UChicago Econometrics Notes: 20510

Notes by Agustín Esteva, Lectures by Murilo Ramos

Academic Year 2024-2025

Contents

1 Lectures

1.1 Monday, June 16: Intro to Probability

Lemma 1. (Jensen) Suppose X is a random variable. Let $g : \mathbb{R} \to \mathbb{R}$ be convex. Then

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$$

Proof. Since g is convex, then for any $x, y \in \mathbb{R}$, and for any $t \in (0, 1)$

$$g(tx + (1-t)y) \leq tg(x) + (1-t)g(y).$$

Let $x = X$ and $y = \mathbb{E}[X]$, we see that

$$g(tX + (1-t)\mathbb{E}[X]) \leq tg(X) + (1-t)g(\mathbb{E}[X]).$$

Taking expected value,

$$\mathbb{E}[g(tX + (1-t)\mathbb{E}[X]] \leq t\mathbb{E}[g(X)] + (1-t)g(\mathbb{E}[X])$$

□

Definition 1. Let X be a random variable. We say that the **k*th* moment** of X is $\mathbb{E}[X^k]$. We say that the **k*th* centered moment** of X is $\mathbb{E}[(X - \mathbb{E}[X])^k]$. We say that the **k*th* standardized moment** of X is

$$\mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}}\right)^k\right]$$

Definition 2. We say that the **skewness** of X is the third standardized moment of X. We say that the **kurtosis** of X is the fourth standardized moment of X.

Lemma 1. Let $s \leq t$. Let X be a positive random variable. If $\mathbb{E}[X^t] < \infty$, then $\mathbb{E}[X^s] < \infty$.

Proof. We can split up X into

$$X^t = \mathbb{1}_{\{X^t \geq 1\}} + \mathbb{1}_{\{X^t < 1\}}.$$

It should now be clear that $X^s < X^t + 1$ almost surely. Taking expectations we are done. □

Definition 3. Let X and Y be random variables. We define the **covariance** of X and Y to be

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Proposition 1. Let X, Y, Z be random variables. Let $a, b, c \in \mathbb{R}$.

(a)
$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$$

(b)
$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$
$$\mathrm{Cov}(X, a) = 0$$

(c) Covariance is bilinear in terms of random variables.

(d)

$$\text{Cov}(a + bX, cY) = bc\text{Cov}(X, Y)$$

Lemma 2.

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

Definition 4. We define the **correlation** of X and Y to be

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

Remark 1. If $\text{Corr}(X, Y) = 0$, we say that X and Y are uncorrelated, and infer that there is no linear association between X and Y.

Lemma 3. (C-S Lemma)

$$(x, y) \leq \|x\|\|y\|$$

Remark 2. Using the L^2 norm, we see that

$$\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}$$

Proof. We prove it for the case of expected value. Let $a \in \mathbb{R}$. Then

$$0 \leq \mathbb{E}[(X - aY)^2] = \mathbb{E}[X^2] - 2a\mathbb{E}[XY] + a^2\mathbb{E}[Y^2]$$

is a quadratic function with respect to a. Optimizing,

$$0 = -2\mathbb{E}[XY] + 2a^*\mathbb{E}[Y^2] = 0 \implies a^* = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$$

Plugging back into (1),

$$0 \leq \mathbb{E}[X^2] - 2\frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} + \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} = \mathbb{E}[X^2] - \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]}$$

□

Remark 3. We see that if $X = aY$, then by (1), equality in C-S happens. This is an iff.

Theorem 1. For any X, Y random variables,

$$|\text{Corr}(X, Y)| \leq 1$$

with equality if and only if $Y = a + bX$ almost surely for some $a, b \in \mathbb{R}$.

Proof. It suffices to show that $0 \leq (\text{Corr}(X, Y))^2 \leq 1$. But

$$(\text{Corr}(X, Y))^2 = \frac{\text{Cov}(X, Y)^2}{\text{Var}(X)\text{Var}(Y)}$$

By definition, it suffices to show that

$$\text{Cov}(X, Y)^2 = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]^2 \leq \mathbb{E}[(X - \mathbb{E}[X])^2]\mathbb{E}[(Y - \mathbb{E}[Y])^2] = \text{Var}(X)\text{Var}(Y)$$

and so we are done by C-S inequality. Equality comes from equality in C-S. □

Definition 5. Recall that the **conditional probability** of X given Y is defined to be

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_X(x)}$$

The **conditional expectaion** of X given Y is defined to be

$$\mathbb{E}[X \mid Y = y] = \int_{\mathbb{R}} x f_{X|Y}(x \mid y) = \frac{\int_{\mathbb{R}} x f_{XY}(x,y)}{\int_{\mathbb{R}} f_Y(y)}$$

Definition 6. Recall that the **mean squared error** of $\hat{X}$ is

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

Theorem 2. $\mathbb{E}[Y \mid X]$ is the best predictor for Y given X in an MSE sense. That is, it is the best estimator in the sense that it minimizes the MSE. In other words,

$$\mathbb{E}[Y \mid X] = \min_{g(X)} \mathbb{E}[(Y - g(X))^2]$$

Theorem 3.

$\mathbb{E}[Y \mid X]$ is the best predictor for Y given X in an MSE sense: That is,

$$\mathbb{E}[Y \mid X] = \min_{g(X)} \mathbb{E}[(Y - g(X))^2]$$

Proposition 2. Let X, Y, Z be random variables, let g, f be functions, and let $a, b \in \mathbb{R}$ Then the following hold:

(a) $\mathbb{E}[g(X) + h(X)Y \mid X] = g(X) + g(X)\mathbb{E}[Y \mid X]$

(b) $\mathbb{E}[aY = bZ + c \mid X] = a\mathbb{E}[Y \mid X] + b\mathbb{E}[Z \mid X] + c$

(c) If $Y \leq Z$ almost surely, then $\mathbb{E}[Y \mid X] \leq \mathbb{E}[Z \mid X]$

(d) $\boxed{\text{(Tower Law) } \mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]}$

Definition 7. We say that X is **mean independent** of Y if $\mathbb{E}[X \mid Y] = c$ almost surely.

Remark 4. Note that this notion is not symmetric.

Lemma 4. If X is mean independent of Y, then

- $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$
- $\mathbb{E}[YX] = \mathbb{E}[Y]\mathbb{E}[X]$
- $\text{Corr}(Y, X) = 0$

Proof. Easy!

- $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[c] = c = \mathbb{E}[X \mid Y]$
- $\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY \mid Y]] = \mathbb{E}[Y\mathbb{E}[X \mid Y]] = c\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y]$
- Clear from ii and the fact that $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

$\square$

Remark 5. Independence implies mean independence implies zero covariance. The converses are in general false.

- Zero covariance but not mean independent: Let X be a random variable taking values $\{-1, 0, 1\}$ with equal probability:

$$\begin{aligned} X = -1 &= 1/3 \\ X = 0 &= 1/3 \\ X = 1 &= 1/3 \end{aligned}$$

  Let $Y = X^2$.

- Let X be a random variable taking values $\{-1, 1\}$ with:

$$\begin{aligned} X = -1 &= 0.5 \\ X = 1 &= 0.5 \end{aligned}$$

  Let Y be defined such that:

  – If $X = 1$, Y takes values $\{-1, 1\}$ with $Y = -1|X = 1 = 0.5$ and $Y = 1|X = 1 = 0.5$.

  – If $X = -1$, Y takes values $\{-2, 2\}$ with $Y = -2|X = -1 = 0.5$ and $Y = 2|X = -1 = 0.5$.

Definition 8. We say that the **conditional variance** of Y given X is

$$\text{Var}(Y \mid X) = \mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2 \mid X]$$

Lemma 5. Let X and Y be r.v. and g, h be functions. Then

(a) $\text{Var}(Y \mid X) = \mathbb{E}[Y^2 \mid X] - \mathbb{E}[Y \mid X]^2$

(b) $\text{Var}(g(X) + h(X)Y \mid X) = \text{Var}(h(X)Y \mid X) = h^2(X)\text{Var}(Y \mid X)$

(c) (Law of Total Variance) $\text{Var}(Y) = \mathbb{E}[\text{Var}(Y \mid X)] + \text{Var}(\mathbb{E}[Y \mid X])$

Proof. First,

$$\begin{aligned} \mathbb{E}[\text{Var}(Y \mid X)] &= \mathbb{E}[\mathbb{E}[Y^2 \mid X] - \mathbb{E}[Y \mid X]^2] \\ &= \mathbb{E}[\mathbb{E}[Y^2 \mid X]] - \mathbb{E}[\mathbb{E}[Y \mid X]^2] \\ &= \mathbb{E}[Y^2] - \mathbb{E}[\mathbb{E}[Y \mid X]^2] \end{aligned}$$

For the second term,

$$\begin{aligned} \text{Var}(\mathbb{E}[Y \mid X]) &= \mathbb{E}[\mathbb{E}[Y \mid X]^2] - \mathbb{E}[\mathbb{E}[Y \mid X]]^2 \\ &= \mathbb{E}[\mathbb{E}[Y \mid X]^2] - \mathbb{E}[Y]^2 \end{aligned}$$

Combining we conclude. $\square$

Wednesday, June 18: Intro to Statistics

We will assume that if we are sampling without replacement with simple random samples, then for a large population, we will treat it as i.i.d. samples.

Definition 9. Recall that an **estimator** is a function of the sample such that

$$\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$$

Remark 6. Note that an estimator is a random variable, as compared to parameters (e.g, means of populations or variances of populations), which are numbers.

The sample mean is a the most frequently used estimator.

Recall the analogy principle, where we use $\frac{1}{n}\sum \cdot$ to mim $\mathbb{E}[\cdot]$. For example, if $\theta = \mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$, then

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

As another example, consider

$$\theta = \mathbb{P}\{X \le x\} = \mathbb{E}[\mathbb{1}_{X \le x}]$$

then

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i \le x}$$

Definition 10. Let $\hat{\theta}$ be an estimator for θ. We define the **bias** to be

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

Example 1.1. The sample mean is unbiased:

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}[\frac{1}{n}\sum X_i] = \frac{1}{n}\sum \mathbb{E}[X_i] = \mathbb{E}[X]$$

Example 1.2. Consider $\theta = \mathrm{Var}(X)$ with $\hat{\theta} = \frac{1}{n}\sum(X_i - \overline{X}_n)^2$. Then consider that by the previous example,

$$\begin{aligned}
\hat{\theta_n} &= \frac{1}{n}\sum(X_i - \overline{X}_n)^2 \\
&= \frac{1}{n}\sum((X_i - \mathbb{E}[X_i]) - (\overline{X}_n + \mathbb{E}[\overline{X}_n]))^2 \\
&= \frac{1}{n}\sum(X_i - \mathbb{E}[X_i])^2 - (\overline{X}_n + \mathbb{E}[\overline{X}_n])^2 \\
\mathbb{E}[\hat{\theta}_n] &= \mathbb{E}\left[\frac{1}{n}\sum(X_i - \mathbb{E}[X_i])^2 - (\overline{X}_n + \mathbb{E}[\overline{X}_n])^2\right] \\
&= \frac{1}{n}\sum \mathrm{Var}(X_i) - \frac{1}{n}\mathrm{Var}(\overline{X}_n) \\
&= \mathrm{Var}(X) - \frac{1}{n}\mathrm{Var}(X) = \frac{(n-1)}{n}\mathrm{Var}(X)
\end{aligned}$$

Normalize $\frac{n}{n-1}$ to make it unbiased.

That is a stupid ass proof. Convince yourself of the following steps:

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_n] &= \mathbb{E}[\frac{1}{n}\sum(X_i - \bar{X})^2] \\
&= \frac{1}{n}\sum \mathbb{E}[X_i^2] - \frac{1}{n}\sum \mathbb{E}[(\bar{X})^2] \\
&= \mathbb{E}[X^2] - \mathbb{E}[(\bar{X})^2] \\
&= \mathrm{Var}(X) - \mathbb{E}[X]^2 - (\mathrm{Var}(\bar{X}) - \mathbb{E}[\bar{X}]^2) \\
&= \mathrm{Var}(X) - \frac{1}{n}\mathrm{Var}(X) \\
&= \frac{n-1}{n}\mathrm{Var}(X)
\end{aligned}$$

1.2 Final, June 20: Estimator Theory

Remark 7. Recall that linear combination of normal variables is normal, and thus if $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, then

$$\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \implies \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0,1)$$

Definition 11. We say that an estimator $\hat{\theta}_n$ **consistent** if it converges in probability. That is, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\{|\hat{\theta}_n - \theta| > \epsilon\} = 0$$

and we write

$$\hat{\theta}_n \underset{\mathbb{P}}{\to} \theta$$

Lemma 6. (Chebyshev). If $1 \leq p < \infty$, then for any $\lambda > 0$, we have that

$$\mathbb{P}\{|X| \geq \lambda\} \leq \frac{\mathbb{E}[|X|^p]}{\lambda^p}$$

Proof. Letting $A = \{\omega \mid |X| \geq \lambda\}$, we see that

$$\mathbb{E}[|X|^p] = \int_\Omega |X|^p d\mathbb{P} \geq \int_A |X|^p d\mathbb{P} \geq \lambda^p \mathbb{P}\{A\}$$

□

Theorem 4.

Weak Law of Large Numbers. Let $X_1, \dots, X_n \sim F$ be i.i.d. Suppose $\mathbb{E}[X_1^2] < \infty$, then

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}[X_1].$$

Proof. Note that

$$\mathrm{Var}(\overline{X}_n) = \mathbb{E}[(\overline{X}_n - \mathbb{E}[\overline{X}_n])^2] = \mathbb{E}[(\overline{X}_n - n\mathbb{E}[X_1])^2].$$

Hence,

$$\begin{aligned}\mathbb{P}\{|\overline{X}_n - \mathbb{E}[X_1]| > \epsilon\} &= \mathbb{P}\{(\overline{X}_n - \mathbb{E}[X_1])^2 > \epsilon^2\} \\ &\leq \frac{\mathrm{Var}(\overline{X}_n)}{\epsilon^2} \\ &= \frac{\mathrm{Var}(X_1)}{n\epsilon^2} \\ &\to 0\end{aligned}$$

where we use the fact that $\mathbb{E}[X_1^2] < \infty$ to say that $\mathrm{Var}(X_1) < \infty$. □

Proposition 3. Let $X_1, \dots, X_n \sim F$ be i.i.d. Then $\overline{X}_n$ is consistent.

Proof. As $n \to \infty$, we know by the law of large numbers that $\overline{X}_n \to \mu$ in probability, and so we are done. □

Theorem 5.

Continuous Mapping Theorem.

(a) Suppose $X_n \to x$ in probability and g is continuous. Then

$$g(X_n) \underset{\mathbb{P}}{\to} g(x)$$

(b) Suppose $X_n \to X$ in distribution and g is continuous. Then

$$g(X_n) \underset{\mathscr{D}}{\to} g(X)$$

Proposition 4. Let $X_1, \ldots, X_n \sim F$ be i.i.d. Then

$$\hat{\sigma}_x^2 = \frac{1}{n} \sum (X_i - \overline{X})^2$$

is consistent.

Proof. Note that

$$\hat{\sigma}_x^2 = \frac{1}{n} \sum (X_i - \overline{X}_n)^2 = \left[\frac{1}{n} \sum X_i^2\right] - (\overline{X}_n)^2 \to \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

by the weak law of large numbers and the continuous mapping theorem using $g(w, x) = w - z^2$ and so we are done. To see the big step, we open up the parenthesis:

$$\begin{aligned}
\hat{\sigma}_x^2 &= \frac{1}{n} \sum (X_i - \overline{X}_n)^2 \\
&= \frac{1}{n} \sum X_i^2 - 2X_i\overline{X} + \overline{X}^2 \\
&= \frac{1}{n} X_i^2 - \frac{1}{n} 2\overline{X} \sum X_i + \overline{X}^2 \\
&= \frac{1}{n} X_i^2 + \overline{X}^2
\end{aligned}$$

$\square$

Example 1.3. Let $(X_1, Y_1), \cdots \sim (X, Y)$ be i.i.d with $X, Y \in L^2$. Let $\theta = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$. By the anaology principle,

$$\hat{\theta}_n = \frac{1}{n} \sum (X_i - \overline{X})(Y_i - \overline{Y})$$

Letting $g(w, z, t) = w - zt$ and noting that

$$\hat{\theta}_n = \frac{1}{n} \sum X_i Y_i - \overline{XY},$$

we can use the CMT and the WLLN to show that $\hat{\theta}_n$ is consistent.

Definition 12. We say that X_n **converges in distribution** to X if $F_{X_n} \to F_X(x)$

Theorem 6.

Central Limit Theorem. Let $X_1, \ldots, X_n \sim F$ be i.i.d. with mean μ and $\mathbb{E}[X^2] < \infty$ and variance σ^2. Then

$$\frac{S_n}{\sqrt{n}} \to N(\mu, \sigma^2)$$

in distribution

In other words, we have that for large n,

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$$

Lemma 2.

Slutsky Suppose $X_n \to X$ in distribution and $Y_n \to y$ in probability. Then

(a) $X_n Y_n \to Xy$ in distribution

(b) $X_n + Y_n \to X + y$ in distribution

(c) $\frac{X_n}{Y_n} \to \frac{X}{y}$ if $y \neq 0$

(d) If g is continuous, then $g(X_n, Y_n) \to g(X, y)$

Example 1.4. Suppose $X_1, \ldots, X_n \sim X$ i.i.d. with $\mathbb{E}[X^2] < \infty$ and $\sigma_X^2 > 0$. Recall that the CLT implies that

$$\frac{\sqrt{n}(\overline{X}_n - \mu_X)}{\sigma_X} \underset{d}{\to} N(0,1).$$

In general, we don't observe σ_X^2, so we use an estimate $\hat{\sigma}_X^2 \underset{\mathbb{P}}{\to} \sigma_X^2$ and thus by the CMT

$$\hat{\sigma}_X \to \sigma_X.$$

Hence, a more feasible statistic for hypothesis tests is

$$\frac{\sqrt{n}(\overline{X}_n - \mu_X)}{\hat{\sigma}_X} = \left(\frac{\sqrt{n}(\overline{X}_n - \mu_X)}{\sigma_X}\right)\frac{\sigma_X}{\hat{\sigma}_X} \to N(0,1)$$

by Slutsky.

Remark 8.

Hypothesis Testing

(a) (*Step 1*) State H_0 and H_a.

(b) (*Step 2*) Test statistic and call it

$$T_n = g(X_1, \ldots, X_n)$$

a function of the data.

- Z score could be

$$Z = \frac{\sqrt{n}(\overline{X}_n - \mu_X)}{\hat{\sigma}_X}$$

(c) (*Step 3*) Outline rejection region R and critical values. I.e, $\alpha = 0.05$.

(d) (*Step 4*) Conclude (Reject or fail to reject H_0)

Definition 13. We say that a **Type I Error** is when the null hypothesis is falsely rejected (H_0 is true but it is rejected). We say that a **Type II Error** is when the failed to be failed to be rejected (H_0 is false but it was failed to be rejected)

Remark 9. The convention is to choose some $\alpha \in \mathbb{R}$ such that

$$\mathbb{P}\{\text{Type I error}\} = \mathbb{P}\{T_n \in R \mid H_0\} = \alpha.$$

We call α our significance level.

Example 1.5. (Two sided) Suppose $0 < \text{Var}(X) < \infty$ and $H_0 : \mathbb{E}[X] = \mu_0$ and $H_a : \mathbb{E}[X] \neq \mu_0$. We let

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\hat{\sigma}_X} \underset{d}{\to} N(0,1)$$

where the convergence happens under the null. We set $\alpha = 0.05$, and thus

$$\mathbb{P}\{\left|\frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\hat{\sigma}_X}\right| \geq c \mid H_0\} = \alpha = 2\,(1 - \Phi(c))$$

by the symmetric of the normal distribution. Solving,

$$c = \Phi^{-1}(1 - \frac{\alpha}{2})$$

Definition 14. We define the $p-$**value** to be the smallest α for which we reject H_0.

Example 1.6. $H_0 : \mathbb{E}[X] \geq 10$, $H_a : \mathbb{E}[X] < 10$. Found $T_n = -1.5$. Then the $p-$value is

$$\mathbb{P}\{Z \leq -1.5\} = p.$$

We reject if $p < \alpha$ Suppose now $H_0 : \mathbb{E}[X] = 10$ and $H_a : \mathbb{E}[X] \neq 10$. Then

$$2\mathbb{P}\{Z \leq -1.5\} = p.$$

More generally, we saw in a $2-$sided test that $c = \Phi^{-1}(1 - \frac{\alpha}{2})$ and we reject if $|T_n| > c$, and thus reject if

$$\alpha > 2(1 - \Phi(|T_n|)) = 2\mathbb{P}\{|T_n|\} = p$$

1.3 Monday, June 23: Introducing the SLR

Definition 15. (SLR Model) We say that y is a simple linear regression if

$$Y_i = \beta_0 + \beta_1 X_i + \sigma U_i,$$

where we call β_0 to be our intercept parameter, β_1 to be our slope parameter, and U_i is the error term.

Remark 10. There are three ways to interpret the regressors, and an analysis of these interpretations will yield some insight in why we assume some things:

(a) (Linear Conditional Expectation) Suppose that for some Y and X r.v,

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

We can define

$$U = Y - \mathbb{E}[Y \mid X].$$

Hence, by definition,

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

Thus, we see that

$$\mathbb{E}[U \mid X] = \mathbb{E}[Y - \mathbb{E}[Y \mid X] \mid X] = 0,$$

implying that U is mean independent of X and thus

$$\mathrm{Cov}(U, X) = 0$$

and moreover,

$$\mathbb{E}[U] = \mathbb{E}[\mathbb{E}[U \mid X]] = 0$$

(b) (Best Linear Predictor (BLP)). Suppose $Y = \beta_0 + \beta_1 X + U = \mathrm{BLP}(Y \mid X) + U$

- Suppose we want to find the best linear predictor for Y as a function of X in the sense that in minimizes MSE. That is,

$$\mathrm{BLP}(Y \mid X) = \min_{(b_0, b_1) \in \mathbb{R}^2} \mathbb{E}[(Y - b_0 - b_1 X)^2]$$

  Taking FOC, we find that

$$\mathrm{BLP}_1 = (\beta_0, \beta_1)$$

- Suppose we want to find the best linear predictor for $\mathbb{E}[Y \mid X]$. We want to find

$$\mathrm{BLP}_2 = \min_{(b_0, b_1) \in \mathbb{R}^2} \mathbb{E}[(\mathbb{E}[Y \mid X] - b_0 - b_1 X)^2]$$

We claim that $\mathrm{BLP}_1 = \mathrm{BLP}_2$.

Proof. Computing,

$$\begin{aligned}
\mathbb{E}[(Y - b_0 - b_1 X)^2] &= \mathbb{E}[Z^2] \\
&= \mathbb{E}[((Y - \mathbb{E}[Y \mid X) + (\mathbb{E}[Y \mid X] - b_0 + b_1 X))^2] \\
&= \mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2] + \mathbb{E}[(\mathbb{E}[Y \mid X] - b_0 + b_1 X))^2]
\end{aligned}$$

where we can use orthogonality since

$$\begin{aligned}
&\mathbb{E}[(Y - \mathbb{E}[Y \mid X)(\mathbb{E}[Y \mid X] - b_0 + b_1 X)] = \\
&= \mathbb{E}[Y \mathbb{E}[Y \mid X]] - \mathbb{E}[\mathbb{E}[Y \mid X]^2] - b_0 \mathbb{E}[V] - b_1 \mathbb{E}[VX] \\
&= 0 - 0 - b_1 \mathbb{E}[\mathbb{E}[VX \mid X]] = 0
\end{aligned}$$

Hence, we minimize by taking derivatives and the first term drops out, yielding our result. ☐

So we minimized $\mathbb{E}[(Y - b_0 - b_1 X)^2] = \mathbb{E}[Z^2]$ and to do this explicitly,

$$\frac{\partial f}{\partial b_0} = -2\mathbb{E}[Y - b_0 - b_1 X] \implies \mathbb{E}[U] = 0 \qquad \frac{\partial f}{\partial b_1} = -2\mathbb{E}[X(Y - b_0 - b_1 X)] \implies \mathbb{E}[UX] = 0$$

Thus,

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

and

$$\text{Cov}(U, X) = \mathbb{E}[UX] - \mathbb{E}[U]\mathbb{E}[X] = 0$$

and thus the BLP satisfies the conditions in the previous example.

(c) (Causal Interpretation) Suppose our BLP is of the form

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

where *we assume* that $\mathbb{E}[U] = 0$ and $\text{Cov}(X, U) = \mathbb{E}[XU] = 0$. Then the causal model is of the form

$$Y = \gamma_0 + \gamma_1 X + V,$$

where V is called the causal error (alive!) and can be explained by everything that causes Y which is not encoded in X, implying that $\text{Cov}(X, V) \neq 0$. We define γ_1 to be

$$\left.\frac{\partial Y}{\partial X}\right|_{\text{keeping everything constant}} = \gamma_1.$$

It is easy to estimate β_0, β_1, but it is much harder to compute γ_0, γ_1.

1.4 Wednesday, June 25: SLR Coefficient Theory

Lemma 3. The following equalities hold:

(a)
$$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum(X_i - \bar{X})Y_i$$

(b)
$$\sum(X_i - \bar{X})^2 = \sum(X_i - \bar{X})X_i$$

(c)
$$\sum(X_i - \bar{X}) = 0$$

Remark 11.

SLR Setup in the population Let X, Y, U be r.v. such that

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

and assume

(a) $\mathbb{E}[U] = 0$

(b) $\mathbb{E}[XU] = \text{Cov}(X, U) = 0$

(c) $0 < \text{Var}(X) < \infty$

(d) $(X_1, Y_1), \ldots, (X_n, Y_n) \sim (X, Y)$ i.i.d.

From (a), we have that

$$\begin{aligned} 0 &= \mathbb{E}[U] \\ &= \mathbb{E}[Y - \beta_0 - \beta_1 X] \\ &= \mathbb{E}[Y] - \beta_0 - \beta_1 \mathbb{E}[X] \end{aligned}$$

From (b), we have that

$$\begin{aligned} 0 &= \mathbb{E}[X(Y - \beta_0 - \beta_1 X)] \\ &= \mathbb{E}[X(Y - \mathbb{E}[Y]) - \beta_1(X - \mathbb{E}[X])] \end{aligned}$$

and hence

$$\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}(Y))] = \mathbb{E}[X(Y - \mathbb{E}[Y])] = \beta_1 \mathbb{E}[X(X - \mathbb{E}[X])] = \beta_1 \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}(X))]$$

Thus,

$$\boxed{\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}} \tag{1}$$

$$\boxed{\beta_0 = \mathbb{E}[Y] - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \mathbb{E}[X]} \tag{2}$$

Example 1.7. Consider the special case when X is Bernoulli so that $X_1 \sim \text{Bernoulli}(p)$. Note that to compute $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, we compute

$$\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y \mid X]] = p\mathbb{E}[Y \mid X = 1]$$

$$\mathbb{E}[X]\mathbb{E}[Y] = p\mathbb{E}[Y] = p\mathbb{E}[\mathbb{E}[Y \mid X]] = p\left(p\mathbb{E}[Y \mid X = 1] + (1-p)\mathbb{E}[Y \mid X = 0]\right)$$

Thus, we have that

$$\begin{aligned}\text{Cov}(X, Y) &= p\mathbb{E}[Y \mid X = 1] - p^2\mathbb{E}[Y \mid X = 1] - p(1-p)\mathbb{E}[Y \mid X = 0] \\ &= p(1-p)\left(\mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0]\right)\end{aligned}$$

Hence,

$$\beta_1 = \mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0]$$

Computing, we see that

$$\begin{aligned}\beta_0 &= \mathbb{E}[Y] - \beta_1\mathbb{E}[X] \\ &= \mathbb{E}[\mathbb{E}[Y \mid X]] - \beta_1 p \\ &= p(\mathbb{E}[Y \mid X = 1]) + (1-p)\mathbb{E}[Y \mid X = 0] - \left(\mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0]\right)p \\ &= \mathbb{E}[Y \mid X = 0]\end{aligned}$$

Tautological, we have that

$$\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid X = 0] + \left(\mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0]\right)$$

implying that by definition,

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

$\boxed{\text{Thus, if } X \text{ is Bernoulli, thenE[Y} \mid X] \text{ is linear in } X \text{ and thus mean independent}}$.

Proposition 5.

SLR Setup in the Sample Let X, Y, U be r.v. such that

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

and assume

(a) $\mathbb{E}[U] = 0$

(b) $\mathbb{E}[XU] = \text{Cov}(X, U) = 0$

(c) $0 < \text{Var}(X) < \infty$

(d) $(X_1, Y_1), \ldots, (X_n, Y_n) \sim (X, Y)$ i.i.d.

Then

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \tag{3}$$

$$\hat{\beta}_2 = \bar{Y} - \hat{\beta}_1\bar{X} \tag{4}$$

Proof. This is clear using the analogy principle on (1) and (2).

For another derivation, recall that

$$\mathbb{E}[Y - \beta_0 - \beta_1 X] = 0 \qquad \mathbb{E}[(Y - \beta_0 - \beta_1 X)X] = 0$$

are the first order conditions for $\min_{(b_0,b_1)} \mathbb{E}[(Y - b_0 - b_1 X)^2]$. Within the sample, See full derivation in PSET $\square$

Definition 16. Consider a sample regression model such that

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{U}_i.$$

We call $\hat{U}_i$ the **residual**, and note that $\hat{\beta}_i$ are both random variables. We define the **residual** to be

$$\hat{U}_i = Y_i - \hat{Y}_i,$$

where $\hat{Y}$ is the **fitted value** such that

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Remark 12. Recall conditions (a) and (b) in the basic setup. We showed in the above proof the sample equivalents of them for the first order conditions. That is,

$$\boxed{\frac{1}{n}\sum \hat{U}_i = 0} \tag{5}$$

$$\boxed{\frac{1}{n}\sum X_i \hat{U}_i = 0} \tag{6}$$

Notice that these hold always in the OLS, since they are major assumptions. These should not hold in general in causal models.

Example 1.8. Suppose $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{U}_i$ where X_i is Bernoulli. Calling n_0 the number of times X_i fails and n_1 the number of successes, then $n = n_0 + n_1$ is the number in the sample. We call

$$\bar{Y}_0 = \frac{\sum_{i=1}^n Y_i(1 - X_i)}{\sum_{i=1}^n (1 - X_i)} = \frac{1}{n_0} \sum_{i:X_i=0} Y_i$$

$$\bar{Y}_1 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i} = \frac{\sum_{i:X_i=1} Y_i}{n_1}$$

Thus, we find that (see PSET)

$$\hat{\beta}_0 = \bar{Y}_0 \qquad \hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$$

Definition 17. We say that R^2 is the **measure of fit** if it is

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

where

$$\text{Total Sum of Squares (TSS)} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\text{Explained Sum of Squares (ESS)} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$\text{Sum of Squared Residuals (SSR)} = \sum_{i=1}^n \hat{u}_i^2$$

Proposition 6. The following hold,

(a)
$$\text{TSS} = \text{ESS} + \text{SSR}$$

(b)
$$R^2 = 1 - \frac{\text{SST}}{\text{TSS}}$$

(c) $R^2 \in [0, 1]$.

Proof. (a) We compute from the RHS,

$$\begin{aligned}
\sum(\hat{y}_i - \bar{y})^2 + \sum \hat{u}_1^2 &= \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2 \\
&= \sum(\hat{y}_i - \bar{y} - \hat{y}_i + y_i)^2 - 2\sum(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \\
&= \sum(y_i - \bar{y}_i)^2 - 2\sum(\hat{y}_i - \bar{y})\hat{u}_i \\
&= \sum(y_i - \bar{y}_i)^2 - 2(\sum \hat{y}_i u_i - \bar{y}\sum \hat{u}_i) \\
&= \sum(y_i - \bar{y}_i)^2 - 2(\sum(\beta_0 + \beta_1 x_i)u_i - \bar{y}\sum \hat{u}_i) \\
&= \sum(y_i - \bar{y}_i)^2 \\
&= \text{TSS}
\end{aligned}$$

Where we use Remark 12

(b) Dividing by TSS in (a), we see that
$$1 = R^2 + \frac{\text{SSR}}{\text{TSS}}$$

(c) From (b), it suffices to see that TSS $\geq$ SSR, but this follows directly from (a) □

Remark 13. Suppose $R^2 = 0$, then ESS $= 0$ and SSR $=$ TSS. That is, $\hat{Y}_i = \bar{Y}$. Terrible model!!

Suppose $R^2 = 1$, then ESS $=$ TSS. and SSR $= 0$ and thus $\hat{u}_i = 0$ and $\hat{y}_i = y_i$. Goated model.

<u>R^2 does NOT IMPLY CAUSATION.</u>

Proposition 7.

(Properties of $\hat{\beta}$) Let X, Y, U be r.v. such that

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

and assume

(a) $\mathbb{E}[U] = 0$

(b) $\mathbb{E}[XU] = \mathrm{Cov}(X, U) = 0$

(c) $0 < \mathrm{Var}(X) < \infty$

(d) $(X_1, Y_1), \ldots, (X_n, Y_n) \sim (X, Y)$ i.i.d.

Then the following hold

(a) If $\mathbb{E}[U \mid X] = 0$ (alternatively, we have shown that this condition is equivalent to X being binary or to $\mathbb{E}[Y \mid X]$ being linear in X), then

$$\mathbb{E}[\hat{\beta}_0] = \beta_0 \qquad \mathbb{E}[\hat{\beta}_1] = \beta_1$$

(b) If $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$, then

$$\hat{\beta}_0 \underset{\mathbb{P}}{\longrightarrow} \beta_0 \qquad \hat{\beta}_1 \underset{\mathbb{P}}{\longrightarrow} \beta_1$$

(c) If $\mathbb{E}[X^4] < \infty$ and $\mathbb{E}[Y^4] < \infty$, then

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \underset{\mathscr{D}}{\longrightarrow} N(0, \sigma_1^2)$$

Proof. (a) We will first show all these results for $\hat{\beta}_1$. Note that

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X} \\
&= \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} \\
&= \frac{\sum (X_i - \bar{X})(\beta_0 + \beta_1 X_i + U_i)}{\sum (X_i - \bar{X})^2} \\
&= \beta_1 + \frac{\sum (X_i - \bar{X}) U_i}{\hat{\sigma}_X^2}
\end{aligned}$$

We note that

$$\hat{\beta}_1 = \beta_1 + \frac{\sum (X_i - \bar{X}) U_i}{\sum (X_i - \bar{X})^2} \tag{7}$$

Taking $\mathbb{E}[\hat{\beta}_1 \mid X_1, \ldots, X_n]$ in (7) and using the assumption that $\mathbb{E}[U \mid X] = 0$ and then LIE we conclude. Moreover,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_0] &= \mathbb{E}[\bar{Y} - \hat{\beta}_1 \bar{X}] \\
&= \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] \\
&= \beta_0
\end{aligned}$$

(b) Under the condition of the second moments, we have showed (Proposition 4 and Example 1.3) that the estimators for covariance and variance are consistent. Thus, using (d) and the CMT for $g(s,t) = \frac{s}{t}$, we see that

$$\hat{\beta}_1 = g(\hat{\sigma}_{XY}, \hat{\sigma}_X) \underset{\mathbb{P}}{\to} g(\sigma_{XY}, \sigma_X) = \beta_1$$

Moreover, we use the CMT again with $g(w,s,t) = w - st$ to show that

$$\hat{\beta}_0 = g(\bar{Y}, \hat{\beta}_1, \bar{X}) \underset{\mathbb{P}}{\to} g(\mathbb{E}[Y], \beta_1, \mathbb{E}[X]) = \beta_0$$

(c) From (7), we see that

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_1 - \beta_1) &= \frac{\frac{1}{\sqrt{n}}\sum(X_i - \bar{X})U_i}{\frac{1}{n}\sum(X_i - \bar{X})^2} \\
&\underset{\mathbb{P}}{\to} \frac{1}{\sigma_X^2}\left[\frac{1}{\sqrt{n}}\sum(X_i - \bar{X})U_i\right] \\
&= \frac{1}{\sigma_X^2}\left[\frac{1}{\sqrt{n}}\sum(X_i - \mathbb{E}[X] + \mathbb{E}[X] - \bar{X})U_i\right] \\
&= \frac{1}{\sigma_X^2}\left[\left(\frac{1}{\sqrt{n}}\sum(X_i - \mathbb{E}[X])U_i\right) + \frac{1}{\sqrt{n}}\sum(\mathbb{E}[X] - \bar{X})U_i\right] \\
&\underset{\mathscr{D}}{\to} \frac{1}{\sigma_X^2} N(0, \mathrm{Var}((X - \mathbb{E}[X])U)) \\
&= N(0, \frac{1}{(\sigma_X^2)^2}\mathrm{Var}((X - \mathbb{E}[X])U))
\end{aligned}$$

where we use Slutsky's Lemma for the last convergence, noting that we use the CLT for the first term and the convergence of $\bar{X} \to \mu_X$ in probability for the second.

□

1.5 Friday, June 27: OVB, Homo/heteroskedasticity, and Inference

Example 1.9. (Omitted Variable Bias) Causal Model:

$$\text{wages}_i = \gamma_0 + \gamma_1 \text{educ}_i + V_i$$

where V_i is alive and $\text{Cov}(V_i, X_i) \neq 0$.

BLP Model:

$$\text{wages}_i = \beta_0 + \beta_1 \text{educ}_i + U_i$$

such that $\text{Cov}(X_i, U_i) = 0$. Thus, $\gamma_1 \neq \beta_1$ and $\gamma_0 \neq \beta_0$.

Does β_1 over/underestimate γ_1? Compare to (7), and we see that

$$\begin{aligned}
\hat{\beta}_1 &\underset{\mathbb{P}}{\to} \beta_1 \\
&= \frac{\sigma_{XY}}{\sigma_X^2} \\
&= \frac{\text{Cov}(X, \gamma_0 + \gamma_1 X + V)}{\text{Var}(x)} \\
&= \frac{\text{Cov}(X, \gamma_0) + \gamma_1 \text{Cov}(X, X) + \text{Cov}(X, V)}{\text{Var}(X)} \\
&= \gamma_1 + \frac{\text{Cov}(X, V)}{\text{Var}(X)}
\end{aligned}$$

(OVB) If $\text{Cov}(X, V) > 0$, then $\hat{\beta}_1$ overestimates γ_1. If $\text{Cov}(X, V) < 0$, then it underestimates. If $\text{Cov}(X, V) = 0$, then $\hat{\beta}_1 \to \gamma_1$ in probability.

Remark 14. In samples, it is often unfeasable to know what σ_1^2 is. Thus, we often don't use (c) in proposition 7. We estimate using

$$\hat{\sigma}_1^2 = A\hat{\text{Var}}(\hat{\beta}_1) = \frac{\frac{1}{n}\sum(X_i - \bar{X})\hat{U}_i^2}{(\hat{\sigma}_X^2)^2}$$

and we know that $\hat{\sigma}_1^2 \underset{\mathbb{P}}{\to} \sigma_1^2$.

Definition 18. If U is **homoskedastic**, then $\mathbb{E}[U \mid X] = 0$ and $\text{Var}(U \mid X) = \text{Var}(U)$. If U is heteroskedastic, then $\mathbb{E}[U \mid X] = h(X)$

Proposition 8. Suppose U is homoskedastic, then $\sigma_1^2 = \frac{\text{Var}(U)}{\text{Var}(X)}$

Proof. We have that

$$\begin{aligned}
\text{Var}((X - \mathbb{E}[X])U) &= \mathbb{E}[(X - \mathbb{E}[X])^2 U^2] \\
&= \mathbb{E}[(X - \mathbb{E}[X])^2 \mathbb{E}[U^2 \mid X]] \\
&= \mathbb{E}[(X - \mathbb{E}[X])^2 \left(\mathbb{E}[U^2 \mid X] - \mathbb{E}[U \mid X]^2\right)] \\
&= \mathbb{E}[(X - \mathbb{E}[X])^2 \text{Var}(U \mid X)] \\
&= \mathbb{E}[(X - \mathbb{E}[X])^2 \text{Var}(U)] \\
&= \text{Var}(U)\text{Var}(X)
\end{aligned}$$

□

Example 1.10. (Hetero or Homo?) Suppose Y is Bernoulli(p) and $\mathbb{E}[U \mid X] = 0$ and $Y = \beta_0 + \beta_1 X + U$. Recall that we have showed that $\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$. First, Note that $Y^2 = Y$. Next, note that

$$\text{Var}(Y \mid X) = \mathbb{E}[Y^2 \mid X] - \mathbb{E}[Y \mid X]^2 = \mathbb{E}[Y \mid X] - \mathbb{E}[Y \mid X]^2 = \mathbb{E}[Y \mid X][1 - \mathbb{E}[Y \mid X]]$$

Thus,

$$\text{Var}(Y \mid X) = (\beta_0 + \beta_1 X)(1 - \beta_0 - \beta_1 X)$$

But we also have that

$$\text{Var}(Y \mid X) = \text{Var}(\beta_0 + \beta_1 X + U \mid X) = \text{Var}(U \mid X)$$

which depends on X, and so the error term U is never homoskedastic.

Remark 15. (Hypothesis Testing)

(a) $H_0 : \beta_1 = a$ and $H_1 : \beta_1 \neq a$

(b) $T_n = \frac{\hat{\beta}_1 - \beta_1^{H_0}}{\text{SE}(\hat{\beta}_1)}$

(c) Same as before

Example 1.11. (Hypothesis Test) Test whether $\beta_1 = 1$ or $\beta_1 \neq 1$ at $\alpha 0.05$ We know that

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_1} = \frac{0.6350 - 1}{0.0214} = -17.05$$

If sample is larger, then $t_\alpha = 1.96$ and we definitely reject.

Remark 16. (Log Level Regression) Recall the Maclaurin expansions:

$$e^x = 1 + x + \frac{1}{2}x^2 + \cdots \approx 1 + x, \qquad x << 1$$

$$\log(1 + x) = 0 + x + O(x) \approx x, \qquad x << 1.$$

Thus, if $Y = \exp\{\beta_0 + \beta_1 X + U\}$, then $\text{Log}(Y) \approx \beta_0 + \beta_1 X + U$ We know that

$$\beta_1 = \frac{d \log Y}{dX} = \frac{1}{Y}\frac{dY}{dX} \approx \frac{\Delta Y}{\Delta X}\frac{1}{Y}$$

And hence

$$\frac{\Delta Y}{Y} \approx \beta_1 \Delta X.$$

Thus,

$$\boxed{\% \Delta Y \approx 100 \beta_1 \Delta X}$$

Remark 17. (Log Log Model) Suppose $\log(Y) = \beta_0 + \beta_1 \text{Log}(X) + U$ and thus

$$\beta_1 = \frac{d\text{Log}Y}{d\text{Log}X} = \frac{1}{Y}\frac{1}{\frac{1}{X}}\frac{dY}{dX} \approx \frac{\% \Delta Y}{\% \Delta X}$$

Thus,

$$\boxed{\% \Delta Y \approx \beta_1 \% \Delta X}$$

Remark 18. (Level Log Model) Similarly to before, if $Y = \beta_0 + \beta_1 \log X + U$, then

$$\boxed{\Delta Y \approx \frac{\beta_1}{100} \% \Delta X}$$

1.6 Monday, June 30: Vector Statistics

Remark 19. Recall that A^{-1} exists if $\det(A) \neq 0$ or if the columns of A are linearly independent or the rows are. Recall that a vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ is linearly dependent if there exists scalars $\mathbf{c} = (c_1, \ldots, c_n)$ such that

$$\mathbf{c}\mathbf{x} = c_1 x_1 + \ldots c_n x_n = 0.$$

Suppose X is a random vector such that

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$$

Then $\mathbb{E}[X]$ is the expected value of each of its entries. We have that

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$$

and the covariance matrix is

$$\begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots \\ \vdots & & \end{bmatrix}$$

and thus $\mathrm{Var}(X)$ is symmetric. As an example, suppose $X_{n\times 1}$ is a r.v, and $A_{m\times n}$ is a matrix of constants and $b_{m\times 1}$ is a column, then $\mathrm{Var}(AX + b) = A\mathrm{Var}(X)A^T$

Theorem 7.

(The big 4) Suppose $X_1, \ldots, X_n \sim X_{k\times 1}$ are i.i.d. Then the following hold,

(a) (**WLLN**) We have

$$\bar{X} \underset{\mathbb{P}}{\to} \mathbb{E}[X]$$

(b) (**CMT**) Suppose $X_n \underset{\mathbb{P}}{\to} x$ and $Y_n \underset{\mathbb{P}}{\to} y$, and g is continuous then

$$g(X_n, Y_n) \underset{\mathbb{P}}{\to} g(x, y)$$

(c) (**CLT**) Suppose the second moment of each element in X is finite. Then

$$\sqrt{n}(\bar{X} - \mathbb{E}[X]) \sim N(0, \mathrm{Var}(X))$$

(d) (**Slutsky's**) If $X_n \xrightarrow{d} X$ where X is a random matrix and $Y_n \xrightarrow{\mathbb{P}} y$ is a constant matrix. Then

(i) $X_n Y_N \xrightarrow{d} Xy$ when Xy is defined.

(ii) $X_n + Y_n \xrightarrow{d} X + y$ when $X + y$ is defined.

(iii) $X_n Y_n^{-1} \xrightarrow{d} Xy^{-1}$ when Xy is defined and $\det(y) \neq 0$.

Remark 20. If $X_{m\times 1} \sim \mathscr{N}(\mathbb{E}[X]_{m\times 1}, \mathrm{Var}(X)_{m\times m})$ then $AX + b$ is also multivariate normal with

$$AX + b \sim \mathscr{N}(A\mathbb{E}[X] + b, A\mathrm{Var}(X)A^T)$$

Theorem 8. If $X_{m\times 1} \sim \mathscr{N}(0_{m\times 1}, y_{m\times m})$ and $\det(y) \neq 0$. Then $g(X, y) = X^T y^{-1} X \sim \chi^2_{\dim X}$. Moreover, suppose $X_n \xrightarrow{X}_{m\times 1} \sim \mathscr{N}(0, y_{m\times m})$ with y invertible and $y_n \xrightarrow{\mathbb{P}} y_{m\times m}$. Then

$$X_n^T y_n^{-1} X_n \xrightarrow{d} X^T y^{-1} X \sim \chi^2_{\dim(X)}$$

Remark 21. Suppose

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + U$$

Define

$$X_{k+1\times 1} = (1, X_1, \ldots, X_k)^T$$

and

$$\beta_{k+1\times 1} = (\beta_0, \beta_1, \ldots, \beta_k)^T.$$

Then

$$Y = X^T \beta + U_{|X|}$$

Remark 22. Again, there are three interpretations for the SLR:

(a) (Linear) Assume $\mathbb{E}[Y \mid X] = X^T\beta$. Define $U = Y - \mathbb{E}[Y \mid X] = Y - X^T\beta$. Then $Y = X^T\beta + U = \mathbb{E}[Y \mid X] + U$. Then the β are not casual. But then

$$\mathbb{E}[U \mid X] = \mathbb{E}[Y - \mathbb{E}[Y \mid X] \mid X] = 0$$

and thus U is mean independent of X. Moreover,

$$\mathbb{E}[XU] = \mathbb{E}[\mathbb{E}[XU \mid X]] = \mathbb{E}[X\mathbb{E}[U \mid X]] = 0.$$

Note that this is enough (from PSET) to say that

$$\mathbb{E}[U] = 0.$$

(b) (BLP) The BLP $(Y \mid X)$ is the function that solves

$$\min_{b\in\mathbb{R}^{k+1}} \mathbb{E}[(Y - X^T b)^2]$$

which can be shown to be equivalent to

$$\min_{b\in\mathbb{R}^{k+1}} \mathbb{E}[(\mathbb{E}[Y \mid X] - X^T b)^2].$$

Then, once we find $\text{BLP}(Y \mid X) = X^T\beta$, we define $U = Y - X^T\beta$. Rewriting the minimization problem, we have that

$$\min_{b\in\mathbb{R}^{k+1}} \mathbb{E}[(Y - X^T b)^2]$$

Taking derivative with respect to b, we see that

$$\text{FOC}_b: \qquad -2\mathbb{E}[(Y - X^T\beta)X^T] = 0$$

applying the transpose and ignoring the -2, we see that (since $Y - X^T\beta$ is a scalar and is therefore its own transpose),

$$\mathbb{E}[X(Y - X^T\beta)^T] = \mathbb{E}[X(Y - X^T\beta)] = \mathbb{E}[XU] = 0.$$

So we get for free that $\mathbb{E}[XU] = 0$, and thus $\mathbb{E}[U] = 0$ and $\text{Cov}(X_j, U) = 0$ for any $j \in [k]$.

(c) (Causal Model) Assume

$$Y = g(X, U),$$

where X are the observed covariates of Y and U are the unobserved covariates. That is, if $g(X,U) = X^T\beta + U$, then $Y = X^T\beta + U$, where $\beta_j = \frac{\partial Y}{\partial X_j}$ is the causal effect of X_j on Y, holding X_{-j} and U constant. Thus,

$$Y = \beta_0 + X_{-0}^T\beta_{-0} + U = (\beta_0 + \mathbb{E}[U]) + X_0^T\beta_{-0} + (U - \mathbb{E}[U]) = \beta_0' + X_{-0}^T\beta_{-0} + U'$$

Hence,

$$\mathbb{E}[U'] = 0 \qquad \text{Cov}(X_j, U) \neq 0$$

1.7 Monday, July 7: Interactions

Definition 19. (Notation) We notate

$$X_{-j} = (1, X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_k)^T$$

Example 1.12. (Non Linear) Suppose

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_1 X_1^3 + \beta_4 X_2 + U_i$$

Hence,

$$\frac{\partial \text{BLP}}{\partial X_1} = \beta_1 + 3\beta_1$$

can be interpreted as the effect of X_1 on Y, here β_1 is the effect when $X_1 = 0$, β_2 is the sensitivity of the Y with respect to X_1 (if positive, then X_1 has an increasing effect on Y).

Example 1.13. (Interactions)

(a) (Dummy + Cont) Suppose $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$, where X_1 is 1 or 0 and X_2 is a continuum and let's assume a causal model. Then β_1 is the effect of X_1 on Y regardless of X_2. And vice-versa for β_2. The problem is that there is no way of measuring the interaction between X_1 and X_2. Consider now

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + U.$$

Suppose $X_1 = 0$, then $Y = \beta_0' + \beta_2' X_2 + u$. If $X_1 = 1$, then $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X_2 + U$, which is great because this shows there is a different intercept $(\beta_0 + \beta_1)$ and slope $(\beta_2 + \beta_3)$ for different X_1. This is able to capture the interaction much better.

(b) (Cont + Cont) Suppose $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + U$. Then

$$\frac{\partial \text{BLP}}{\partial X_1} = \beta_1 + \beta_3 X_2$$

Then β_3 is the sensitivity of X_1 on Y with respect to X_2.

(c) (Dummy + Dummy/Difference in Differences) Suppose

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + U,$$

where both X_1 and X_2 are binary

X_1/X_2	0	1	Diff
0	β_0	$\beta_0 + \beta_2$	β_2
1	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Diff	β_1	$\beta_1 + \beta_3$	β_3

Table 1: Interactions between Dummies

β_3 is known as the difference between differences coefficient.

Definition 20. We say that $X_{k+1\times 1}$ is **perfectly colinear** if there exists some $\mathbf{c} = (c_1, \dots, c_{k+1})^T \neq \mathbf{0}$ such that

$$\mathbf{c}X = 0$$

Lemma 4. Suppose X is not perfectly colinear, then $\mathbb{E}[XX^T]$ is invertible.

Proof. Suppose not. Then there exists some $\mathbf{c} \neq 0$ such that

$$\begin{aligned}
0 &= \mathbb{E}[XX^T]\mathbf{c} \\
&= \mathbf{c}^T\mathbb{E}[XX^T]\mathbf{c} \\
&= \mathbb{E}[\mathbf{c}^T XX^T\mathbf{c}] \\
&= \mathbb{E}[(\mathbf{c}X)^2]
\end{aligned}$$

Implying that $\mathbf{c}X = 0$ and thus X is perfectly colinear, a contradiction. $\square$

Remark 23. Let X_1, X_2 be binary. Then if X contains X_1 and X_2, then X is perfectly colinear, as $0 = 1 - (X_1 + X_2)$. If X_1 and X_2 are colinear. DO NOT build a regressionn with

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

, instead do difference in means

$$Y = \beta_0 + \beta_1 X_2 + U$$

where β_0 and β_1 are the difference in means

or do

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

1.8 Wednesday, July 9: β Theory

Theorem 9.

(Deriving β) let $Y_{1\times 1}, X_{k+1\times 1}, U_{k+1\times 1}$ be R.V.s with $Y = X^T\beta + U$ such that

(a) $\mathbb{E}[XU] = 0$ (which implies $\mathbb{E}[U] = 0, \mathrm{Cov}(X_j, U) = 0$).

(b) No perfect co-linearity in X.

(c) $\mathbb{E}[XX^T] < \infty$ (which implies $\mathbb{E}[X_j^2] < \infty$ and $\mathbb{E}[X_jX_s] < \infty$)

Let $(Y^1, (X^1)^T), \ldots, (Y^n, (X^n)^T) \sim (Y, X^T)$ i.i.d. Then

$$\beta = \mathbb{E}[XX^T]^{-1}\mathbb{E}[XY] \tag{8}$$

Proof. From (a), we see that

$$\begin{aligned} 0 &= \mathbb{E}[XU] \\ &= \mathbb{E}[X(Y - X^T\beta)] \\ &= \mathbb{E}[XY] - \mathbb{E}[XX^T]\beta \end{aligned}$$

Rearranging, we see that $\mathbb{E}[XX^T]\beta = \mathbb{E}[XY]$, and thus we use Lemma 4 to conclude. □

Theorem 10.

(**Frisch-Waugh-Lovell**) With the same assumption as in Theorem 9, define

(a) $Y^* := Y - \mathrm{BLP}(Y \mid X_{-j})$

(b) $X_j^* := X_j - \mathrm{BLP}(X_j \mid X_{-j})$

If

$$Y^* = \beta_0^* + \beta_j^* X_j^* + U^*,$$

then $\beta_j^* = \beta_j$

Proof. From the PSET,

- $\mathrm{Cov}(X_j^*, X_\ell) = 0$ for all $\ell \neq j$.
- $\mathrm{Cov}(X_j^*, X_j) = \mathrm{Var}(X_j^*)$ (decompose X_j with BLP)
- $\mathrm{Cov}(X_j^*, U) = 0$ open up X_j^*

Denoting $\text{BLP}(X_j \mid X_{-j}) = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{j-1}X_{j-1} + \alpha_{j+1}X_{j+1} + \cdots + \alpha_k X_k$. Then computing,

$$\begin{aligned}
\beta_j^* &= \frac{\text{Cov}(X_j^*, Y^*)}{\text{Var}(X_j^*)} \\
&= \frac{\text{Cov}(X_j^*, Y) - \text{Cov}(X_j^*, \text{BLP}(Y \mid X_{-j}))}{\text{Var}(X_j^*)} \\
&= \frac{\text{Cov}(X_j^*, Y)}{\text{Var}(X_j^*)} && \text{by (a)} \\
&= \frac{\text{Cov}(X_j^*, \beta_j X_j + X_{-j}^T \beta_{-j} + U)}{\text{Var}(X_j^*)} \\
&= \frac{\beta_j \text{Cov}(X_j^*, X_j) + \text{Cov}(X_j^*, X_{-j}^T \beta_{-j}) + \text{Cov}(X_j^*, U)}{\text{Var}(X_j^*)} \\
&= \beta_j
\end{aligned}$$

□

Remark 24. By the FWL thm, we interpret β_j as the $\text{BLP}(Y \mid X)$ as partial statistical association between X_j and Y, controlling for X_{-j}. We loosely call this partial correlation:

$$Z_y = \frac{Y - \mu_y}{\sigma_Y}, \quad Z_{X_j} = \frac{X_j - \mu_{X_j}}{\sigma_{X_j}},$$

and

$$Z - y = \beta_1 Z_{X_1} + \cdots + \beta_k Z_{X_k} + U,$$

where β_1 is the partial correlation between X_1 and Y.

Moreover, FWL <u>also works in the sample</u>. That is, definiting all equal to the above but in the sample,

$$\hat{\beta}_j^* = \hat{\beta}_j$$

Theorem 11.

(**Estimating** β) With the assumptions of Theorem 9, the OLS estimator of β is given by

$$\hat{\beta} = \left(\frac{1}{n}\sum X^i (X^i)^T\right)^{-1} \left(\frac{1}{n}\sum X^i Y^i\right) \tag{9}$$

Proof. We seek

$$\min_{b \in \mathbb{R}^{k+1}} \frac{1}{n}\sum (Y^i - (X^i)^T b)^2 = 0.$$

Taking derivative, we see that the first order conditions imply

$$\begin{aligned}
(0)^T &= \left(\sum (Y^i - (X^i)^T \beta)(X^i)^T\right)^T \\
&= \sum X^i (Y^i - (X^i)^T \beta) \\
&= \sum X^i Y^i - \beta \sum X^i (X^i)^T
\end{aligned}$$

Rearranging we get the result. □

Remark 25. In the step when we apply the transpose, we can see that

$$\frac{1}{n}\sum X^i\hat{U}^i = 0 \tag{10}$$

which are the mechanical equations of the OLS. This of course implies that

$$\frac{1}{n}\sum \hat{U}^i = 0 \tag{11}$$

This is similar to how $\mathbb{E}[XU] = \mathbb{E}[U] = 0$.

Definition 21. We say that the **fitted/predicted** value is

$$\hat{Y}^i := (X^i)^T\hat{\beta}.$$

We define the **residual** is

$$\hat{U}^i := Y^i - \hat{Y}^i = Y^i - (X^i)^T\hat{\beta}$$

Remark 26. We still define

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

with $R^2 \in (0,1)$. However, we claim that R^2 never decreases with the inclusion of a new regressor. That is, SSR never increases with this inclusion. Thus, we motivate the definition of the **adjusted** R^2

$$\overline{R}^2 := 1 - \frac{(n-1)}{(n-k-1)}\frac{\text{SSR}}{\text{TSS}}$$

Theorem 12.

(Deriving β) let $Y_{1\times1}, X_{k+1\times1}, U_{k+1\times1}$ be R.V.s with $Y = X^T\beta + U$ such that

(a) $\mathbb{E}[XU] = 0$ (which implies $\mathbb{E}[U] = 0, \text{Cov}(X_j, U) = 0$).

(b) No perfect co-linearity in X.

(c) $\mathbb{E}[XX^T] < \infty$ (which implies $\mathbb{E}[X_j^2] < \infty$ and $\mathbb{E}[X_jX_s] < \infty$)

Let $(Y^1, (X^1)^T), \ldots, (Y^n, (X^n)^T) \sim (Y, X^T)$ i.i.d. The OLS estimator for β satisfies the following:

(a) (**Unbiased**) If $\mathbb{E}[U \mid X] = 0$, then

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

(b) (**Consistent**) If $\mathbb{E}[Y_j^2] < \infty$ for all $j = 1, 2, \ldots, k+1$, then

$$\hat{\beta} \xrightarrow[\mathbb{P}]{} \beta$$

(c) (**Asymptotic Distribution**) If $\mathbb{E}[X_j^4] < \infty$ and $\mathbb{E}[Y_j^4] < \infty$, then

$$\sqrt{n}(\hat{\beta} - \beta) \sim N(0, \Sigma)$$

where

$$\Sigma = \mathbb{E}[XX^T]^{-1}\text{Var}(XU)\mathbb{E}[XX^T]^{-1}$$

Proof. (a) We have that

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{n}\sum X^i(X^i)^T\right)^{-1}\left(\frac{1}{n}\sum X^iY^i\right)\\ &= \left(\frac{1}{n}\sum X^i(X^i)^T\right)^{-1}\left(\frac{1}{n}\sum X^i((X^i)^T\beta + U^i)\right)\\ &= \beta + \left(\frac{1}{n}\sum X^i(X^i)^T\right)^{-1}\frac{1}{n}\sum X^iU^i\end{aligned}$$

Taking conditional expectation of

$$\hat{\beta} = \beta + \left(\frac{1}{n}\sum X^i(X^i)^T\right)^{-1}\frac{1}{n}\sum X^iU^i, \tag{12}$$

we see that if we use (a) on (12),

$$\mathbb{E}[\hat{\beta} \mid X_1, \dots, X_n] = \beta_1 + \left(\frac{1}{n}\sum X^i(X^i)^T\right)^{-1}\frac{1}{n}\sum X^i\mathbb{E}[U^i \mid X_1, \dots, X_n] = \beta_1$$

Conclude with LIE.

(b) By the WLLN, we have the convergence of

$$\frac{1}{n}\sum X^i(X^i)^T \underset{\mathbb{P}}{\to} \mathbb{E}[XX^T]$$

$$\frac{1}{n}\sum X^iY^i \underset{\mathbb{P}}{\to} \mathbb{E}[XY]$$

so then using the CMT with $g(A,B) = A^{-1}B$ and Lemma 4, we conclude that

$$\hat{\beta} = g\left(\frac{1}{n}\sum X_iX_i^T, \frac{1}{n}\sum X_iY_i\right) \underset{\mathbb{P}}{\to} g(\mathbb{E}[X_iX_i^T], \mathbb{E}[X_iY_i]) = \beta$$

(c) From equation (12) we use the WLLN, CLT, CMT, and Slutsky to see that

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta) &= \left(\frac{1}{n}\sum X^i(X^i)^T\right)^{-1}\frac{1}{\sqrt{n}}\sum X^iU^i\\ &\underset{\mathbb{P}}{\to} \mathbb{E}[XX^T]^{-1}\frac{1}{\sqrt{n}}\sum X^iU^i\\ &\underset{\mathscr{D}}{\to} \mathbb{E}[XX^T]^{-1}N(0, \mathrm{Var}(XU))\\ &= N\left(0, \mathbb{E}[XX^T]^{-1}\mathrm{Var}(XU)\mathbb{E}[XX^T]^{-1}\right)\end{aligned}$$

Where the last equality is due to $\mathbb{E}[XX^T]$ being invertible and thus symmetric.

□

Remark 27. (Omitted Variable Bias) Suppose

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + U$$

satisfies the conditions of Theorem 12 (b). Suppose it is difficult to measure X_2 directly, so consider estimating with OLS the new

$$Y = \alpha_0 + \alpha_1X_1 + V.$$

We know by Proposition 7 that

$$\begin{aligned}
\hat{\alpha}_1 \xrightarrow[\mathbb{P}]{} \alpha_1 &= \frac{\text{Cov}(X_1, Y)}{\text{Cov}(X_1)} \\
&= \frac{\text{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U)}{\text{Var}(X_1)} \\
&= \beta_1 + \beta_2 \frac{\text{Cov}(X_1, X_2)}{\text{Cov}(X_2)}
\end{aligned}$$

(OVB)

$\beta_2/(\text{Cov}(\text{X}_1, X_2)$	$+$	$-$	0
$+$	$+$	$-$	0
$-$	$-$	$+$	0
0	0	0	0

Table 2: Effects of OVB

Example 1.14. Suppose a causal MLR of

$$\text{muscle mass} = \beta_0 + \beta_1 \text{gymtime} + \beta_2 \text{genes} + U$$

Genes are hard to measure, so we consider estimating the following with OLS

$$\text{muscle mass} = \alpha_0 + \alpha_1 \text{gymtime} + V.$$

By OLS, We require that $\text{Cov}(V, \text{gymtime}) = 0$. The following step would be to investigate β_2 and $\text{Cov}(X_1, X_2)$ and use Table 2.

Remark 28. (**Measurement Error**)

- A measurement error in Y is benign, and doesn't cause OVB, only increases $\text{Var}(\hat{\beta}_j)$
- Measurement error in X is bad. Common case is classical measurement error (CME). To see this, suppose
  $$Y = \beta_0 + \beta_1 X + U,$$
  but a researcher estimates
  $$Y = \alpha_0 + \alpha_1 X^* + V$$
  with OLS, where X^* is a ill-measured X with $X^* = X + Z$, where $\mathbb{E}[Z] = \text{Cov}(X, Z) = \text{Cov}(U, Z) = 0$. We know that
  $$\begin{aligned}
  \hat{\alpha}_1 \xrightarrow[\mathbb{P}]{} \alpha_1 &= \frac{\text{Cov}(X^*, Y)}{\text{Cov}(X^*)} \\
  &= \frac{\text{Cov}(X + Z, \beta_0 + \beta_1 X + U)}{\text{Var}(X + Z)} \\
  &= \frac{\beta_1 \text{Var}(X) + \text{Cov}(X, U) + \beta_1 \text{Cov}(Z, X) + \text{Cov}(Z, U)}{\text{Var}(X) + \text{Var}(Z) + 2\text{Cov}(X, Z)}
  \end{aligned}$$
  Using various assumptions, we have derived

(Attenuation Bias)

$$\hat{\alpha_1} \xrightarrow[\mathbb{P}]{} \beta_1 \underbrace{\frac{\text{Var}(X)}{\text{Var}(X) + \text{Var}(Z)}}_{\text{attenuation bias, } \leq 1}$$

Measurement error always pulls $\hat{\alpha}_1$ towards 0.

Proposition 9. If U is homoskedastic, then

$$\Sigma = \mathbb{E}[XX^T]^{-1}\text{Var}(U).$$

Proof. Since U is homoskedastic, then $\mathbb{E}[U \mid X] = 0$ and $\text{Var}(U \mid X) = \text{Var}(U)$. Hence,

$$\begin{aligned}
\text{Var}(XU) &= \mathbb{E}[XX^T U^2] - \mathbb{E}[XU]\mathbb{E}[XU]^T \\
&= \mathbb{E}[XX^T \mathbb{E}[U^2 \mid X]] \\
&= \mathbb{E}[XX^T \text{Var}(U \mid X)] \\
&= \mathbb{E}[XX^T \text{Var}(U)] \qquad \text{(homosk)} \\
&= \text{Var}(U)\mathbb{E}[XX^T]
\end{aligned}$$

Thus,

$$\Sigma = \mathbb{E}[XX^T]^{-1}\text{Var}(U)\mathbb{E}[XX^T]\mathbb{E}[XX^T]^{-1} = \text{Var}(U)\mathbb{E}[XX^T]^{-1}$$

□

Remark 29. If U is not homosk, then using the analogy principle,

$$\hat{\Sigma} = \left[\frac{1}{n}\sum X^i(X^i)^T\right]^{-1} \left[\frac{1}{n}\sum (\hat{U}^i)^2 X^i(X^i)^T\right] \left[\frac{1}{n}\sum X^i(X^i)^T\right]^{-1} \tag{13}$$

1.9 Friday, July 11: Hypothesis Testing

Lemma 5. Suppose $X_n \underset{\mathscr{D}}{\longrightarrow} N(0, y_n)$ and $y_n \underset{\mathbb{P}}{\to} y$. Then

$$X_n^T y^{-1} X_n \underset{\mathscr{D}}{\longrightarrow} X^T y^{-1} X \sim \chi^2_{\dim(X)}$$

Remark 30. (Hypothesis Testing for Linear Combinations of β)

(1) $H_0 : R\beta = r$, $H_a : R\beta \neq r$, where, usually $r = \mathbf{0}$ and R is the matrix testing the linear combinations.

(2) Since $\sqrt{n}(\hat{\beta} - \beta) \underset{\mathscr{D}}{\longrightarrow} N(0, \Sigma)$, then by Slutsky's Lemma

$$\sqrt{n}(R\hat{\beta} - R\beta) \underset{\mathscr{D}}{\longrightarrow} N(0, R\Sigma R^T).$$

As usual, Σ is unattainable, so we use (13) as a consistent estimator. From Lemma 5,

(Wald's Statistic)

$$T_n = n(R\hat{\beta} - R\beta)^T (R\hat{\Sigma} R^T)^{-1} (R\hat{\beta} - R\beta) \underset{\mathscr{D}}{\longrightarrow} \chi^2_{\dim(R\beta)}$$

(3) The usual end for hypothesis testing.

Example 1.15.

Example 1.16. Suppose $Y = \beta_0 + \beta_1 X_1 + \beta_2 + U$, with

$$H_0 : \beta_1 + 2\beta_2 = 0 \quad H_a : \beta_1 + 2\beta_2 \neq 0.$$

Then

$$R = \begin{bmatrix} 0 & 1 & 2 \end{bmatrix}$$

and $r = \mathbf{0}$.

Example 1.17.

$$H_0 : \beta_1, \beta_2 = 0 \quad H_a : \beta_1 \neq 0 \cup \beta_2 \neq 0$$

, then

$$R = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \implies R\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

Under H_0, $R\beta = r$, where $r = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Recall that

$$\sqrt{n}(\hat{\beta} - \beta) \to N(0, \Sigma) \implies \sqrt{n}(R\hat{\beta} - R\beta) \to N(0, R\Sigma R^T)$$

we can measure the scalar of how far away the data is from the values:

$$n(R\hat{\beta} - R\beta)^T (R \sum R^T)^{-1} (R\hat{\beta} - R\beta)$$

Let

$$T_n = n(R\hat{\beta} - R\beta^{H_0})^T (R\hat{\Sigma} R^T)^{-1} (R\hat{\beta} - R\beta) \underset{d}{\to} \chi^2_{\dim(R\beta)}$$

Remark 31. (**F-Statistic Hypothesis Testing**) Assume U is homoskedastic., and consider the SLR

$$Y^i = \beta_0 + \beta_1 X_1^i + \beta_2 X_2^i + U^i \quad \text{(unrestricted model)}.$$

(1)
$$H_0 : \beta_1, \beta_2 = 0 \qquad H_a : \beta_1 \neq 0 \cup \beta_2 \neq 0$$

(2) Under H_0 :
$$Y^i = \beta_0 + U^i \quad \text{(restricted model)}$$

Compute
$$F_n = \frac{\frac{1}{q}\left(\mathrm{SSR}_r - \mathrm{SSR}_{ur}\right)}{\frac{1}{(n-k_{ur}-1)}\mathrm{SSR}_{ur}} = \frac{\frac{1}{q}\left(R_{ur}^2 - R_r^2\right)}{\frac{1}{(n-k_{ur}-1)}(1-R_{ur}^2)},$$

where q is the number of constraints, and K_{ur} is the number of regressor in the ur model.

(3) F is distrbuted as $F_{q,\, n-K_{ur}-1}$

1.10 Monday, July 14: Instrument Variables in SLR

Definition 22. We say that X_j is **endogenous** if $\text{Cov}(X_j, U) \neq 0$. Else, we say that X_j is **exogenous.**

Until otherwise stated, we consider

$$Y = X^T\beta + U,$$

where there is at least one endogenous variable. From Remark 22, (c), we note that $\mathbb{E}[U] = 0$. how do we estimate β in this scenario?

Remark 32. There are three main sources of endogeneity:

(a) OVB (when U is alive, see Example 1.9)

(b) Measurement Error (When X is not quite right, see Remark 28)

(c) Simultaneity Bias (See Example 1.19)

Definition 23. Suppose $Y = X^T\beta + U$. We say that a r.v. Z is an **instrument** if

- $\text{Cov}(Z, U) = \mathbb{E}[ZU] = 0$ (instrument exogeneity)
- $\text{Cov}(X, Z) \neq 0$ (instrument relevance)

Example 1.18. (Civil conflict in Africa) Consider

$$\text{conflict}_i = \beta_0 + \beta_1 \text{growth}_i + U_i,$$

and consider

$$Z_i : \text{rainfall s.t. } \text{Cov}(X, Z) \neq 0, \text{Cov}(Z, U) = 0$$

Remark 33. (**Case 1: SLR**)

Definition 24. Suppose $Y = \beta_0 + \beta_1 X + U$. We say that a r.v. Z is an **instrument** if

- $\text{Cov}(Z, U) = \mathbb{E}[ZU] = 0$ (instrument exogeneity)
- $\text{Cov}(X, Z) \neq 0$ (instrument relevance)

Suppose $Y = \beta_0 + \beta_1 X + U$, where X is endogenous.

We derive

$$\begin{aligned} 0 &= \mathbb{E}[U] \\ &= \mathbb{E}[Y] - \beta_0 - \beta_1 \mathbb{E}[X] \end{aligned}$$

gives us

$$\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] \tag{14}$$

Plugging in (14)

$$\begin{aligned} 0 &= \mathbb{E}[ZU] \\ &= \mathbb{E}[Z(Y - \beta_0 - \beta_1 X)] \\ &= \mathbb{E}[Z(Y - \mathbb{E}[Y] + \beta_1 \mathbb{E}[X] - \beta_1 X)] \\ &= \mathbb{E}[Z(Y - \mathbb{E}[Y])] - \beta_1 \mathbb{E}[Z(X - \mathbb{E}[X])] \\ &= \text{Cov}(Z, Y) - \beta_1 \text{Cov}(Z, X) \end{aligned}$$

give us

$$\beta_1 = \frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(Z, X)} \tag{15}$$

Note that when X is binary, this nicely simplifies to

(Local Average Treatment)

$$\beta_1 = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[X \mid Z = 1] - \mathbb{E}[X \mid Z = 0]} = \frac{\pi_1}{\alpha_1}$$

where π_1 and α_1 are the coefficients from regressing Y and X on Z, respectively.

Theorem 13.

(Estimation with SLR Instrument Variables) let Y, X, U, Z be R.V.s with $Y = \beta_0 + \beta_1 X + U$ such that

(a) $\mathbb{E}[U] = 0$.

(b) $\mathbb{E}[ZU] = 0$ (Exogeneity).

(c) $\mathrm{Cov}(X, Z) \neq 0$

Let $(Y^1, X^1, Z^1), \ldots, (Y^n, X^n, Z^n) \sim (Y, X, Z)$ i.i.d. Then the IV estimator for β given by

$$\hat{\beta}_1^{IV} = \frac{\hat{\sigma}_{ZY}}{\hat{\sigma}_{ZX}} \tag{16}$$

$$\hat{\beta}_0^{IV} = \bar{Y} - \hat{\beta}_1^{IV} \bar{X} \tag{17}$$

satisfies

(a) The sample equivalents of (a) and (b):

$$0 = \frac{1}{n} \sum (Y^i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} X^i) = \frac{1}{n} \sum \hat{U}^i$$

$$0 = \frac{1}{n} \sum Z^i (Y^i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} X^i) = \frac{1}{n} \sum Z^i \hat{U}^i$$

(b) (**Consistent**) If $\mathbb{E}[Y^2], \mathbb{E}[Z^2], \mathbb{E}[X^2] < \infty$, then both $\hat{\beta}_0^{IV}$ and $\hat{\beta}_1^{IV}$ are consistent.

(c) (**Asymptotic Distribution**) If $\mathbb{E}[X^4], \mathbb{E}[Y^4], \mathbb{E}[Z^4] < \infty$, then

$$\sqrt{n}(\hat{\beta}_1^{IV} - \beta_1) \sim N(0, \sigma_{1,IV}^2)$$

where

$$\sigma_{1,IV}^2 = \frac{\mathrm{Var}((Z - \mathbb{E}[Z])U)}{\mathrm{Cov}^2(X, Z)}$$

Proof. (a) From the first equation, we can derive $\hat{\beta}_0^{IV}$.

$$\begin{aligned} 0 &= \frac{1}{n} \sum (Y^i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} X^i) \\ &= \bar{Y} - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} \bar{X} \end{aligned}$$

Similarly, we plug in (17) into the second equation to see how

$$\begin{aligned}
0 &= \frac{1}{n}\sum Z^i(Y^i - \hat\beta_0^{IV} - \hat\beta_1^{IV}X^i) \\
&= \frac{1}{n}\sum Z^i(Y^i - (\bar Y - \hat\beta_1^{IV}\bar X) - \hat\beta_1^{IV}X^i) \\
&= \frac{1}{n}\sum (Y^i - \bar Y)Z^i - \hat\beta_1^{IV}\frac{1}{n}\sum (X^i - \bar X)Z^i
\end{aligned}$$

yields the answer.

(b) We know that $\hat\sigma_{YZ} \underset{\mathbb{P}}{\to} \sigma_{YZ}$ and $\hat\sigma_{XZ} \underset{\mathbb{P}}{\to} \sigma_{XZ}$. use continuous mapping theorem to conclude.

(c) We can write

$$\hat\beta_1^{IV} = \frac{\sum(Z^i - \bar Z)Y^i}{\sum(Z^i - \bar Z)X^i} = \frac{\sum(Z^i - \bar Z)(\beta_0 + \beta_1 X^i + U^i)}{\sum(Z^i - \bar Z)X^i} = \beta_1 + \frac{\sum(Z^i - \bar Z)U^i}{\sum(Z^i - \bar Z)X^i} \tag{18}$$

From (18) we see that, by using a combination of WLLN, CLT, then WLLN, then Slutsky's Lemma,

$$\begin{aligned}
\sqrt{n}(\hat\beta_1^{IV} - \beta_1) &= \frac{\frac{1}{\sqrt{n}}\sum(Z^i - \bar Z)U^i}{\frac{1}{n}\sum(Z^i - \bar Z)X^i} \\
&\underset{\mathbb{P}}{\to} \frac{\frac{1}{\sqrt{n}}\sum(Z^i - \mathbb{E}[Z])U^i}{\frac{1}{n}\sum(Z^i - \mathbb{E}[Z])X^i + \frac{1}{n}\sum(\mathbb{E}[Z] - \bar Z)X^i} \\
&\underset{\mathbb{P}}{\to} \frac{\frac{1}{\sqrt{n}}\sum(Z^i - \mathbb{E}[Z])U^i}{\mathrm{Cov}(Z, X)} \\
&\underset{\mathscr{D}}{\to} \frac{N(0, \mathrm{Var}((Z - \mathbb{E}[Z])U))}{\mathrm{Cov}(Z, X)}
\end{aligned}$$

As usual, we usually estimate this variance by

$$\hat\sigma_{1,IV}^2 = \frac{\sum(Z_i - \bar Z)\hat U_i}{\sum(Z_i - \bar Z)(X_i - \bar X)}$$

$\square$

Remark 34. (**Biased**) What happens to the bias? In order to establish unbiasness, it is clear from (18) that we would need $\mathbb{E}[U^i \mid X^i, Z^i] = 0$. But then using the LIE it is clear that $\mathbb{E}[X^iU^i] = 0$. But this then implies that $\mathrm{Cov}(X^i, U^i) = 0$, a contradiction to X^i being endogenous.

Example 1.19. (**Simultaneity Problem**) Let

$$Q^d(p) = \beta_0 + \beta_1 p + U^d$$

$$Q^s(p) = \alpha_0 + \alpha_1 p + \alpha_2 Z + U^s$$

assume $\beta_1 < 0, \alpha_1 > 0$, and $\mathrm{Cov}(U^d, U^s) = 0$ and Z is a supply shifter with $\mathrm{Cov}(Z, U^s) = \mathrm{Cov}(Z, U^d) = 0$ and $\mathrm{Var}(Z) > 0$. Suppose further that $\alpha_2 \neq 0$.

Solving for P in $Q(P) = Q(P) = Q$ gives

$$\frac{1}{\alpha_1 - \beta_1}(\beta_0 - \alpha_0 - \alpha_2 Z + U^d - U^s),$$

and hence P is endogenous in both $Q^d(P)$ and $Q^s(P)$ with

$$\mathrm{Cov}(P, U^d) = \frac{\mathrm{Var}(U^d)}{\alpha_1 - \beta_1} > 0, \qquad \mathrm{Cov}(P, U^d) = \frac{-\mathrm{Var}(U^s)}{\alpha_1 - \beta_1} < 0$$

From OVB analysis, the above imply that (Example 1.9) the BLP coefficients overestimate β_1 and underestimate α_1.

It can be worked out that $\mathrm{Cov}(Z, P) \neq 0$ given the assumptions, implying that Z is an instrument variable and we can estimate a consistent estimator!

Remark 35. (**Case 2: MLR**) For the following case, we will consider the model $Y = X^T\beta + U$, where X_1 is endogenous and X_{-1} are exogenous and (WLOG) $\mathbb{E}[U] = 0$.

Definition 25. Suppose $Y = X^T\beta + U$. We say t We say that a r.v. Z is an **instrument** if

- (instrument exogeneity) $\mathrm{Cov}(Z, U) = \mathbb{E}[ZU] = 0$
- (instrument relevance) Letting $W = (1, Z, W_2, \dots, W_k)$ not be perfectly colinear and $\pi = (\pi_0, \pi_1, \dots, \pi_k)$ such that $\pi_1 \neq 0$, then
$$X_1 = \mathrm{BLP}(X_1 \mid W) + V = W^T\pi$$

Proposition 10. Suppose $Y = X^T\beta + U$ and X_1 is the only endogeneous variable. Then the following are equivalent:

(a) Z is relevant;

(b) $\mathbb{E}[WX^T]$ is invertible;

(c) $\mathbb{E}[WW^T]^{-1}\mathbb{E}[WX^T]$ is invertible

Proof. ($a \iff c$) Suppose Z is relevant. Consider regressing X by W. then the BLP coefficient is given by (see Theorem 9)

$$\alpha = \mathbb{E}[WW^T]^{-1}\mathbb{E}[WY] = \begin{pmatrix} 1 & \pi_0 & 0 & \cdots & 0 \\ 0 & \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \pi_k & 0 & 0 & 1 \end{pmatrix}$$

which is clearly if and only if invertible when $\pi \neq 0$.

($c \iff b$) This is a result from linear algebra. More generally, if A is an invertible matrix, then B is invertible if and only if $A^{-1}B$ is invertible. For the forward direction we see that since A and B are invertible, then $B^{-1}A$ is the inverse of $A^{-1}B$. For the backward direction, we note that products of invertible matrices are invertible, so then $B = A(A^{-1}B)$ is invertible. $\square$

With the assumptions above, we derive

$$0 = \mathbb{E}[WU] = \mathbb{E}[W(Y - X^T\beta)] = \mathbb{E}[WY] - \mathbb{E}[WX^T]\beta$$

and thus

$$\beta = \mathbb{E}[WX^T]^{-1}\mathbb{E}[WY] \tag{19}$$

Theorem 14.

(Estimation with MLR Instrument Variables) let $Y_{1\times 1}, X_{k+1\times 1}, U_{k+1\times 1}, Z_{k+1\times 1}$ be R.V.s with $Y = \beta_0 + \beta_1 X + U$ and $W = (1, Z, W_1, X_2, X_k)$ such that

(a) $\mathbb{E}[WU] = 0$ (Exogeneity)

(b) $\mathbb{E}[WX^T]$ is invertible (Relevance).

(c) W has no perfect co-linearity

Let $(Y^1, X^1, Z^1), \ldots, (Y^n, X^n, Z^n) \sim (Y, X, Z)$ i.i.d. Then the IV estimator for β given by

$$\hat{\beta}^{IV} = \left(\frac{1}{n}\sum W^i (X^i)^T\right)^{-1} \frac{1}{n}\sum W^i Y^i \tag{20}$$

satisfies

(a) The sample equivalents of (a):

$$0 = \frac{1}{n}\sum W^i (Y^i - (X^i)^T \hat{\beta}^{IV}) = \frac{1}{n}\sum W^i \hat{U}^i$$

(b) (**Consistent**) If $\mathbb{E}[Y^2], \mathbb{E}[Z^2], \mathbb{E}[X^2] < \infty$, then $\hat{\beta}^{IV}$ is consistent.

(c) (**Asymptotic Distribution**) If $\mathbb{E}[X^4], \mathbb{E}[Y^4], \mathbb{E}[Z^4] < \infty$, then

$$\sqrt{n}(\hat{\beta}^{IV} - \beta) \sim N(0, \Sigma_{IV})$$

where

$$\Sigma_{IV} = \mathbb{E}[WX^T]^{-1}\mathrm{Var}(WU)(\mathbb{E}[WX^T]^{-1})^T$$

Proof.

- From the equation in (a), it is a simple rearrangement to recuperate (20).
- Using the WLLN,

$$\frac{1}{n}\sum W^i (Y^i)^T \underset{\mathbb{P}}{\to} \mathbb{E}[WY^T], \qquad \frac{1}{n}\sum W^i Y^i \underset{\mathbb{P}}{\to} \mathbb{E}[WY]$$

  Using the continuous mapping theorem along with (b), we arrive at the result.
- Expanding (20),

$$\begin{aligned}\hat{\beta}^{IV} &= \left(\frac{1}{n}\sum W^i (X^i)^T\right)^{-1} \frac{1}{n}\sum W^i Y^i \\ &= \left(\frac{1}{n}\sum W^i (X^i)^T\right)^{-1} \frac{1}{n}\sum W^i ((X^i)^T \beta + U^i)\end{aligned}$$

  and thus

$$\hat{\beta}^{IV} = \beta + \left(\frac{1}{n}\sum W^i (X^i)^T\right)^{-1} \frac{1}{n}\sum W^i U^i \tag{21}$$

Rearranging (21),

$$\begin{aligned}\sqrt{n}(\hat{\beta}^{IV}-\beta) &= \left(\frac{1}{n}\sum W^i(X^i)^T\right)^{-1}\sqrt{n}\sum W^iU^i \\ &\xrightarrow[\mathbb{P}]{} \mathbb{E}[WX^T]^{-1}N(0,\mathrm{Var}(WU)) \\ &= N(0,\mathbb{E}[WX^T]^{-1}\mathrm{Var}(WU)(\mathbb{E}[WX^T]^{-1})^T)\end{aligned}$$

As usual (since we have that $\mathrm{Var}(WU)=\mathbb{E}[WW^TU^2]$), we estimate this variance by

$$\hat{\Sigma}_{IV} = \left(\frac{1}{n}\sum W^i(X^i)^T\right)^{-1}\left(\frac{1}{n}\sum W^i(W^i)^T\hat{U}^i\right)\left(\left(\frac{1}{n}\sum W^i(X^i)^T\right)^{-1}\right)^T$$

□

Remark 36. (**Testing Relevance**) One can run an OLS on

$$X_1 = W^T\pi = \pi_0 + \pi_1 Z + \cdots + \pi_k X_k$$

to test $H_0 : \pi_1 = 0$ and $H_a : \pi_0 \neq 0$. Use an F statistics where the rule of thumb is that $F > 10$ implies a relevant instrument, while $F \leq 10$ is an weak instrument that can inflate $SE(\hat{\beta}^{IV})$

Remark 37. **(2SLS)** Applying Remark 35 to the SLR case, we consider $Y = X^{\beta}+U$, where X is endogenous, and Z is an instrument such that

$$X = \mathrm{B\hat{L}P}(X \mid Z) + V = \underbrace{\hat{\pi}_0 + \hat{\pi}_1 Z}_{\hat{X}} + \hat{\varepsilon}, \qquad \text{(first stage)}$$

Because this is an OLS, then X^* is exogenous. Thus, we can run the OLS

$$Y = \mathrm{BLP}(Y \mid \hat{X}) + U^* = \hat{\beta}_0^{2SLS} + \beta_1^{2SLS}\hat{X} + \hat{U} \qquad \text{(second stage).}$$

With the assumptions of Theorem 13, estimating β gives

$$\hat{\beta}_1^{SLS} = \frac{\mathrm{Cov}(\hat{X},Y)}{\mathrm{Var}(\hat{X})} \qquad \hat{\beta}_1^{SLS} = \bar{Y} - \hat{\beta}_1^{SLS}\bar{\hat{X}}$$

with consistency results:

$$\hat{\beta}_0^{2SLS} \xrightarrow[\mathbb{P}]{} \beta_0 \qquad \hat{\beta}_1^{2SLS} \xrightarrow[\mathbb{P}]{} \beta_0$$

and

$$\hat{\beta}_1^{SLS} = \hat{\beta}_1^{IV}$$

1.11 Friday, July 16: Causal Inference

Remark 38. (**Rubin Causal Model**)

- Assignment mechanism is first come first serve,
- Define y_0^i to be the potential outcome of individual i if it does not receive the treatment.
- Define y_1^i as above.
- Let
$$D^i = \begin{cases} 1 & i \text{ did received treatment} \\ 0 & \text{else} \end{cases}$$
- Define $\tau^i = y_1^i - y_0^i$ to be the potential treatment difference between person i.

(Fundamental Problem fo Causal Inference) τ^i cannot be observed since we cannot clone people.

- Define the observed outcome to be
$$Y^i = y_0^i + D^i(y_1^i - y_0^i)$$
- Define the **average treatment effect** is defined by
$$ATE = \mathbb{E}[y_0^i - y_1^i]$$
- Define the the **average treatment effect on treated** is
$$ATT = \mathbb{E}[y_1^i - y_0^i \mid D^i = 1]$$
- Define the **average treatment effect on untreated is**
$$ATU = \mathbb{E}[y_1^i - y_0^i \mid D^i = 0]$$
- Define the naive treatment effect is
$$\theta = \mathbb{E}[Y^1 \mid D^i = 1] - \mathbb{E}[Y^i \mid D^i = 0] = \mathbb{E}[y_1^i \mid D_i = 1] - \mathbb{E}[y_0^i \mid D_i = 0]$$

Clearly, we can estimate
$$\hat{\theta} = \bar{Y}_T - \bar{Y}_C \xrightarrow[\mathbb{P}]{} \theta,$$
but $\theta \neq$ any AT$(\cdot)$ above!

Remark 39. (**Selection Bias an Treatment Effects**) Notice that
$$\begin{aligned} \theta &= \mathbb{E}[y_1^i \mid D_i = 1] - \mathbb{E}[y_0^i \mid D_i = 0] \\ &= \mathbb{E}[y_1^i \mid D_i = 1] - \mathbb{E}[y_0^i \mid D_i = 1] + \mathbb{E}[y_0^i \mid D_i = 1] - \mathbb{E}[y_0^i \mid D_i = 0] \\ &= ATT + \underbrace{\mathbb{E}[y_0^i \mid D_i = 1] - \mathbb{E}[y_0^i \mid D_i = 0]}_{SB_0 \text{ selection bias in } y_0} \end{aligned}$$

Similarly,
$$\boxed{\theta = \begin{cases} ATT + SB_0 \\ ATU + SB_1 \end{cases}} \tag{22}$$

By (22), we see that if $SB_0, SB_1 > 0$, then $\theta > ATU, ATU$. Also by 22, we see that
$$ATT + SB_0 = ATU + SB_1$$

Example 1.20. (The Golden Standard in RCM) In an experiment with randomization such that $y_0^i, y_1^i \perp D^i$, we see that

$$SB_0 = \mathbb{E}[y_0^i \mid D_i = 1] - \mathbb{E}[y_0^i \mid D_i = 0] = \mathbb{E}[y_0^i] - \mathbb{E}[y_0^i] = 0$$

and same for SB_1.

Moreover,

$$\begin{aligned} ATE &= \mathbb{E}[y_1^i - y_0^i] \\ &= \mathbb{E}[\mathbb{E}[y_1^i - y_0^i \mid D^i]] \\ &= p(ATT) + (1-p)(ATU) \\ &= p(\theta - SB_0) + (1-p)(\theta - SB_1) \\ &= \theta \end{aligned}$$

Thus, in a randomized experiment, $\boxed{ATT = ATU = ATE = \theta}$

Remark 40. (**Difference in Differences Model**) We consider the model

$$\mathrm{Y}_t^i = \beta_0 + \beta_1 \mathrm{D}^i + \beta_2 \mathrm{Post}_t + \beta_3 (\mathrm{D}^i \times \mathrm{Post}_t) + U^i$$

where

- Y_{it}: Outcome for unit i at time t
- D^i: Treatment group indicator (1 if treated, 0 otherwise)
- Post_t: Post-treatment period indicator
- β_3: DiD estimator (treatment effect)

$$\begin{aligned} \text{ATT} &= E[Y_{1i} - Y_{0i} \mid \text{Treat}_i = 1] \\ \beta_3 &= \underbrace{(\text{Treatment}_{\text{Post}} - \text{Treatment}_{\text{Pre}})}_{\text{Treatment group change}} - \underbrace{(\text{Control}_{\text{Post}} - \text{Control}_{\text{Pre}})}_{\text{Control group change}} \end{aligned}$$

These two are equal if:

(a) **Parallel Trends**: Control group represents counterfactual trend
(b) **No Anticipation**: Treatment doesn't affect pre-period outcomes
(c) **SUTVA**: No interference between units

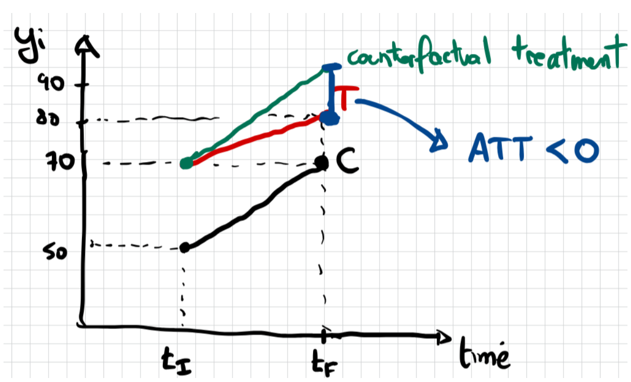

Figure 1: PTA Visualized

2 Commandments of Econometrics

(First Commandment) Never assume homoskedasticity, always compute the robust

$$SE(\hat{\beta})$$

Reason:

almost never is, and hard to see!

(Second Commandment) Never build a model with perfect colinearity in $\mathbf{X}$.

Reason:

$\hat{\beta}$ won't exist!

(Third Commandment) Never use OLS to estimate supply and demand. Use instrument variables instead.

Reason:

P is endogenous!