  where  $u$  is a mean-zero random error independent of  $Z$  and  $e$ .

- ▶ This measurement error is called the classical measurement error.
- ▶ Now you consider  $Y = X\beta^* + v$  where  $\mathbb{E}[Xv] \neq 0$ .
- ▶ Derive the OLS estimator  $\hat{\beta}^*$  and its probability limit.
- ▶ Compare this with  $p \lim \hat{\beta}$ . Which one is closer to zero?

## Exercise

- ▶ On page 25, what condition of the IV was used?
- ▶ Explain why we cannot identify  $\beta$  in the under-identified case.

## SW12. Instrumental Variables Regression: Part 2

Ryu Myeonggyu

June 24, 2025

## IV estimator

- The reduced form coefficients can be estimated by OLS:

$$\hat{\Gamma} = \left( \frac{1}{n} \sum_{i=1}^n Z_i Z_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n Z_i X_{2i}' \right)$$

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

## IV estimator

- ▶ In the just-identified case ( $\ell = k$ ), the IV estimator is

$$\hat{\beta}_{IV} = \left( \frac{1}{n} \sum_{i=1}^n Z_i X'_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n Z_i Y'_i \right)$$

- ▶ Equivalently, the indirect least square (ILS) estimator is

$$\hat{\beta}_{ILS} = \hat{\Gamma}^{-1} \hat{\lambda} = ((Z'Z)^{-1}(Z'X))^{-1}((Z'Z)^{-1}Z'Y)$$

## IV estimator

- ▶ In the over-identified case ( $\ell > k$ ), the 2SLS estimator is

$$\hat{\beta}_{\text{2SLS}} = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'Y)$$

- ▶ This formula seems quite lengthy, but

$$\hat{\beta}_{\text{2SLS}} = (X'P_Z X)^{-1}(X'P_Z Y)$$

- ▶ Since  $P_Z$  is symmetric and idempotent, we see that

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

## IV estimator

- ▶ The 2SLS estimation is done in two stages:

1st stage Regress  $X$  on  $Z$  and get fitted values

$$\hat{X} = P_Z X$$

2nd stage Regress  $Y$  on  $\hat{X}$

$$\hat{\beta} = (\hat{X}' \hat{X})^{-1} (\hat{X}' Y) = \hat{\beta}_{2SLS}$$

- ▶ The key idea is to **kill** the endogenous variation and use the exogenous variation of  $X$  by the instrument.

## Asymptotics of 2SLS estimator

- ▶ Under some regularity conditions,  $\hat{\beta}_{2SLS}$  is consistent and asymptotically normal.
- ▶ (consistency)

$$\hat{\beta}_{2SLS} \xrightarrow{P} \beta$$

- ▶ (asymptotic normality)

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

where

$$V_\beta = (Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1} (Q'_{ZX} Q_{ZZ}^{-1} \Omega Q_{ZZ}^{-1} Q_{ZX}) (Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1},$$

$Q_{ZX} = \mathbb{E}[Z_i X'_i]$ ,  $Q_{ZZ} = \mathbb{E}[Z_i Z'_i]$  and  $\Omega = \mathbb{E}[Z_i Z'_i e_i^2]$

## Standard Errors

- ▶ Under homoskedasticity,

$$V_\beta^0 = (Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1} \mathbb{E}[e_i^2]$$

- ▶ For the robust estimator, using the analogy principle,

$$\hat{V}_\beta = (\hat{Q}'_{ZX} \hat{Q}_{ZZ}^{-1} \hat{Q}_{ZX})^{-1} (\hat{Q}'_{ZX} \hat{Q}_{ZZ}^{-1} \hat{\Omega} \hat{Q}_{ZZ}^{-1} \hat{Q}_{ZX}) (\hat{Q}'_{ZX} \hat{Q}_{ZZ}^{-1} \hat{Q}_{ZX})^{-1}$$

where  $\hat{Q}_{ZX} = \frac{1}{n} \sum_{i=1}^n Z_i X'_i$  and *mutatis mutandis*.

- ▶ Do not use second stage standard error.

## Threats to IV

- ▶ Endogeneity test
- ▶ Consider the structural equation and reduced form

$$Y = X_1' \beta_1 + X_2' \beta_2 + u$$

$$X_2 = \Gamma_1' Z_1 + \Gamma_2' Z_2 + e$$

- ▶ IV assumption:  $\mathbb{E}[Zu] = 0$
- ▶  $X_2$  is endogenous  $\iff e$  and  $u$  are correlated.
- ▶ Consider the linear projection of  $u$  on  $e$

$$u = e'\alpha + \nu, \quad \mathbb{E}[e\nu] = 0$$

## Endogeneity Test

- ▶ Substituting into the structural equation<sup>1</sup>

$$Y = X_1' \beta_1 + X_2' \beta_2 + e' \alpha + \nu$$

with  $\mathbb{E}[X_1 \nu] = \mathbb{E}[X_2 \nu] = \mathbb{E}[e \nu] = 0$ . (exercise)

- ▶ Since  $e$  is unobserved, replace  $e$  by  $\hat{e} = X_2 - \hat{\Gamma}'_1 Z_1 - \hat{\Gamma}'_2 Z_2$
- ▶ Endogeneity test
- ▶  $H_0 : \mathbb{E}[X_2 u] = 0$  vs.  $H_1 : \mathbb{E}[X_2 u] \neq 0$ .
- ▶ Test  $\alpha = 0$  vs.  $\alpha \neq 0$ .

---

<sup>1</sup>This method is called the control function regression.

## Threats to IV

- ▶ Over-identification: number of IV > number of endogenous regressors.
- ▶ Are the IV's (jointly) valid?
- ▶ We use the information of over-identification to test the joint validity of instruments. (more moment conditions than free parameters)
- ▶ Recall the IV exogeneity condition

$$\mathbb{E}[Zu] = 0 \iff \mathbb{E}[Z(Y - X'\beta)] = 0$$

- ▶ Assumption:  $\beta$  is uniquely determined.

## Threats to IV

- ▶ Sargan test (1958)
- ▶  $H_0 : \mathbb{E}[Zu] = 0$  vs  $\mathbb{E}[Zu] \neq 0$ .
- ▶ Assume conditional homoskedasticity

$$\mathbb{E}[u^2|Z] = \sigma^2$$

- ▶ Consider a linear regression of  $u$  on  $Z$

$$Z = u'\alpha + \nu$$

- ▶ Then  $H_0 \iff \alpha = 0$ .
- ▶ Sargan's test statistic is based on score

$$S := \hat{\alpha}'(\hat{V}ar(\hat{\alpha}))^\dagger \hat{\alpha} = \frac{\hat{e}'Z(Z'Z)^{-1}Z'\hat{e}}{\hat{\sigma}^2} \stackrel{A}{\sim} \chi_{\ell-k}^2$$

where  $\hat{\sigma}^2 = \frac{1}{n}\hat{e}'\hat{e}$ .

## Weak Instrument

- ▶ Instruments are called weak if the reduced form coefficients are of small magnitude.
- ▶ That is,  $Z$  has small explanatory power of  $X$  so that fairly big part of  $X$  is killed.
- ▶ For simplification, assume there is no included exogenous variables so that our model can be written as

$$Y = X'\beta + u$$

$$X = \Gamma'Z + e$$

- ▶ Staiger, Stock (1997)'s local-to-zero model

$$\Gamma = \frac{C}{\sqrt{n}}$$

where  $C$  is a free matrix.

## Weak Instrument

- ▶ One can show that

$$\hat{\beta}_{2SLS} - \beta = (X' P_Z X)^{-1} (X' P_Z u)$$

$$\xrightarrow{d} ((Q_Z C + \xi_2)' Q_Z^{-1} (Q_Z C + \xi_2))^{-1} (Q_Z C + \xi_2)' Q_Z^{-1} \xi_u$$

where

- ▶  $Q_Z = \mathbb{E}[Z_i Z_i']$ ,  $\xi_e = d \lim \frac{1}{\sqrt{n}} Z' u$  and  
 $\xi_2 = d \lim \frac{1}{\sqrt{n}} Z_i' e_i \sim N(0, Z_i Z_i' e_i^2)$ .
- ▶ This shows  $\hat{\beta}_{2SLS}$  is inconsistent and CLT does not hold in its standard form

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta)$$

- ▶ Invalidates the standard asymptotic inference

## Weak Instrument

- ▶ Stock, Wright, Yogo (2002), Stock, Yogo (2005)
- ▶ They propose F tests for the excluded instruments in the reduced form regressions with non-standard critical values.
- ▶ When there is one endogenous regressor and a single instrument the Stock-Yogo test rejects the null of weak instruments when this F statistic exceeds 10.<sup>2</sup>
- ▶ Other tests are available when there are more than one endogenous variable, such as Cragg and Donald (1996) or Kleibergen and Paap (2006).

---

<sup>2</sup>This is the famous rule-of-thumb!

## Weak Instrument

Table 1. Selected Critical Values for Weak Instrument Tests for TSLS  
Based on the First-stage F statistic

Number of instruments (K)	Relative bias > 10%		Actual size of 5% test > 15%	
	Threshold $\mu^2/K$	F statistic 5% critical value	Threshold $\mu^2/K$	F statistic 5% critical value
1			1.82	8.96
2			4.62	11.59
3	3.71	9.08	6.36	12.83
5	5.82	10.83	9.20	15.09
10	7.41	11.49	15.55	20.88
15	7.94	11.51	21.69	26.80

NOTE: The second column contains the smallest values of  $\mu^2/K$  that ensure that the bias of TSLS is no more than 10% of the inconsistency of OLS. The third column contains the 5% critical values applicable when the first-stage F statistic is used to test the null that  $\mu^2/K$  is less than or equal to the value in the second column against the alternative that  $\mu^2/K$  exceeds that value. The final two columns present the analogous weak-instrument thresholds and critical values when weak instruments are defined so that the usual nominal 5% TSLS t test of the hypothesis  $\beta = \beta_0$  has size potentially exceeding 15%. (Source: Stock and Yogo 2001.)

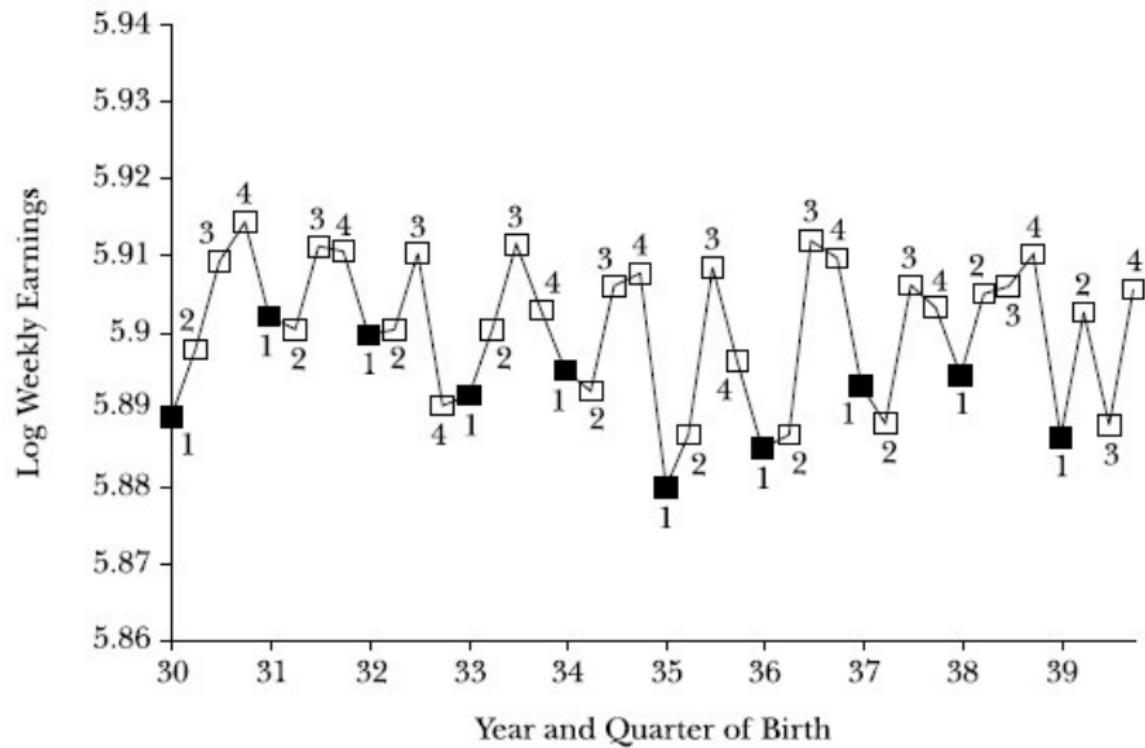
## Examples of IV

- ▶ 1. Angrist, Krueger (1991). Quarter of Birth
- ▶ 2. Card (1995). College Proximity
- ▶ 3. Angrist, Evans (1998). Twin IV
- ▶ 4. Bartik (1991). Shift-Share IV

## Quarter of Birth Instrument

- ▶ Angrist, Krueger (1994). Does Compulsory School Attendance Affect Schooling and Earnings?, *Quarterly Journal of Economics*.
- ▶  $X = \text{schooling}$
- ▶  $Y = \text{student outcomes (earnings, test score, ...)}$
- ▶ School dropout is NOT random. ( $\rightarrow$ selection bias)
- ▶ Compulsory schooling law-youth remain in school until their sixteenth or seventeenth birthday

# Quarter of Birth



# College Proximity

- ▶ Card (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling.
- ▶  $X = \text{education}$
- ▶  $Y = \log(\text{earning})$
- ▶ Education is NOT random!
- ▶  $Z = \text{indicator of college in the region}$
- ▶ Relevance? Exogeneity?

## Twin IV

- ▶ Angrist, Evans (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size, *American Economic Review*.
- ▶  $Y$  =labor supply
- ▶  $X$  =childbearing (more than two children)
- ▶ Simultaneous equation bias, selection bias
- ▶  $Z$  =twins at second birth

**Fact** Parents' prefer mixed sibling-sex composition

- ▶  $\tilde{Z}$  =first two children are of the same sex

## Bartik IV

- ▶ Elasticity in local labor markets

$$Y_i = \beta X_i + W'_i \gamma + \varepsilon_i$$

- ▶  $Y_i$  = outcome variable (e.g. wage growth)
- ▶  $X_i$  = explanatory variable (e.g. employment growth)
- ▶  $W_i$  = vector of observed controls
- ▶  $\varepsilon_i$  = idiosyncratic error (possibly correlated with  $X_i$ )
- ▶ There are  $K$  industries indexed by  $k$ .
- ▶ Simultaneous equation bias, omitted variable bias (e.g. changing local amenities) ...

## Bartik IV

- ▶ Shift-Share IV

$$Z_i = \sum_{k=1}^K s_{ik} g_k$$

where

- ▶  $s_{ik}$  =exposure shares varying across units (e.g. employment shares)
- ▶  $g_k$  =common shifts (industry-level changes)
- ▶ Idea

$$\underbrace{X_i}_{\text{employment growth}} = \frac{X_{i1} - X_{i0}}{X_{i0}} = \sum_{k=1}^K \underbrace{\frac{X_{ik0}}{X_{i0}}}_{s_{ik}} \cdot \underbrace{\frac{X_{ik1} - X_{ik0}}{X_{ik0}}}_{g_{ik}}$$

- ▶  $Z$  replaces  $g_{ik}$  by  $g_k$  (industry-wide shock)

- ▶ By construction,  $Z$  satisfies relevance condition.
- ▶ Assumption:

$$\mathbb{E}[g_k \varepsilon_i] = 0$$

- ▶ That is, shocks in local labor market and industry-level labor market are not correlated.
- ▶ Or we assume that the initial share  $s_{ik}$  is not correlated with  $\varepsilon_i$ .
- ▶ Either one is sufficient for exogeneity (Goldsmith-Pinkham et al. (2020), Borusyak and Jaravel (2022))

## Example of Bartik IV

- ▶ Autor, Dorn, Hanson (2013). The China Syndrome: Local Labor Market Effects of Import Competition in the United States, *American Economic Review*.
- ▶  $Y_i = \Delta$ local manufacturing employment
- ▶  $X_i = \Delta$ local exposure to Chinese imports
- ▶ Reverse causality:
  1. more exposure to Chinese imports → less production in the US → unemployment ↑
  2. local recession → unemployment ↑ → relies on cheaper products (=Chinese product) → exposure to Chinese imports ↑
- ▶  $S_{ik} = \frac{\text{Employment}_{ik}}{\text{Employment}_i}$
- ▶  $g_k = \Delta$ imports from China in countries **other than the US**.  
(proxy of Chinese productivity shift)

# Exercise

- ▶ Replicate some columns of the Table below.

Table 12.1: Instrumental Variable Wage Regressions

	OLS	IV(a)	IV(b)	2SLS(a)	2SLS(b)	LIML
education	0.074 (0.004)	0.132 (0.049)	0.133 (0.051)	0.161 (0.040)	0.160 (0.041)	0.164 (0.042)
experience	0.084 (0.007)	0.107 (0.021)	0.056 (0.026)	0.119 (0.018)	0.047 (0.025)	0.120 (0.019)
experience <sup>2</sup> /100	-0.224 (0.032)	-0.228 (0.035)	-0.080 (0.133)	-0.231 (0.037)	-0.032 (0.127)	-0.231 (0.037)
Black	-0.190 (0.017)	-0.131 (0.051)	-0.103 (0.075)	-0.102 (0.044)	-0.064 (0.061)	-0.099 (0.045)
south	-0.125 (0.015)	-0.105 (0.023)	-0.098 (0.0284)	-0.095 (0.022)	-0.086 (0.026)	-0.094 (0.022)
urban	0.161 (0.015)	0.131 (0.030)	0.108 (0.049)	0.116 (0.026)	0.083 (0.041)	0.115 (0.027)
Sargan				0.82	0.52	0.82
p-value				0.37	0.47	0.37

Figure: Card (1995)

## Exercise

- ▶ The regressors are education, experience(=age-education-6), experience<sup>2</sup>/100, black, south (indicator for residence in southern part of US) and urban (residence in a standard metropolitan statistical area). Drop observations for which wage is missing. The variables you will need for this exercise include lwage76, ed76 , age76, smsa76r, reg76r, black, nearc2, nearc4, nearc4a, nearc4b. See the description file for definitions.
- ▶ IV(b) uses 4-yr college, age, and age<sup>2</sup>/100 as instruments for education, experience, and experience<sup>2</sup>/100.
- ▶ 2SLS(a) uses public and private as instruments for education.

## Exercise

- (a) Replicate columns named OLS, IV(b) and 2SLS(a). Use HC1 robust standard errors. (do not report Sargan statistic and p-value)
- (b) For IV(b) and 2SLS(a), report standard errors calculated at the second stage. Are they equal to the reported standard errors?
  - ▶ Do (c) and (d) for 2SLS(a) only.
- (c) Using the Stock-Yogo's rule of thumb, are the instruments strong or weak?
- (d) Test the hypothesis that edu is exogenous for the structural return to schooling.

# Exercise

- ▶ Stock and Watson Ch.12
- ▶ Exercise: 3, 8, 9

# SW13. Experiments and Quasi-Experiments: Part 1

Ryu Myeonggyu

July 2, 2025

# Contents

Potential Outcomes Framework

LATE theorem

Quasi-Experiments

Exercise

# Potential Outcomes Framework

- ▶ Neyman-Rubin Causal Framework
- ▶ Also called ‘treatment effect’ framework
- ▶  $i$  indexes for individuals.
- ▶  $Y_i$  is the observed outcome variable of interest.
- ▶ There is a treatment  $D_i$  which is binary.
- ▶  $D_i = 1$  if treated,  $D_i = 0$  if untreated.
- ▶  $Y_i(1)$  and  $Y_i(0)$  denote the **potential** outcomes of individual  $i$ .
- ▶  $Y_i(1)$  =potential outcome if treated,  $Y_i(0)$  =potential outcome if untreated

# Potential Outcomes Framework

e.g.  $Y$  =income,  $D$  =marriage

- ▶  $Y_i(1)$  =potential income if married,  $Y_i(0)$  =potential income if not married.
- ▶ Does marriage has positive or negative effect on income?
- ▶ What is the causal effect of marriage on income?
- ▶ We need to compare the two:  $Y_i(1)$  vs.  $Y_i(0)$ .
- ▶ However, it is impossible to observe the both since an individual is either married or not married.

# Potential Outcomes Framework

- ▶ Hence we consider average causal parameters instead of individual causal parameters.

**ATE** The Average Treatment effect (ATE) is

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

**ATT** The Average Treatment effect on the treated (ATT) is

$$ATT := \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1]$$

## Difference-in-means estimator

- ▶ How can we estimate those causal parameters?
- ▶ One choice would be the difference-in-means estimator

$$\begin{aligned}\mathbb{E}[Y_i|D=1] - \mathbb{E}[Y_i|D=0] &= \mathbb{E}[Y_i(1)|D=1] - \mathbb{E}[Y_i(0)|D=0] \\ &= \mathbb{E}[Y_i(1) - Y_i(0)|D=1] \\ &\quad + \underbrace{\mathbb{E}[Y_i(0)|D=1] - \mathbb{E}[Y_i(0)|D=0]}_{\text{counterfactual}}\end{aligned}$$

i.e. compare the means between those who are treated and untreated.

- ▶ The last term is called the selection bias.

## Ideal of RCT

- ▶ Suppose the treatment  $D$  is randomly assigned. That is,

$$\{Y_i(1), Y_i(0)\} \perp D$$

- ▶ Also, observe that

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

- ▶ Then,

$$\begin{aligned}\mathbb{E}[Y_i|D=1] - \mathbb{E}[Y_i|D=0] &= \mathbb{E}[Y_i(1)|D=1] - \mathbb{E}[Y_i(0)|D=0] \\ &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[Y_i(1) - Y_i(0)] = ATE\end{aligned}$$

## Ideal of RCT

- ▶ In the experimental setting, also called **randomized controlled trial (RCT)**, the independence condition is satisfied.
- ▶ In social science research this is rarely true, since people are not randomly assigned to treatment.
  - e.g. Those who can afford a marriage life (i.e. richer people) are more likely to get married.
  - ▶ In this case, the condition would hold if people gets married by flipping a coin.

# Conditional Independence Assumption

- ▶ One solution is to use conditioning on observables.
- ▶ Suppose  $X$  is observed.

**CIA** We say the conditional independence assumption (CIA) holds if

$$\{Y_i(1), Y_i(0)\} \perp D|X$$

- ▶ Also called “unconfoundedness.”
- ▶ By similar argument, we can recover (exercise)

$$\mathbb{E}[Y_i(1) - Y_i(0)|X]$$

- ▶ This is called the conditional average treatment effect (CATE)
- ▶ ATE can be calculated via LIE.

## The role of CIA

- ▶ Suppose CIA is violated.
- ▶ Rewrite the potential outcome as a regression form.

$$\begin{aligned}Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\&= (Y_i(1) - Y_i(0))D_i + Y_i(0) \\&= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)]}_{=\beta_1} D_i + \underbrace{\mathbb{E}[Y_i(0)]}_{=\beta_0} \\&\quad + (Y_i(1) - Y_i(0) - \mathbb{E}[Y_i(1) - Y_i(0)])D_i + Y_i(0) - \mathbb{E}[Y_i(0)] \\&= \beta_0 + \beta_1 D_i + u_i;\end{aligned}$$

- ▶ Then

$$\mathbb{E}[D_i u_i] = \text{Cov}(D, Y(1)) \neq 0$$

- ▶  $D$  is endogenous!

# The meaning of CMI

- ▶ Recall that  $\hat{\beta}_1$  has causal interpretation if  $\mathbb{E}[u|D] = 0$ .
- ▶  $0 = \mathbb{E}[u_i|D_i = 0] = \mathbb{E}[Y_i(0)] - \mathbb{E}[Y_i(0)|D_i = 0]$
- ▶  $0 = \mathbb{E}[u_i|D_i = 1] = \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(1)]$
- ▶ Meaning?

**Note** The above holds if

$$\{Y_i(1), Y_i(0)\} \perp D$$

# Threats to Validity in RCT

- ▶ Even if you're conducting an RCT, the analysis can be invalid.
- ▶ Threats to internal validity:
  - ▶ Failure to randomize (imperfect randomization)
  - ▶ Failure to follow treatment protocol (or partial compliance)
  - ▶ Attrition (some subjects drop out)
  - ▶ Experimental effects (e.g. experimenter bias)
- ▶ Threats to external validity: cannot extend the result to population
  - ▶ Nonrepresentative sample
  - ▶ Nonrepresentative “treatment”
  - ▶ General equilibrium effects

## Wald IV

- ▶ Consider a binary instrument  $Z$
- ▶ The (binary) treatment  $D$  is potential

$$D = ZD(1) + (1 - Z)D(0)$$

i.e.  $D$  is a function of  $Z$ .

- ▶ The Wald estimand is

$$\frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]}$$

- ▶ This is an IV estimator for binary treatment and instrument.

## LATE theorem

- ▶ The question is: what causal parameter does Wald estimand identify?
- ▶ Imbens, Angrist (1994). Identification and Estimation of Local Average Treatment Effects, *Econometrica*.
- ▶ Under some assumptions, the IV estimator identifies the local average treatment effect. (LATE)

## LATE assumptions

- We can divide population into four groups:

$$\begin{cases} \text{Always-taker if } D(1) = D(0) = 1 \\ \text{Never-taker if } D(1) = D(0) = 0 \\ \text{Complier if } D(1) = 1, D(0) = 0 \\ \text{Defier if } D(1) = 0, D(0) = 1 \end{cases}$$

- The average treatment effect for the compliers is called the LATE

$$LATE = \mathbb{E}[Y_i(1) - Y_i(0)|\text{Compliers}]$$

## LATE assumptions

- ▶ Suppose that the assumptions A1-A4 hold:

- A1 (Exclusion)  $Y$  is constant with respect to  $Z$  controlling for  $D$ .
- A2 (Independence) Potential outcomes and potential treatments are independent of  $Z$ .

$$\{Y_i(1), Y_i(0), D_i(1), D_i(0)\} \perp Z$$

- A3 (Relevance)  $Z$  is correlated with  $D$ .

$$\mathbb{E}[D_i(1) - D_i(0)] \neq 0$$

- A4 (Monotonicity) There is no defier.

$$\mathbb{P}(D_i(1) - D_i(0) \geq 0) = 1$$

## Proof of LATE theorem

- ▶ First observe that

$$\begin{aligned}\mathbb{E}[D|Z=1] &= \mathbb{E}[D(1)Z + D(0)(1-Z)|Z=1] \\ &= \mathbb{E}[D(1)|Z=1] \\ &= \mathbb{E}[D(1)] \quad (\because \text{independence})\end{aligned}$$

- ▶ Similarly  $\mathbb{E}[D|Z=0] = \mathbb{E}[D(0)]$ .
- ▶ Thus

$$\begin{aligned}\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0] &= \mathbb{E}[D(1) - D(0)] \\ &= \mathbb{E}[\mathbb{E}[D(1) - D(0)|\text{Types}]] \quad (\because \text{LIE}) \\ &= \mathbb{E}[\underbrace{D(1) - D(0)}_{=1} | \text{Complier}] \mathbb{P}(\text{Complier}) \\ &= \mathbb{P}(\text{Complier})\end{aligned}$$

## Proof of LATE theorem

- ▶ For the numerator,

$$\begin{aligned}\mathbb{E}[Y|Z=1] &= \mathbb{E}[Y(0) + D(Y(1) - Y(0))|Z=1] \\ &= \mathbb{E}[Y(0)|Z=1] + \mathbb{E}[D(1)(Y(1) - Y(0))|Z=1] \\ &= \mathbb{E}[Y(0)] + \mathbb{E}[(Y(1) - Y(0))D(1)] (\because \text{independence})\end{aligned}$$

- ▶ Similarly  $\mathbb{E}[Y|Z=0] = \mathbb{E}[Y(0)] + \mathbb{E}[(Y(1) - Y(0))D(0)]$ .
- ▶ Therefore

$$\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] = \mathbb{E}[(Y(1) - Y(0))(D(1) - D(0))]$$

and by conditioning on  $D(1) - D(0)$ , this equals to

$$\mathbb{P}(\text{Complier})\mathbb{E}[Y(1) - Y(0)|\text{Complier}]$$

## Proof of LATE theorem

► Hence

$$\begin{aligned} Wald &= \frac{\mathbb{P}(\text{Complier})\mathbb{E}[Y(1) - Y(0)|\text{Complier}]}{\mathbb{P}(\text{Complier})} \\ &= \mathbb{E}[Y(1) - Y(0)|\text{Complier}] = LATE. \end{aligned}$$

► The IV regression identifies the average treatment effect on the **compliers**.

**Corollary** If treatment effects are homogeneous, i.e.  $Y(1) - Y(0) = \delta$ , then

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

# Quasi-Experiments

- ▶ In social sciences research, experiments are almost impossible to implement.  
e.g. Can you conduct an experiment to investigate the causal effect of marriage on income?
- ▶ There are situations where the treatment is as good as random.
- ▶ They are called **natural experiments** or **quasi-experiments**.
- ▶ These include difference-in-differences (DiD) and regression discontinuity design (RDD).

# SUTVA

- ▶ In any causal inference, we (implicitly) assume **SUTVA**.
- ▶ SUTVA=Stable Unit Treatment Value Assumption.
- ▶ First, the treatment is received in homogeneous doses to all units. (does NOT mean the treatment effect is homogeneous)
- ▶ Second, there are no externalities: one's treatment does not affect the treatment on the others.  
*cf.* general equilibrium effect

# Difference-in-Differences

- ▶ Consider a binary treatment  $D$ .
- ▶ There are two time periods, say  $t = 0, 1$ .
- ▶ At  $t = 1$ , some units are treated and some are not.
- ▶ What is the effect of treatment on outcome variable  $Y$ ?
- ▶ Here  $Y_0$  denotes the outcome at  $t = 0$ ,  $Y_1$  the observed outcome at  $t = 1$  and  $Y_1(0)$ ,  $Y_1(1)$  the potential outcomes at  $t = 1$  so that

$$Y_1 = DY_1(1) + (1 - D)Y_1(0)$$

# Difference-in-Differences

- ▶ A naive estimator

$$\mathbb{E}[Y_1|D = 1] - \mathbb{E}[Y_1|D = 0]$$

- ▶ What if  $D$  is determined based on  $Y_0$ ?

e.g.  $D$ =public health center,  $Y$ =income level

- ▶ Introduces a selection bias!
- ▶ Compare the differences.
- ▶ DiD estimator

$$\mathbb{E}[Y_1 - Y_0|D = 1] - \mathbb{E}[Y_1 - Y_0|D = 0]$$

$$\begin{aligned} &= (\mathbb{E}[Y_1|D = 1] - \mathbb{E}[Y_0|D = 1]) - ((\mathbb{E}[Y_1|D = 0] - \mathbb{E}[Y_0|D = 0])) \\ &= (\mathbb{E}[Y_1|D = 1] - \mathbb{E}[Y_1|D = 0]) - ((\mathbb{E}[Y_0|D = 1] - \mathbb{E}[Y_0|D = 0])) \end{aligned}$$

# Difference-in-Differences

- ▶ In regression form,

$$\Delta Y_i = \beta_0 + \beta_1 D_i + u_i$$

or

$$Y_{it} = \beta_0 + \beta_1(D_i \cdot t) + \beta_2 t + \beta_3 D_i + u_i$$

- ▶ One can think of this as

$D$  = group fixed effect

$t$  = time fixed effect

- ▶ Automatically calculates the standard error!

## Parallel Trends

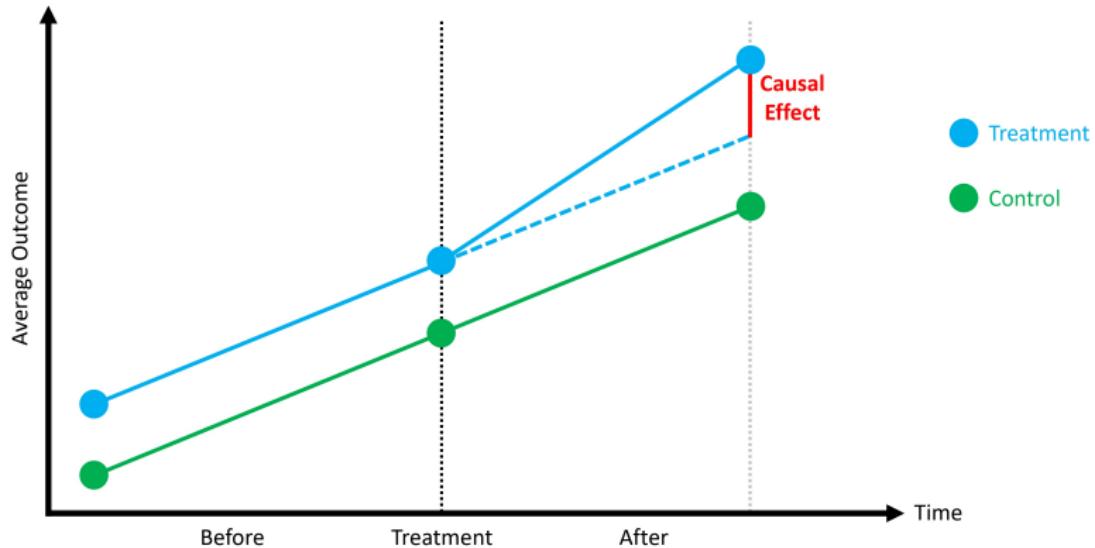
- ▶ What DiD estimator identifies?
- ▶ The DiD estimator

$$\begin{aligned} & \mathbb{E}[Y_1 - Y_0 | D = 1] - \mathbb{E}[Y_1 - Y_0 | D = 0] \\ &= \mathbb{E}[Y_1(1) | D = 1] - \mathbb{E}[Y_1(0) | D = 0] \\ &\quad - (\mathbb{E}[Y_0 | D = 1] - \mathbb{E}[Y_0 | D = 0]) \\ &= \underbrace{\mathbb{E}[Y_1(1) - Y_1(0) | D = 1]}_{ATT} \\ &\quad + \mathbb{E}[Y_1(0) - Y_0 | D = 1] - \mathbb{E}[Y_1(0) - Y_0 | D = 0] \end{aligned}$$

- ▶ If the last term vanishes, we say **parallel trends assumption** holds:

$$\underbrace{\mathbb{E}[Y_1(0) - Y_0 | D = 1]}_{\text{counterfactual}} = \mathbb{E}[Y_1(0) - Y_0 | D = 0]$$

# Parallel Trends



## Parallel Trends

- ▶ Parallel trends assumption means that the potential outcome of treatment group **would have changed parallel** to the control group if they were not treated.
- ▶ Testing parallel trends assumption is essentially impossible since it needs counterfactual information.
- ▶ Alternatively, one can use **pretrend analysis** to see if the parallel trends assumption makes sense.
- ▶ If the treatment group and control group showed different pretrend, it is not likely that the parallel trends hold.
- ▶ This argument is quite heuristic but widely used in practice.

# Regression Discontinuity Design

- ▶ Regression discontinuity design (RDD) aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned.
  - ▶ An intervention  $X_i$  is treated if and only if  $v_i \geq c_0$ .
  - ▶ We call  $v_i$  the running variable<sup>1</sup> and  $c_0$  the cutoff.
  - ▶ The key assumption is that the expected potential outcomes change continuously as a function of the running variable through the cutoff.
  - ▶ The only thing that causes the outcome to change abruptly at the cutoff is the treatment.
- e.g. Lee (2008). Randomized experiments from non-random selection in U.S. House elections, *Journal of Econometrics*.

---

<sup>1</sup>or assignment variable

# Regression Discontinuity Design

- ▶ Question: is there an “incumbency advantage” in the election?
  - ▶  $Y$ =probability of being elected in the next election
  - ▶  $X$ =(Democrat's) incumbency
  - ▶ Regressing  $Y$  on  $X$  causes endogeneity!
- e.g. unobserved competitiveness of politician.  
attractive politician → higher probability of winning elections
- ▶ In the US house election, there are (*de facto*) only two parties: Democrats and Republicans.

# Regression Discontinuity Design

- ▶  $v = (\text{vote share of Democrat}) - (\text{vote share of Republican})$
- ▶ Democrats win ( $X = 1$ )  $\iff v \geq 0$
- ▶ A discrete treatment  $X$  is assigned based on a continuous running variable  $v$  by its cutoff  $c_0 (= 0)$ .
- ▶ Regress

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 v_i + u_i$$

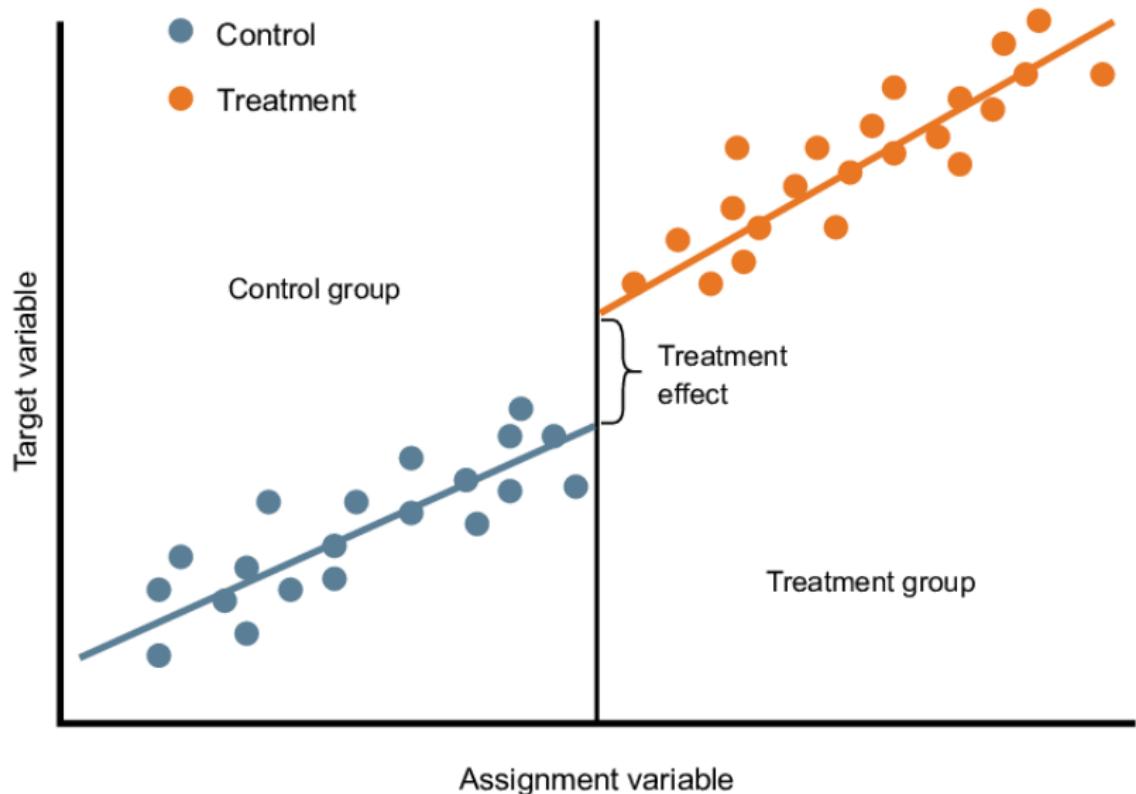
- ▶  $\mathbb{E}[Y_i | v_i = c] = \begin{cases} \beta_0 + \beta_1 + \beta_2 c & \text{if } c \geq 0 \\ \beta_0 + \beta_2 c & \text{if } c < 0 \end{cases}$

- ▶ Thus

$$\lim_{c \searrow c_0} \mathbb{E}[\underbrace{Y_i}_{=Y_i(1)} | v_i = c] - \lim_{c \nearrow c_0} \mathbb{E}[\underbrace{Y_i}_{=Y_i(0)} | v_i = c] = \beta_1$$

- ▶ The parameter  $\beta_1$  measures the degree of discontinuity.

# Regression Discontinuity Design



# Regression Discontinuity Design

- ▶ By RDD we estimate the local average treatment effect near the cutoff  $c_0$ .

$$\delta_{SRD} = \mathbb{E}[Y_i(1) - Y_i(0)|v = c_0]$$

- ▶ Sharp vs. Fuzzy RD
- ▶ Sometimes there is a discontinuity, but it's not entirely deterministic. i.e. there is an increase in the probability of treatment assignment.
- ▶ An IV technique can be used for estimation of fuzzy RDD.  
e.g. Wald-type estimator

$$\beta_{FRD} = \frac{\lim_{c \searrow c_0} \mathbb{E}[Y_i|v_i = c] - \lim_{c \nearrow c_0} \mathbb{E}[Y_i|v_i = c]}{\lim_{c \searrow c_0} \mathbb{E}[X_i|v_i = c] - \lim_{c \nearrow c_0} \mathbb{E}[X_i|v_i = c]}$$

# Summary

- ▶ Next lecture, we will cover additional causal inference methods.
- e.g. DiDiD, Nonparametric RDD, Synthetic control, Staggered DiD, Event study, Matching, Synthetic DiD, DiRD, etc.  
(tentative)

## Exercise

**RCT** Do Stock and Watson Ch13. Exercise E13.1. Download dataset from [https://www.princeton.edu/~mwatson/Stock-Watson\\_3u/Students/Stock-Watson-EmpiricalExercises-DataSets.htm](https://www.princeton.edu/~mwatson/Stock-Watson_3u/Students/Stock-Watson-EmpiricalExercises-DataSets.htm)

**RDD** With the provided dataset, run RDD regression where  $Y$  = outcome variable,  $X$  = running variable, cutoff=50.

- (a) Estimate the RDD we discussed.
- (b) Estimate

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \alpha(X \cdot D) + u$$

How does it differ from (a)? [Hint] Slope

- (c) Find the expression for the causal parameter RDD estimates using the coefficients of (b).

## Exercise

DiD We will simulate an example of DiD estimator.

Step 1 You **MUST** set seed number 1234 by

```
set.seed(1234)
```

and set observation number 500.

Step 2 Set treatment group by the following code

```
sample(0:1, size=500, replace=TRUE)
```

Step 3 Generate pretreatment outcome  $Y_0 \sim N(0, 1)$  and add 1 for treated individuals.

Step 4 Generate post-treatment potential outcomes by

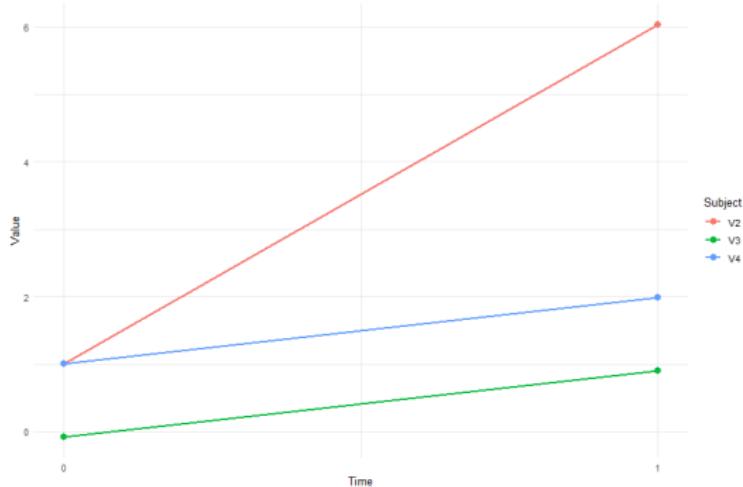
$$Y_1(0) = Y_0 + N(1, 1)$$

and

$$Y_1(1) = Y_1(0) + N(4, 1)$$

## Exercise

- ▶ Through Step 1-Step 4, what you have will be like



- (a) Calculate ATE and ATT.
- (b) Calculate the difference-in-means estimator.
- (c) Compute the DiD estimate. Compare it to the ATT.

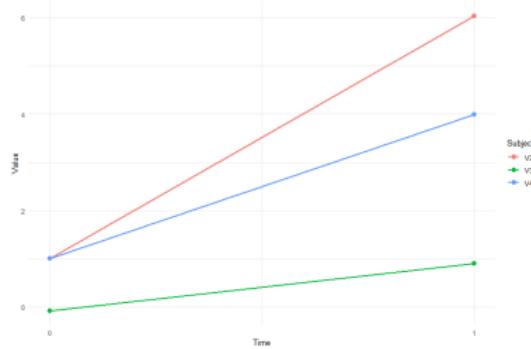
## Exercise

**Step 5** Now consider a situation where parallel trends assumption is violated. Replace the potential outcome if untreated by

$$Y_1(0) + 2 \cdot D;$$

while  $Y_1(1)$  has the same value as before.

- ▶ After Step 5 your data should look like



- (c) Repeat (a) and (b) for non-parallel situation.

# SW13. Experiments and Quasi-Experiments: Part 2

Ryu Myeonggyu

July 8, 2025

# Matching

Recall Conditional Independence Assumption (CIA)<sup>1</sup>

$$\{Y_i(0), Y_i(1)\} \perp D_i | X_i$$

- ▶ Other names: unconfoundedness, ignorability, selection on observables

Def Strong Ignorability: Ignorability + Overlap

$$p(x) := \mathbb{P}(D_i = 1 | X_i = x) \in (0, 1).$$

- ▶  $p(X)$  is called the **propensity score**.

---

<sup>1</sup>Here the symbol  $\perp$  means independence. In many contexts this means zero correlation and in this case,  $\perp\!\!\!\perp$  is used to mean independence.

# Matching

- ▶ Under strong ignorability, how can we find ATE?
- ▶ One can try multiple linear regression, but there is another (huge) literature: matching
- ▶ The key idea is similar: Use (strong) CIA
- ▶ Regression Adjustment (RA) or Regression Imputation
- ▶ Inverse Probability Weighting (IPW)
- ▶ Doubly Robust Estimators (AIPW, IPWRA)
- ▶ Exact Matching, Propensity Score Matching (PSM)

# Matching

- ▶ Regression Adjustment (RA)
- ▶ Notation: for  $d = 0, 1$ ,

$$\mu_d(x) = \mathbb{E}[Y_i(d)|X_i = x]$$

- ▶ Model  $\mu_d(x)$

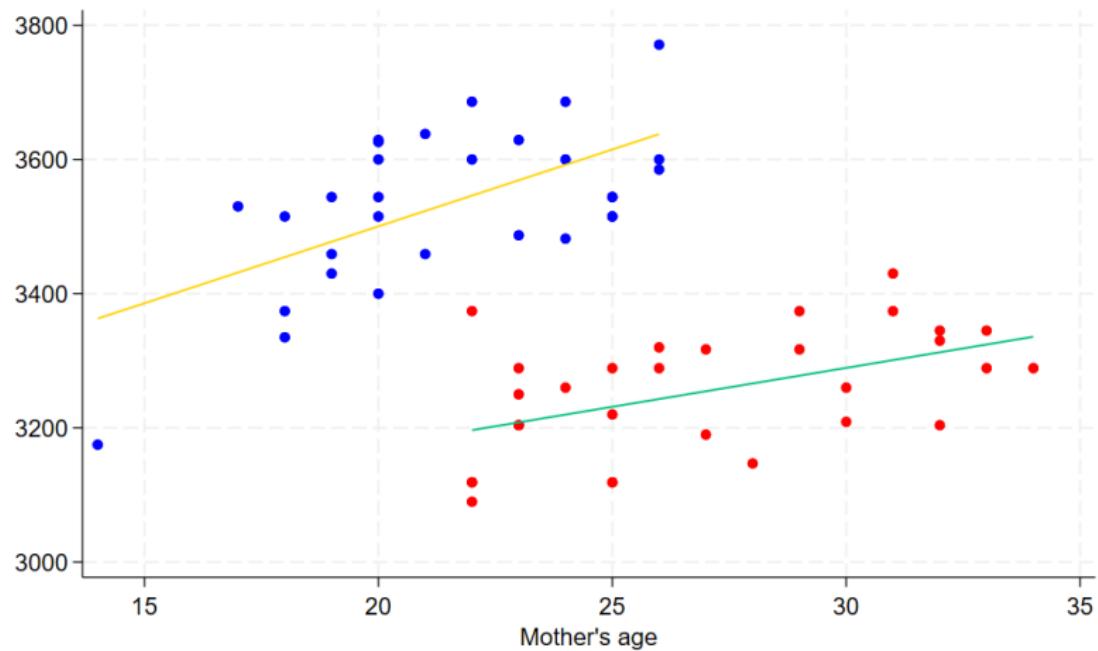
e.g.

$$\mu_d(X_i) = \beta_{0d} + \beta_{1d}X_i + u_{i,d}$$

- ▶ Obtain fitted values  $\hat{\mu}_d(X_i)$
- ▶ Estimate ATE by

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$$

# Regression Adjustment



# Matching

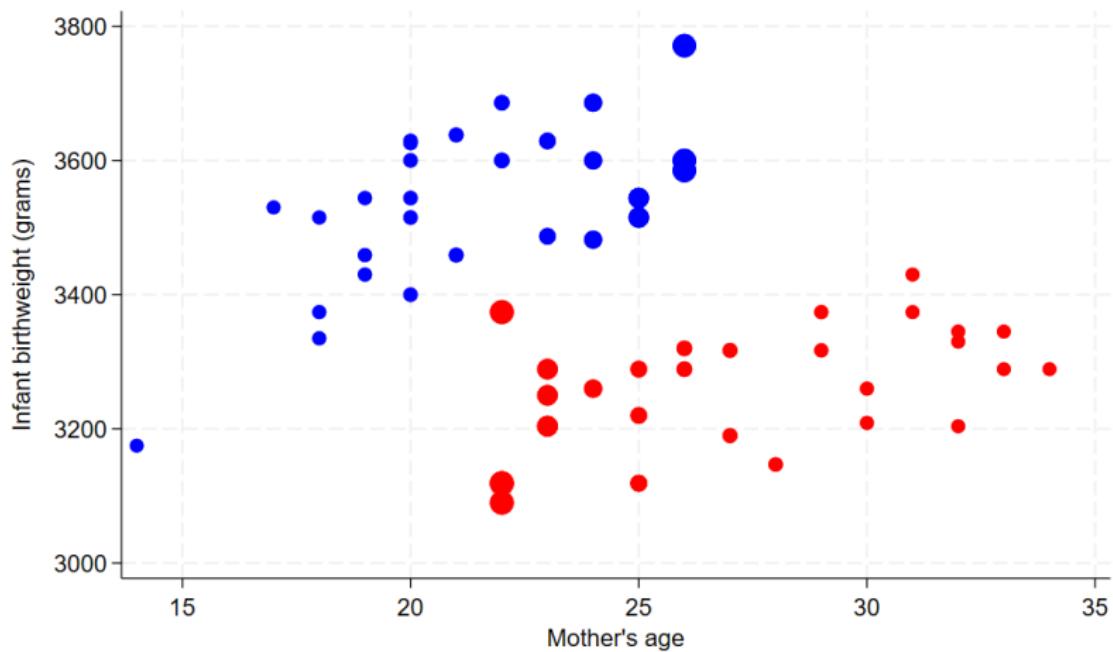
- ▶ Inverse Probability Weighting (IPW)
- ▶ Observe that

$$ATE = \mathbb{E} \left[ \frac{D_i Y_i}{p(X_i)} \right] - \mathbb{E} \left[ \frac{(1 - D_i) Y_i}{1 - p(X_i)} \right]$$

## Proof

- ▶  $p(X_i) = \mathbb{P}(D_i = 1|X_i)$  can be estimated using logit or probit models.

# Inverse Probability Weighting



# Doubly Robust Estimators

- ▶ Augmented IPW (AIPW), IPWRA
- ▶ Basically IPW+RA
- ▶ AIPW

Step 1 Estimate  $\hat{p}(X_i)$

Step 2 Estimate  $\hat{\mu}_d$  and get fitted values

Step 3 Obtain ATE by

$$\begin{aligned} AIPW = & \frac{1}{n} \sum_i \left[ \frac{D_i Y_i}{\hat{p}(X_i)} - \frac{D_i - \hat{p}(X_i)}{\hat{p}(X_i)} \hat{\mu}_1(X_i) \right] \\ & - \frac{1}{n} \sum_i \left[ \frac{(1 - D_i) Y_i}{1 - \hat{p}(X_i)} - \frac{D_i - \hat{p}(X_i)}{1 - \hat{p}(X_i)} \hat{\mu}_0(X_i) \right] \end{aligned}$$

- ▶ This gives a consistent estimator of ATE at least one of  $\hat{p}(X_i)$  or  $(\hat{\mu}_0, \hat{\mu}_1)$  is correctly specified.

# Proof

# Doubly Robust Estimators

## ► IPWRA

Step 1 Estimate  $\hat{p}(X_i)$

Step 2 Using IPW, fit  $\hat{\mu}_d$  by weighted regression.

$$\widehat{ATE} = \frac{1}{n} \sum_i [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$$

# Matching

- ▶ Matching
- ▶ Matching estimators impute the missing potential outcomes using only the outcomes of a few nearest neighbors of the opposite treatment group.
- ▶ For each  $Y_i$ , match closest observations with respect to the covariates  $X$ .
- ▶ 1: $m$ -matching
- ▶ with/without replacement
- ▶ How can we measure “closeness” ?

# Matching

- ▶ Exact matching finds observations that have exactly same covariates.
- ▶ Often there are no feasible matches.
- ▶ Coarsened Exact Matching (CEM): coarsen the covariates
  - e.g. age → age groups (18-29, 30-39, 40-49, ...)
  - e.g. income → income quartile

# Matching

- ▶ If covariate  $X$  is one-dimensional, one can easily measure distances.
- ▶ If  $X$  is multi-dimensional, the choice of distance  $\|X_i - X_j\|$  matters:
- ▶ (Normalized) Euclidean

$$\|X_i - X_j\|^2 = \sum_{k=1}^n (X_{ik} - X_{jk})^2 \left( = \sum_{k=1}^n \left( \frac{X_{ik} - X_{jk}}{\sigma_k} \right)^2 \right)$$

- ▶ Mahalanobis (1936)

$$\|X_i - X_j\|_M^2 = (X_i - X_j)' \Sigma_X^{-1} (X_i - X_j)$$

where  $\Sigma_X$  is the covariance matrix of  $X$ .

# Matching

- ▶ One remarkable method: Propensity Score Matching (PSM)
- ▶ Rosenbaum, Rubin (1983)
- ▶ Strong Ignorability implies

(a)

$$\{Y_i(0), Y_i(1)\} \perp D_i | p(X_i)$$

(b)

$$\mathbb{P}(D_i = 1 | p(X_i)) \in (0, 1)$$

# Proof

- ▶ Implication: one only needs to control the propensity score.

# Matching

- ▶ PSM has become the prevalent method in the matching literature.
- ▶ Estimator of ATT

$$ATT = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} (Y_i(1) - \hat{Y}_i(0))$$

where

$$\hat{Y}_i(0) = \sum_{j \in I_0} \hat{\omega}(i, j) Y_j(0)$$

- ▶  $I_0, I_1$ : the set of control/treatment groups
- ▶  $S_P$ : common support
- ▶  $\hat{\omega}(i, j)$ : weights which depend on the distance of propensity scores for observations  $i$  and  $j$ .

# Matching

- ▶ Nearest neighbor (k-)matching

$$\hat{\omega}(i,j) = \begin{cases} 1/k, & \text{for closest } k \text{ observations } j \text{ w.r.t. } |\hat{p}_i - \hat{p}_j| \\ 0, & \text{otherwise} \end{cases}$$

- ▶ Caliper

$$\hat{\omega}(i,j) = \begin{cases} 1/n_i & \text{if } |\hat{p}_i - \hat{p}_j| < c \\ 0, & \text{otherwise} \end{cases}$$

where  $n_i$ =number of matches for observation  $i$ .

- ▶ Stratification (or Blocking): Consider a partition

$$0 = t_0 < t_1 < \cdots < t_T = 1$$

Then

$$\hat{\omega}(i,j) = \begin{cases} 1/n_i & \text{if } \hat{p}_j \in [t_i, t_{i+1}) \\ 0, & \text{otherwise} \end{cases}$$

# Matching

- ▶ Hwang, Kim, Lee (2025). Is Job Loss Always Bad for Health? Evidence from National Health Screening, *Review of Economics and Statistics*.
- ▶ Impact of job loss on health outcomes
- ▶ Reverse Causality, OVB
- ▶ Uses Mass Layoffs during Global Recession as natural experiments.
- ▶ Perform 1:1 (exact) matching with replacement, using individual characteristics at reference month such as age, industry sector, firm size, and insurance fee.
- ▶ Matching rate: 96.5% for men and 95.9% for women

# Matching

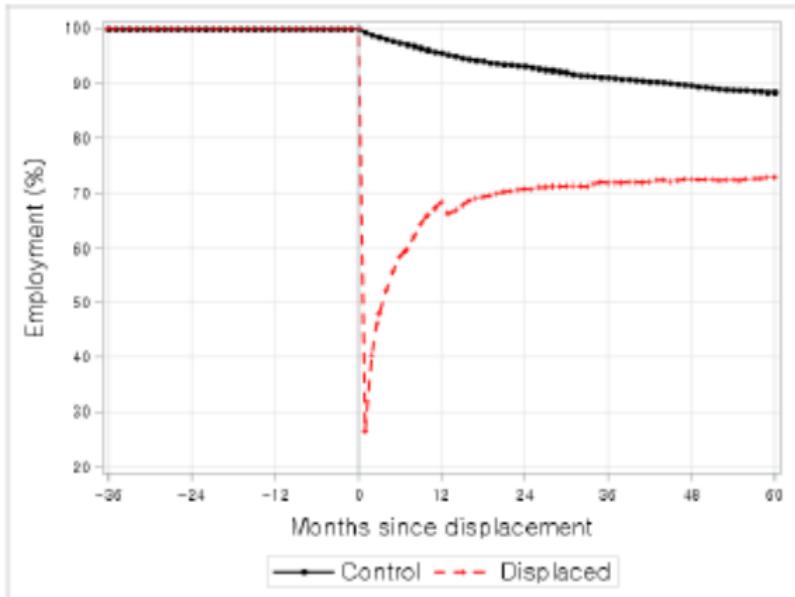
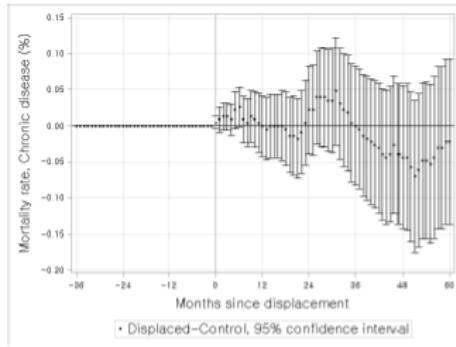


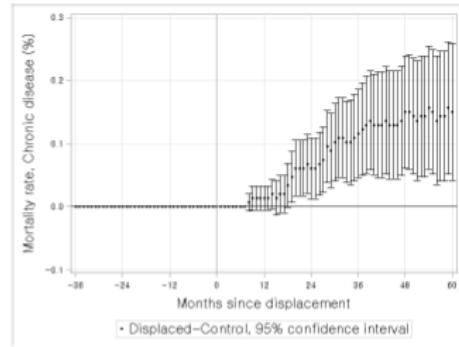
Figure: Employment, Men

# Matching

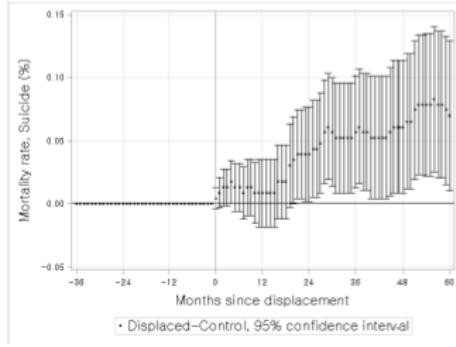
Panel B1. Chronic disease, Men



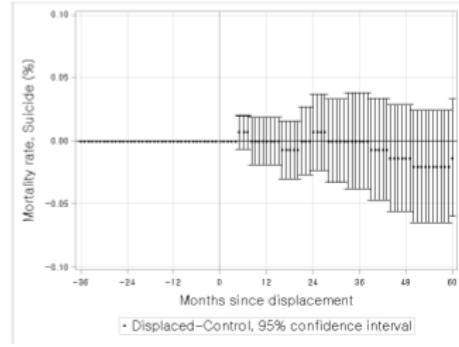
Panel B2. Chronic disease, Women



Panel C1. Suicide, Men



Panel C2. Suicide, Women



- ▶  $Y_i$  =(average) health status of region  $i$
  - ▶  $D_i$  =(local) public health center
- Q Heterogeneous effect by income level?
- ▶ The lower income group may benefit a lot from the public health center more than the high income group.

- ▶ Regression model

$$\begin{aligned}Y_{igt} = & \beta_0 + \beta_1 D + \beta_2 H + \beta_3 Post + \beta_4 (D \cdot Post) \\& + \beta_5 (D \cdot H) + \beta_6 (H \cdot Post) + \beta_7 (D \cdot H \cdot Post) + u_{igt}\end{aligned}$$

where  $g \in \{\text{high, low}\}$ ,  $i$  indexes individual,  $t \in \{0, 1\}$  and  $H = \mathbb{1}(High)$ .

- ▶ Compute

$$\mathbb{E}[Y_{igt}|D, H, Post]$$

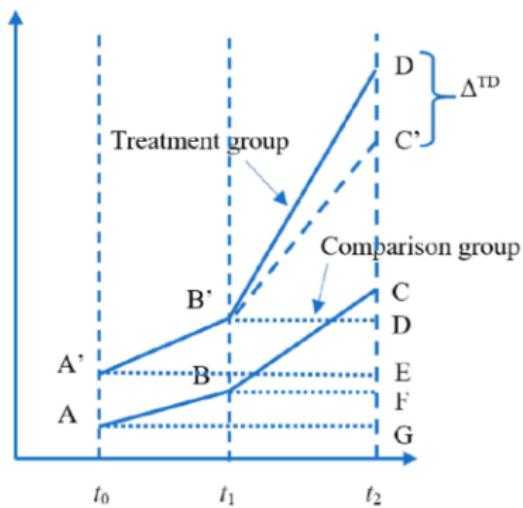
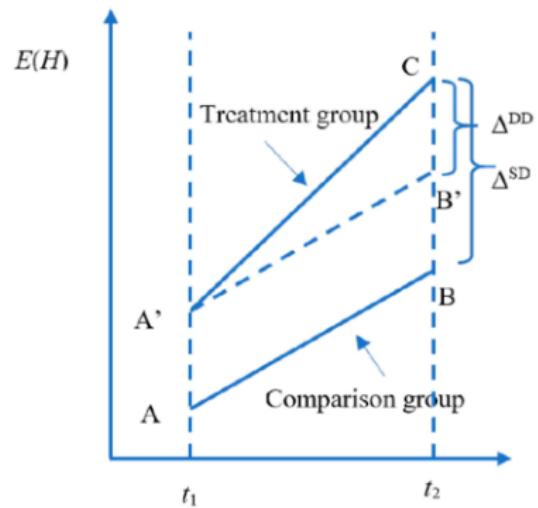
for each value of  $D, H, Post$ .

$$\begin{aligned}\beta_7 &= \mathbb{E}[Y_{igt}|D = 1, H = 1, Post = 1] - \mathbb{E}[Y_{igt}|D = 1, H = 1, Post = 0] \\&\quad - (\mathbb{E}[Y_{igt}|D = 1, H = 0, Post = 1] - \mathbb{E}[Y_{igt}|D = 1, H = 0, Post = 0]) \\&\quad - (\mathbb{E}[Y_{igt}|D = 0, H = 1, Post = 1] - \mathbb{E}[Y_{igt}|D = 0, H = 1, Post = 0]) \\&\quad + (\mathbb{E}[Y_{igt}|D = 0, H = 0, Post = 1] - \mathbb{E}[Y_{igt}|D = 0, H = 0, Post = 0]) \\&= [\mathbb{E}[Y_{igt}|D = 1, H = 1, Post = 1] - \mathbb{E}[Y_{igt}|D = 1, H = 1, Post = 0]] \\&\quad - (\mathbb{E}[Y_{igt}|D = 0, H = 1, Post = 1] - \mathbb{E}[Y_{igt}|D = 0, H = 1, Post = 0]) \\&\quad - [(\mathbb{E}[Y_{igt}|D = 1, H = 0, Post = 1] - \mathbb{E}[Y_{igt}|D = 1, H = 0, Post = 0]) \\&\quad - (\mathbb{E}[Y_{igt}|D = 0, H = 0, Post = 1] - \mathbb{E}[Y_{igt}|D = 0, H = 0, Post = 0])]\end{aligned}$$

- ▶ For the second equality, the first four terms are DiD for high group and the last four terms are DiD for low group.
- ▶ Notice that

$$\text{DiD}_g = \text{ATT}_g + \text{Non-Parallel-Trends}_g$$

- ▶ From this decomposition, we see that  $\beta_7$  basically identifies the difference of ATT by group. ( $\Delta\text{ATT}$ )
- ▶ For identification of  $\Delta\text{ATT}$ , one need the assumption that the two non-parallel trends of two groups cancel out.
- ▶ Called the common parallel trends assumption.
- ▶ The differences in trends between the treated and control groups would have remained parallel in the absence of the treatment.



# Synthetic Control

- ▶ Abadie, Diamond, Hainmueller (2015). Comparative Politics and the Synthetic Control Method, *American Journal of Political Science*.
- ▶ Impact of Reunification on German Economy
- ▶ Impute “synthetic” (West) Germany by “convex” combination of countries in the rest of the world.
- ▶ linear combination vs. convex combination

## Synthetic Control

- ▶ Assume that we have a balanced panel for  $t = 1, \dots, T$  time periods and  $j = 1, 2, \dots, J + 1$  units. ( $j = 1$  is West Germany)
- ▶ The sample includes a positive number of pretreatment periods,  $T_0$ , as well as a positive number of posttreatment periods,  $T_1$ , with  $T = T_0 + T_1$ .
- ▶  $X_1$  is a  $k \times 1$  vector containing the values of the pretreatment characteristics of the treated unit.
- ▶  $X_0$  is the  $k \times J$  matrix collecting the values of the same variables for the units in the control group.
- ▶ Let  $W = (w_2, \dots, w_{J+1})'$  be a vector of nonnegative real numbers such that

$$w_2 + \dots + w_{J+1} = 1$$

# Synthetic Control

- ▶ We choose the weight  $W^*$  such that

$$W^* = \arg \min_W \|X_1 - X_0 W\|_V^2 = \sum_{m=1}^k v_m (X_{1m} - X_{0m} W)^2$$

where  $V$  is a positive semi-definite diagonal matrix.<sup>2</sup>

- ▶ In Abadie et al. (2015), the  $X$ 's include per capita GDP, inflation rate, industry share of value added, investment rate, schooling, and a measure of trade openness.
- ▶ The hyperparameters  $v_m$ 's are selected to minimize RMSPE using cross-validation.

---

<sup>2</sup>Mathematically  $V$  need not be diagonal.  $V$  generates a seminorm  $\|X\|_V = \sqrt{X' V X}$

# Synthetic Control

TABLE 1 Synthetic and Regression Weights for West Germany

Country	Synthetic Control Weight	Regression Weight	Country	Synthetic Control Weight	Regression Weight
Australia	0	0.12	Netherlands	0.09	0.14
Austria	0.42	0.26	New Zealand	0	0.12
Belgium	0	0	Norway	0	0.04
Denmark	0	0.08	Portugal	0	-0.08
France	0	0.04	Spain	0	-0.01
Greece	0	-0.09	Switzerland	0.11	0.05
Italy	0	-0.05	United Kingdom	0	0.06
Japan	0.16	0.19	United States	0.22	0.13

Notes: The synthetic weight is the country weight assigned by the synthetic control method. The regression weight is the weight assigned by linear regression. See text for details.

# Synthetic Control

FIGURE 1 Trends in per Capita GDP: West Germany versus Rest of the OECD Sample

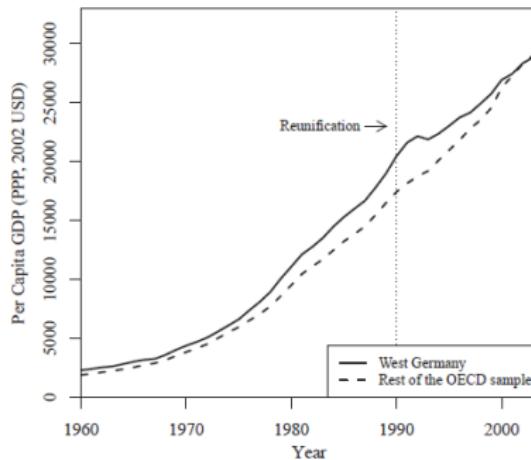
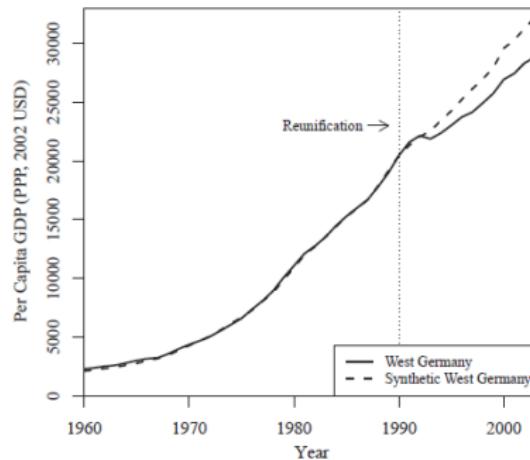


FIGURE 2 Trends in per Capita GDP: West Germany versus Synthetic West Germany



## Next Lecture

- ▶ We will cover more advanced topics of causal inference methods.
- ▶ TWFE-DiD, Staggered DiD, DiRD  
(Difference-in-Regression-Discontinuity) and S-DiD.

## Exercise

- ▶ Suppose that there are  $p$  dummy variables  $X_1, \dots, X_p$ .
- (a) What is the maximum number of distinct values that the conditional expectation function  $\mathbb{E}[Y|X_1, \dots, X_p]$  can take?
- (b) Let  $M$  be the answer to (a). A linear regression model which includes  $M$  variables is called a saturated dummy variable regression model. Is the DiDiD regression model saturated?
- (c) Verify the expression for  $\beta_7$ .

## Exercise

- ▶ In the DiDiD setup, suppose that the high group is not affected by treatment.
- ▶ Then what does DiDiD estimate?

# SW13. Experiments and Quasi-Experiments: Part 3

Ryu Myeonggyu

July 11, 2025

# Contents

TWFE-DiD in Staggered Adoption

Synthetic DiD

Difference-in-Discontinuities

## TWFE-DiD

- ▶ Recall that the Difference-in-differences estimator can be estimated by the following regression

$$Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 \cdot t + \beta_3 (D_i \cdot t) + u_{it}$$

where  $t = 0, 1$ .

- ▶ Notice that this model is equivalent to

$$Y_{it} = \beta_0 + \alpha_i + \gamma_t + \beta_3 D_{it} + u_{it}$$

for  $t = 0, 1$  and  $D_{it} := D_i \cdot t$ .

- ▶ We call this model **the canonical DiD**.
- ▶ The canonical DiD estimator between the treated group and the untreated group is

$$\hat{\beta}_{TU}^{2 \times 2} = (\bar{Y}_T^{POST} - \bar{Y}_T^{PRE}) - (\bar{Y}_U^{POST} - \bar{Y}_U^{PRE})$$

- ▶ The more common situation is one where geographic units receive treatments at different points in time.
- ▶ Now treatments are given within time periods  $t = 1, 2, \dots, T$  and we have a panel data

$$\{Y_{it}, D_{it} | i = 1, \dots, n, t = 1, \dots, T\}$$

- ▶ A natural extension analogous to the above formula is

$$Y_{it} = \beta_0 + \alpha_i + \gamma_t + \beta_3 D_{it} + u_{it}$$

for  $t = 1, 2, \dots, T$ .

- ▶ This estimator is called the **TWFE-DiD**.

# TWFE-DiD

- ▶ TWFE-DiD has been widely accepted as an intuitive extension of DiD where  $T > 2$ .
- ▶ However, recent econometric research casts doubt on this heuristics.

**Notation** There are  $N$  units and  $T$  time periods. There are two treatment timings:  $k$  and  $\ell$  with  $k < \ell$ .

- ▶ We will denote by  $U$  the untreated group.
- ▶ Each timing group's sample share is

$$\eta_j := \sum_i \mathbb{1}_{\{t_i=j\}} / N$$

and the share of time it spends treated is

$$\bar{D}_j := \sum_t \mathbb{1}_{\{t \geq k\}} / T$$

- ▶ The sample mean of  $Y_{it}$  for units treated at time  $b$  during the post period for treatment day  $a$  is:

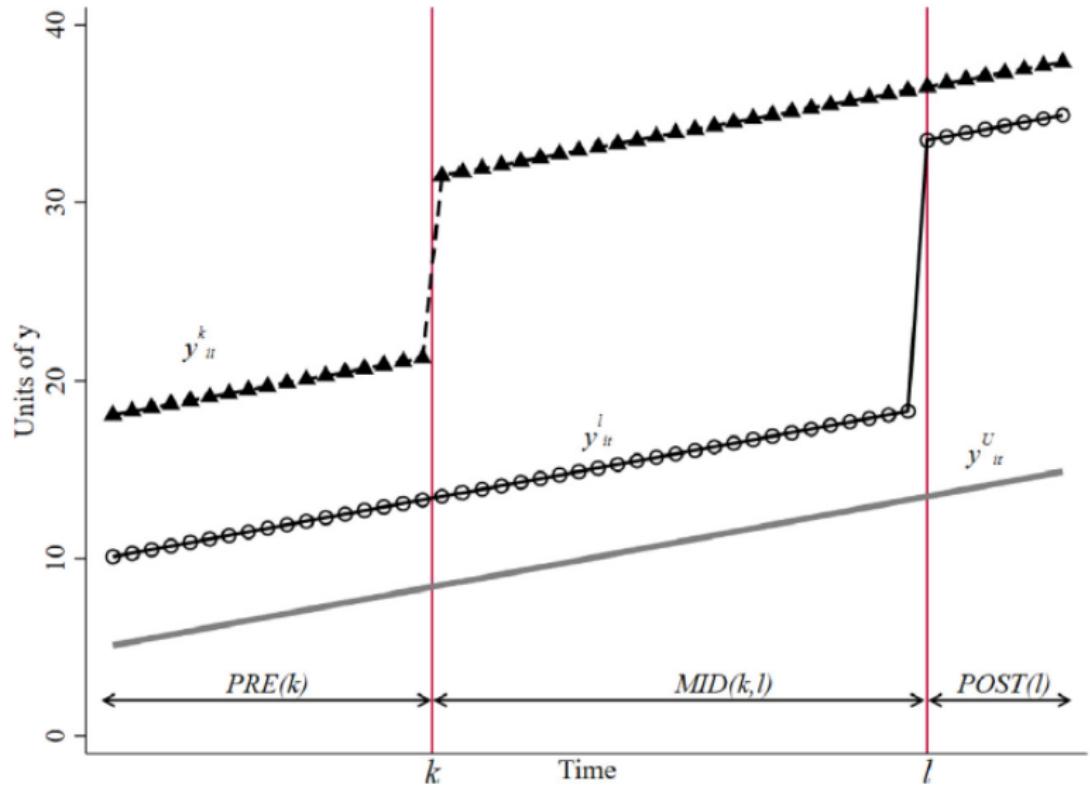
$$\bar{Y}_b^{POST(a)} := \frac{1}{T-a+1} \sum_{t=a}^T \frac{\sum_i Y_{it} \mathbb{1}\{t_i = b\}}{\sum_i \mathbb{1}\{t_i = b\}}$$

- ▶ Similarly,

$$\bar{Y}_b^{PRE(a)} := \frac{1}{a-1} \sum_{t=1}^{a-1} \frac{\sum_i Y_{it} \mathbb{1}\{t_i = b\}}{\sum_i \mathbb{1}\{t_i = b\}}$$

$$\bar{Y}_b^{MID(a)} := \frac{1}{a-b} \sum_{t=a}^{b-1} \frac{\sum_i Y_{it} \mathbb{1}\{t_i = b\}}{\sum_i \mathbb{1}\{t_i = b\}}$$

# TWFE-DiD



## TWFE-DiD

- ▶ By the FWL theorem, one can show that

$$\hat{\beta}_{TWFEDiD} = \frac{\sum_{i,t} Y_{it} \tilde{D}_{it}}{\sum_{i,t} \tilde{D}_{it}^2}$$

where  $\tilde{x}_{it} = (x_{it} - \bar{x}_i) - (\bar{x}_t - \bar{\bar{x}})$ .

- ▶ If there was single treatment timing, say  $j$ , then this reduces to the canonical DiD estimator

$$\hat{\beta}_{jU}^{2\times 2} = (\bar{Y}_j^{POST(j)} - \bar{Y}_j^{PRE(j)}) - (\bar{Y}_U^{POST(j)} - \bar{Y}_U^{PRE(j)})$$

- ▶ In the staggered design, the TWFE-DiD estimator is a weighted average of  $2 \times 2$  canonical DiD's between all possible comparisons.

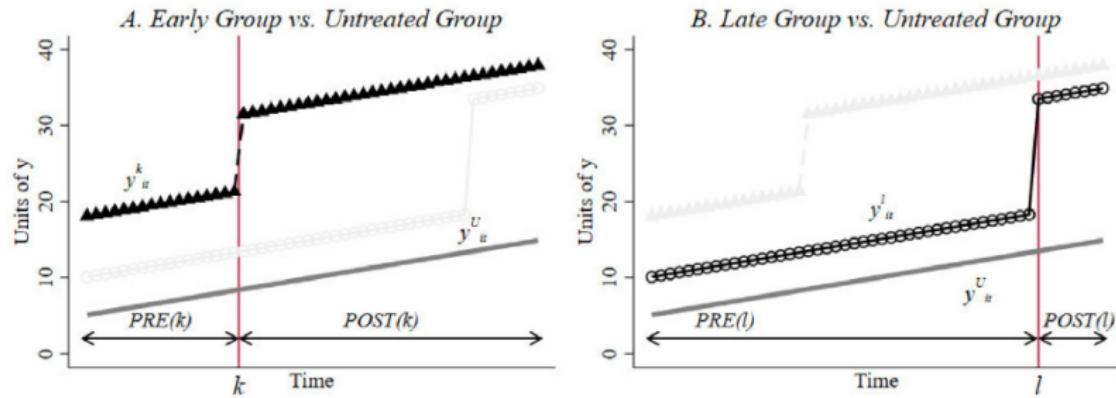


Figure: Comparisons to Untreated Groups

- $\hat{\beta}_{jU}^{2 \times 2} = (\bar{Y}_j^{POST(j)} - \bar{Y}_j^{PRE(j)}) - (\bar{Y}_U^{POST(j)} - \bar{Y}_U^{PRE(j)})$  for  $j = k, \ell$

# TWFE-DiD

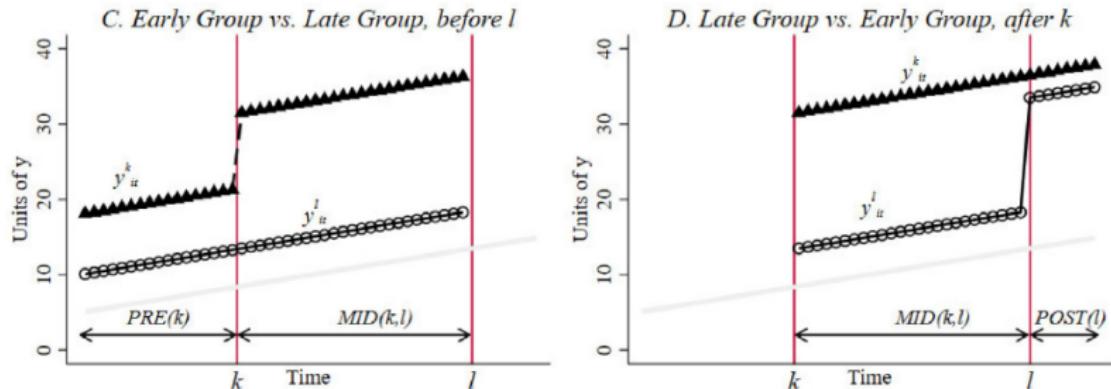


Figure: Comparisons between Early Treated vs. Later Treated

- ▶  $\hat{\beta}_{kl}^{2 \times 2}(k) = (\bar{Y}_k^{MID(k,\ell)} - \bar{Y}_k^{PRE(k)}) - (\bar{Y}_\ell^{MID(k,\ell)} - \bar{Y}_\ell^{PRE(j)})$
- ▶  $\hat{\beta}_{kl}^{2 \times 2}(\ell) = (\bar{Y}_\ell^{POST(\ell)} - \bar{Y}_\ell^{MID(k,\ell)}) - (\bar{Y}_k^{POST(\ell)} - \bar{Y}_k^{MID(k,\ell)})$

# TWFE-DiD

**Thm** (Bacon decomposition theorem) Assume that the data contain  $k = 1, \dots, K$  timing groups of units ordered by the time when they receive a binary treatment  $k \in (1, T]$ . The OLS estimate  $\hat{\beta}_{TWFEDiD}$  is a weighted average of all possible  $2 \times 2$  DiD estimators:

$$\hat{\beta}_{TWFEDiD} = \sum_{k \neq U} s_{kU} \hat{\beta}_{kU}^{2 \times 2} + \sum_{k \neq U} \sum_{\ell > k} [s_{k\ell}(k) \hat{\beta}_{k\ell}^{2 \times 2}(k) + s_{k\ell}(\ell) \hat{\beta}_{k\ell}^{2 \times 2}(\ell)]$$

where the weights are:

$$s_{kU} = \frac{(n_k + n_U)^2 n_{kU} (1 - n_{kU}) \bar{D}_k (1 - \bar{D}_k)}{(1/NT) \sum_{i,t} \tilde{D}_{it}^2}$$

$$\text{with } n_{kU} = \frac{n_k}{n_k + n_U}$$

# TWFE-DiD

$$s_{k\ell}(k) = \frac{((n_k + n_\ell)(1 - \bar{D}_\ell))^2 n_{k\ell} (1 - n_{k\ell})}{(1/NT) \sum_{i,t} \tilde{D}_{it}^2} \frac{\bar{D}_k - \bar{D}_\ell}{1 - \bar{D}_\ell} \frac{1 - \bar{D}_k}{1 - \bar{D}_\ell}$$

and

$$s_{k\ell}(\ell) = \frac{((n_k + n_\ell)\bar{D}_k)^2 n_{k\ell} (1 - n_{k\ell})}{(1/NT) \sum_{i,t} \tilde{D}_{it}^2} \frac{\bar{D}_\ell}{\bar{D}_k} \frac{\bar{D}_k - \bar{D}_\ell}{\bar{D}_k}$$

with

$$\sum_{k \neq U} s_{kU} + \sum_{k \neq U} \sum_{\ell > k} [s_{k\ell}(k) + s_{k\ell}(\ell)] = 1.$$

- ▶ Notice that the weights are determined by (i) the timing of treatment (ii) sample size

## TWFE-DiD

- ▶ Now, what does TWFE-DiD estimator identifies?
- ▶ We set some notations.
- ▶ Let  $Y_{it}(k)$  be the potential outcome of unit  $i$  in period  $t$  if it is treated at  $t_i = k$ .  $Y_{it}(0)$  is the untreated potential outcome.
- ▶ The observed outcome is

$$Y_{it} = D_{it} Y_{it}(t_i) + (1 - D_{it}) Y_{it}(0)$$

- ▶ The ATT for timing group  $k$  at time  $\tau \geq k$  is

$$ATT_k(\tau) := \mathbb{E}[Y_{i\tau}(t_k) - Y_{i\tau}(0) | t_i = k]$$

- ▶ Average ATT for time range  $W$  with window size  $T_W$  is

$$ATT_k(W) := \frac{1}{T_W} \sum_{t \in W_1} \mathbb{E}[Y_{it}(k) - Y_{it}(0) | t_i = k].$$

# TWFE-DiD

- ▶ Finally, define the difference over time in average untreated potential outcomes as

$$\Delta Y_k^0(W_1, W_0) := \frac{1}{T_{W_1}} \sum_{t \in T_{W_1}} \mathbb{E}[Y_{it}(0) | t_i = k] - \frac{1}{T_{W_0}} \sum_{t \in T_{W_0}} \mathbb{E}[Y_{it}(0) | t_i = k]$$

- ▶ Using these notations, we can write the  $2 \times 2$  DiD estimators as:

$$\begin{aligned}\hat{\beta}_{kU}^{2 \times 2} &= ATT_k(POST(k)) + [\Delta Y_k^0(POST(k), PRE(k)) \\ &\quad - \Delta Y_U^0(POST(k), PRE(k))]\end{aligned}$$

$$\begin{aligned}\hat{\beta}_{k\ell}^{2 \times 2}(k) &= ATT_k(MID(k, \ell)) + [\Delta Y_k^0(MID(k, \ell), PRE(k)) \\ &\quad - \Delta Y_\ell^0(MID(k, \ell), PRE(k))]\end{aligned}$$

$$\begin{aligned}\hat{\beta}_{k\ell}^{2 \times 2}(\ell) &= ATT_\ell(POST(\ell)) + [\Delta Y_\ell^0(POST(\ell), MID(k, \ell)) \\ &\quad - \Delta Y_k^0(POST(\ell), MID(k, \ell))] \\ &\quad - [ATT_k(POST(\ell)) - ATT_k(MID(k, \ell))]\end{aligned}$$

# TWFE-DiD

- ▶ Combining this with the Bacon decomposition, we finally get

$$\beta_{TWFEDiD} = p \lim_{N \rightarrow \infty} \hat{\beta}_{TWFEDiD} = VWATT + VWCT + \Delta ATT$$

- ▶ VWATT=variance weighted ATT

$$VWATT = \sum_{k \neq U} \sigma_{kU} ATT_k(POST(k))$$

$$+ \sum_{k \neq U} \sum_{\ell > k} [\sigma_{k\ell}(k) ATT_k(MID(k, \ell)) + \sigma_{k\ell}(\ell) ATT_k(POST(\ell))]$$

where  $\sigma = p \lim s$ .

► VWCT=Variance Weighted Common Trends

$$\begin{aligned} VWCT = & \sum_{k \neq U} \sigma_{kU} \left[ \Delta Y_k^0(\text{POST}(k), \text{PRE}(k)) - \Delta Y_U^0(\text{POST}(k), \text{PRE}(k)) \right] \\ & + \sum_{k \neq U} \sum_{\ell > k} \left[ \sigma_{k\ell}(k) \left\{ \Delta Y_k^0(\text{MID}(k, \ell), \text{PRE}(k)) - \Delta Y_\ell^0(\text{MID}(k, \ell), \text{PRE}(k)) \right\} \right. \\ & \quad \left. + \sigma_{k\ell}(\ell) \left\{ \Delta Y_\ell^0(\text{POST}(\ell), \text{MID}(k, \ell)) - \Delta Y_k^0(\text{POST}(\ell), \text{MID}(k, \ell)) \right\} \right] \end{aligned}$$

- Finally the last term equals the weighted sum of the change in treatment effects within each timing group's before and after a later treatment time

$$\Delta ATT = \sum_{k \neq U} \sum_{\ell > k} \sigma_{k\ell}(\ell) \underbrace{[ATT_k(\text{POST}(\ell)) - ATT_k(\text{MID}(k, \ell))]}_{\text{heterogeneity over time}}$$

- ▶  $\Delta ATT$  equals zero if ATT's are constant, but when they are not, they lead to bias even when VWCT = 0.
- ▶ There can be two sources of heterogeneity of ATT's: across-group and over-time.
- ▶ In this case, what matters is the heterogeneity over time.
- ▶ Time-varying treatment effects, even if they are identical across units, generate cross-group heterogeneity because of the differing post-treatment windows, and the fact that earlier-treated groups are serving as controls for later-treated groups.

- ▶ There are several estimators to correct for the bias.
- ▶ de Chaisemartin and D'Haultfœuille (2020), Callaway and Sant'Anna (2021), Sun and Abraham (2021), Borusyak, Jaravel and Speiss (2023)
- ▶ The key idea of these estimators is: avoid comparison of later treated units to the early treated units as control group.
  - ▶ the **forbidden comparison**

## Synthetic DiD

- ▶ Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager, (2021). Synthetic Difference-in-Differences. *American Economic Review*
- ▶ Synthetic control+DiD
- ▶ DiD needs parallel trends assumption.
- ▶ Synthetic control method (SCM) applies to the case where only a single (or small number) of units exposed and needs a lot of data of pretreatment period.
- ▶ Like SCM, SDID reweights and matches pre-exposure trends to weaken the reliance on parallel trend type assumptions.
- ▶ Like DID, SDID is invariant to additive unit-level shifts, and allows for valid large-panel inference.

## Synthetic DiD

- ▶ Consider a balanced panel data and binary treatment.
- ▶ The first  $1, \dots, N_c$  units are never exposed to the treatment. (control)
- ▶ The last  $N_c + 1, \dots, N$  ( $\#N_t$ ) units are exposed to treatment, after time  $T_{PRE}$ .
- ▶ Similar to SCM, we find weights  $\hat{\omega}^{sdid}$  that align pre-exposure trends in the outcome of unexposed units with those for the exposed units

$$\sum_{i=1}^{N_c} \hat{w}_i^{sdid} Y_{it} \approx \frac{1}{N_t} \sum_{i=1+N_c}^N Y_{it}$$

## Synthetic DiD

- ▶ We also look for time weights  $\hat{\lambda}_t^{sdid}$  sdid that balance pre-exposure time periods with postexposure ones.
- ▶ Using these weights, run a TWFE regression to estimate the ATE  $\tau$

$$(\hat{\tau}^{sdid}, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\tau, \mu, \alpha, \beta} \left[ \sum_{i,t} (Y_{it} - \mu - \alpha_i - \beta_t - D_{it}\tau)^2 \hat{\omega}_i^{sdid} \hat{\lambda}_t^{sdid} \right]$$

## DiD

$$(\hat{\tau}^{did}, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\tau, \mu, \alpha, \beta} \left[ \sum_{i,t} (Y_{it} - \mu - \alpha_i - \beta_t - D_{it}\tau)^2 \right]$$

## SCM

$$(\hat{\tau}^{sc}, \hat{\mu}, \hat{\beta}) = \arg \min_{\tau, \mu, \beta} \left[ \sum_{i,t} (Y_{it} - \mu - \beta_t - D_{it}\tau)^2 \hat{\omega}_i^{scm} \right]$$

# Synthetic DiD

- ▶ The time weights  $\hat{\lambda}_t^{sdid}$  are chosen such that

$$(\hat{\lambda}_0, \hat{\lambda}^{sdid}) = \arg \min_{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda} \ell(\lambda_0, \lambda)$$

where

$$\ell(\lambda_0, \lambda) = \sum_{i=1}^{N_c} \left( \lambda_0 + \sum_{t=1}^{T_{PRE}} \lambda_t Y_{it} - \frac{1}{T_{POST}} \sum_{t=1+T_{PRE}}^T Y_{it} \right)^2$$

and  $\Lambda = \{\sum_{t=1}^{T_{PRE}} \lambda_t = 1, \lambda_t = T_{POST}^{-1} \text{ for } t > T_{PRE}, \lambda_t \geq 0\}$

# Synthetic DiD

---

## ALGORITHM 1—SDID

---

Data:  $\mathbf{Y}, \mathbf{W}$

Result: Point estimate  $\hat{\tau}^{sdid}$

1. Compute regularization parameter  $\zeta$  using (5);
2. Compute unit weights  $\hat{\omega}^{sdid}$  via (4);
3. Compute time weights  $\hat{\lambda}^{sdid}$  via (6);
4. Compute the SDID estimator via the weighted DID regression

$$(\hat{\tau}^{sdid}, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\tau, \mu, \alpha, \beta} \left\{ \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \hat{\omega}_i^{sdid} \hat{\lambda}_t^{sdid} \right\};$$

---

Note  $W_{it} \rightarrow D_{it}$

## Synthetic DiD

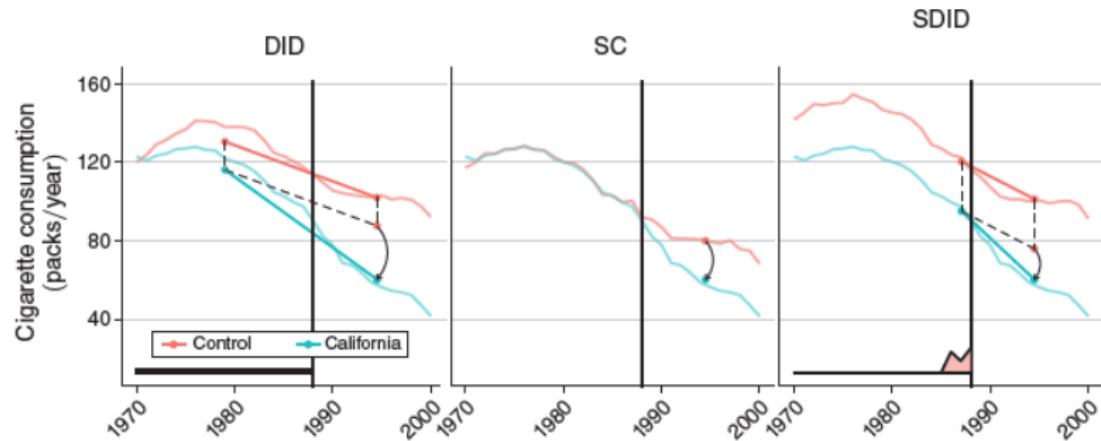
- ▶ Example: California Smoking Cessation Program (Abadie, Diamond and Hainmueller (2010).)
- ▶ the effect of increased cigarette taxes on smoking in California.
- ▶ 9 states (including California) from 1970 through 2000.
- ▶ California passed Proposition 99 increasing cigarette taxes (i.e., is treated) from 1989 onwards.

TABLE 1

	SDID	SC	DID	MC	DIFP
Estimate	-15.6	-19.6	-27.3	-20.2	-11.1
Standard error	(8.4)	(9.9)	(17.7)	(11.5)	(9.5)

*Notes:* Estimates for average effect of increased cigarette taxes on California per capita cigarette sales over 12 posttreatment years, based on SDID, SC, DID, MC, DIFP, along with estimated standard errors. We use the “placebo method” standard error estimator discussed in Section IV.

## Synthetic DiD



- ▶ DID relies on the assumption that cigarette sales in different states would have evolved in a parallel way absent the intervention. Here, preintervention trends are obviously not parallel, so the DID estimate should be considered suspect.

## Synthetic DiD

- ▶ SCM reweights the unexposed states so that the weighted of outcomes for these states match California preintervention as close as possible, and then attributes any postintervention divergence of California from this weighted average to the intervention.
- ▶ What SDID does here is reweight the unexposed control units to make their time trend parallel (but not necessarily identical) to California preintervention, then apply a DID analysis to this reweighted panel.
- ▶ Because of the time weights, we only focus on a subset of the preintervention time periods when carrying out this last step.
- ▶ These time periods were selected so that the weighted average of historical outcomes predicts average treatment period outcomes for control units, up to a constant.

# Difference-in-Discontinuities

- ▶ Difference-in-Discontinuity
  - ▶ RDD+DiD
  - ▶ DiDC, RD-DiD, DiRD, ...
  - ▶ Grembi et al. (2016). Do Fiscal Rules Matter?, *American Economic Journal: Applied Economics*.
  - ▶ Fiscal Rule: impose constraints on fiscal policy by law.
  - ▶ In Italy, fiscal rules are imposed by national government on local governments.
- Q Do fiscal rules effectively reduce debt?

- ▶ Italy's Domestic Stability Pact (DSP): the central government set a target on deficit reduction for all municipal governments in 1999.
  - ▶ relaxed it for municipalities below 5,000 inhabitants in 2001.
  - ▶ Regression Discontinuity?
- No Confounding policy: the salary of the mayor changes sharply at the threshold of 5,000 inhabitants.
- ▶ DiD?
- No Large and small municipalities are typically on differential trends in public policies.

- ▶ Two sources of variation: before/after 2001 and just below/above 5,000 inhabitants.
- ▶ Difference-in-Discontinuities: take the difference between the pretreatment and the posttreatment discontinuity at the cutoff in order to difference out the effect of the salary of the mayor.

# Institutional Background

TABLE 1—RULES OF THE DOMESTIC STABILITY PACT (DSP)

Year	Target of the DSP rules	Covered municipalities
1997	None	All
1998	None	All
1999	Fiscal gap: zero growth	All
2000	Fiscal gap: zero growth	All
2001	Fiscal gap: max 3 percent growth	Above 5,000
2002	Fiscal gap: max 2.5 percent growth	Above 5,000
2003	Fiscal gap: zero growth	Above 5,000
2004	Fiscal gap: zero growth	Above 5,000

*Notes:* The Domestic Stability Pact is a set of fiscal rules imposed by the central government to discipline the fiscal management of local governments. The main target is the *Fiscal gap* (see online Appendix Table A1 for details). The growth of the fiscal gap with respect to its value two years before is constrained to be either 0 or below 2.5 percent/3 percent depending on the year of the DSP.

*Source:* Annual national budget law (*Legge Finanziaria*) from 1999 to 2004

- ▶ Define two treatment variables

$$W_{it} = \begin{cases} 1 & \text{if wage is low} \\ 0 & \text{otherwise} \end{cases}$$

and

$$R_{it} = \begin{cases} 1 & \text{if fiscal rules are relaxed} \\ 0 & \text{otherwise} \end{cases}$$

- ▶ Write  $P_{it}$  = population,  $P_c = 5,000$ .
- ▶ Note that

$$W_{it} = 1 \iff P_{it} < P_c$$

and

$$R_{it} = 1 \iff P_{it} < P_c \quad \& t \geq t_0$$

where  $t_0 = 2001$ .

- ▶ The outcome variable  $Y_{it}$  has four potential outcomes:

$$Y_{it}(W_{it} = w, R_{it} = r) = Y_{it}(w, r).$$

- ▶ We can write

$$\begin{aligned} Y_{it} &= W_{it}[R_{it} Y_{it}(1, 1) + (1 - R_{it}) Y_{it}(1, 0)] \\ &\quad + (1 - W_{it})[R_{it} Y_{it}(0, 1) + (1 - R_{it}) Y_{it}(0, 0)] \end{aligned}$$

- ▶ Our aim is to identify the causal effect of  $R_{it}$  on  $Y_{it}$ .

**Claim** The canonical RD estimator fails identification.

**pf** First observe that

$$\lim_{p \searrow P_c} \mathbb{E}[Y_{it}(1, 0) | P_{it}, t \geq t_0] = \lim_{p \nearrow P_c} \mathbb{E}[Y_{it}(1, 0) | P_{it}, t \geq t_0]$$

assuming continuity in  $w$ . From this, we see that

$$\begin{aligned}\delta_{RD}^{POST} &= \lim_{p \searrow P_c} \mathbb{E}[Y_{it} | P_{it}, t \geq t_0] - \lim_{p \nearrow P_c} \mathbb{E}[Y_{it} | P_{it}, t \geq t_0] \\ &= \lim_{p \searrow P_c} \mathbb{E}[Y_{it}(0, 0) | P_{it}, t \geq t_0] - \lim_{p \nearrow P_c} \mathbb{E}[Y_{it}(1, 1) | P_{it}, t \geq t_0] \\ &= \lim_{p \searrow P_c} \mathbb{E}[Y_{it}(0, 0) | P_{it}, t \geq t_0] - \lim_{p \searrow P_c} \mathbb{E}[Y_{it}(1, 0) | P_{it}, t \geq t_0] \\ &\quad - [\lim_{p \nearrow P_c} \mathbb{E}[Y_{it}(1, 1) | P_{it}, t \geq t_0] - \lim_{p \nearrow P_c} \mathbb{E}[Y_{it}(1, 0) | P_{it}, t \geq t_0]] \\ &= \mathbb{E}[Y_{it}(0, 0) - Y_{it}(1, 0) | P_{it} = P_c, t \geq t_0] \\ &\quad - \mathbb{E}[Y_{it}(1, 1) - Y_{it}(1, 0) | P_{it} = P_c, t \geq t_0]\end{aligned}$$

- ▶ The second term on the last equality is the causal parameter of interest. (the average treatment effect of relaxing fiscal rules in cities where mayors are poorly paid)
- ▶ the first term captures a “selection bias” (ATE of making mayor poorer where fiscal rules are binding)
- ▶ The canonical RD estimator is biased.

- ▶ For the pretreatment period, similarly:

$$\delta_{RD}^{PRE} = \mathbb{E}[Y_{it}(0, 0) - Y_{it}(1, 0) | P_{it} = P_c, t < t_0]$$

- ▶ This corresponds to the selection bias term at  $t < t_0$ !
- ▶ Hence it is natural to think of an estimator

$$\delta_{DiDC} = \delta_{RD}^{POST} - \delta_{RD}^{PRE}$$

► Identifying assumptions:

- A1 All potential outcomes are continuous in  $p$  at  $P_C$ .
- A2 The effect of the confounding policy  $W_{it}$  at  $P_c$ , in the case of no treatment, ( $R_{it} = 0$ ) is constant over time.<sup>1</sup> i.e.

$$\mathbb{E}[Y_{it}(1, 0) - Y_{it}(0, 0) | P_{it} = P_c, t \geq t_0]$$

$$= \mathbb{E}[Y_{it}(1, 0) - Y_{it}(0, 0) | P_{it} = P_c, t < t_0]$$

**Thm** Under assumptions A1 and A2, the DiRD estimator identifies the (local) ATE  $\mathbb{E}[Y_{it}(1, 1) - Y_{it}(1, 0) | P_{it} = P_c]$ .

---

<sup>1</sup>One should notice that this corresponds to the parallel trends assumption of DiD.

- ▶ Moreover if we assume further that:  
The effect of the treatment  $R_{it}$  at  $P_c$  does not depend on the confounding policy  $W_{it}$ , i.e.  
$$Y(1, 1) - Y(1, 0) = Y(0, 1) - Y(0, 0).$$
- ▶ Then the DiRD estimator identifies the (local) causal effect of relaxing fiscal rules in a neighborhood of the threshold.
- ▶ The DiDC is estimated by the following regression:

$$\begin{aligned} Y_{it} = & \beta_0 + \beta_1 P_{it}^* + S_i(\gamma_0 + \gamma_1 P_{it}^*) \\ & + T_t[\alpha_0 + \alpha_1 P_{it}^* + S_i(\mu_0 + \mu_1 P_{it}^*)] + u_{it} \end{aligned}$$

where  $P_{it}^* = P_{it} - P_c$ ,  $S_i$  is an indicator for cities below population 5,000 and  $T_t$  an indicator for posttreatment period.

	Deficit	Fiscal gap
<i>Panel A: Fiscal discipline and expenditures</i>		
Calonico et al. (2014)	17.495 (7.737)	59.468 (32.079)
<i>h</i>	600	513
Observations	2,414	2,136
Cross validation	9.454 (4.343)	48.469 (23.315)
<i>h</i>	1,498	833
Observations	5,858	3,438
<i>Mean</i>	13.393	190.757

# Thank you.

## Useful Materials

- ▶ Jonathan Roth, DiD Resources  
<https://www.jonathandroth.com/did-resources/>.
- ▶ Goldsmith-Pinkham's Applied Empirical Methods class  
<https://github.com/paulgp/applied-methods-phd>
- ▶ Scott Cunningham's *Causal Inference: the Mixtape*  
<https://mixtape.scunning.com/>
- ▶ Prof. Chamna Yoon's Applied Microeconomics lecture  
<https://www.youtube.com/@chamnayoon>