Causal Effects in Policy - Lecture 1

Alice Wu

UW Madison Econ 695

Course Overview

1. Descriptive modeling
   1. Empirical relationships between X and Y: regression analysis
   2. Decomposition techniques
2. Causal modeling: does X "cause" Y?
   1. Potential outcomes framework
   2. Randomized control trials: randomly assign treatments
   3. Natural/Quasi experiments: diff-in-diff, instrument variables, LATE, regression discontinuity
3. Statistical Learning (Machine Learning)
   1. Classification methods
   2. Supervised & Unsupervised learning

1. Descriptive Modeling

Often we are interested in trying to *summarize the relationship* between some "outcome" y and some other variables $x = (x_1, x_2 ... x_J)$.

- we **aren't** necessarily trying to measure the causal effect of x_j on y
- we are trying to take account of the fact that y may be strongly related to some $x_j's$ and only weakly related to others.
- e.g., what is the relationship between earnings (y), gender (x_1), and other characteristics, like education (x_2)?
- our benchmark: the conditional expectation function $E[y|x]$

Benchmark: CEF $E[y|x]$

- CEF itself can be nonlinear.
- we are going to approximate this with a **linear** regression function
  - how does the best linear approximation relate to the CEF?
- we'll consider 2 regression functions in order to get the best linear approx. to CEF:
  - the "population" regression: the function we could estimate with the population (∞ sample)
  - the "sample" regression: the one we can actually estimate with a given sample (drawn randomly from the population?)

2. Causal Modeling

- Many debates in economics amount to disputes over the question: does X "cause" Y?
- *Examples*
  - if a student were to attend Univ. Minnesota instead of UW, would she get a higher-paid job after college?
  - if an unemployed worker has higher UI benefits, will he/she have longer spell of joblessness?
  - does taking Econ 695 increase the chance of getting a data science job after college?

2. Causal Modeling

- *A very empirical notion of causality:* X causes Y if, in an ideal experiment, we could manipulate (randomly change) X, leaving other factors (W) constant, and observe that the **mean** of the distribution of outcomes of Y has changed.

$$E[Y|X = 1, W] - E[Y|X = 0, W]$$

- How can we know if the **distribution** of outcomes has changed? We need to be able to see (or at least estimate) two things:
  - the distribution of Y when X is manipulated (or "treated"): $Y|(X = 1, W) \sim F_1$
  - the distribution in the absence of manipulation – the *counterfactual* $Y|(X = 0, W) \sim F_0$
- Problem: we can't see **both** the outcome under treatment **and** the counterfactual outcome in the absence of treatment
- *How can we resolve the observability problem?* We need a way to infer the counterfactual for the units (people) that are treated

2. Causal Modeling

Possible ideas:

1. ("observational design"): calculate mean outcomes for people who are treated and those who are not in a given data
   - diff-in-diff, matched control group
2. ("pre-post design"): compare outcomes for people who are treated with their outcomes prior to treatment (i.e., the mean change in outcome after treatment):
   - treatment on the treated (TOT), event study
3. RCT - randomly assign treatment, calculate mean outcomes for T's and C's

2. Causal Modeling

4. Instrumental variables: there is some variable Z that shifts X but is unrelated to the unobserved determinants of Y:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + u \\ X &= \pi_0 + \pi_1 Z + \eta \\ Z &\perp u \end{aligned}$$

5. Regression Discontinuity (RD): X shifts discretely ("jumps") when some "running variable" Z passes a threshold. (e.g., eligible for Medicare at age 65, GPA threshold for declaring a major)

Basic Concepts: Population vs. Sample

- Y is a random variable (r.v.):
  - the randomness comes from the act of random sampling from a population distribution.
  - assume the pop distribution has mean μ, variance σ^2
- $\{y_1, y_2, ...y_n\}$ is a random sample of Y from the population
- Sample objects:
  - $\overline{y}_n = n^{-1} \sum_{i=1}^{n} y_i$ is the sample mean, which is a "statistic" (a r.v.)
  - $s_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (y_i - \overline{y}_n)^2$ is the sample variance (note $n-1$)
- Properties of the estimators:
  - $E[\overline{y}_n] = \mu$: the sample mean is an **unbiased** estimator of the pop. mean
  - $Var[\overline{y}_n] = Var[\frac{1}{n} \sum_{i=1}^{n} y_i] = \sigma^2 / n$.
  - $E[s_n^2] = \sigma^2$: the "d.f. corrected" sample var is unbiased for population variance σ^2 (Prove in class)

Convergence in Probability & LLN

- *Convergence in Probability*: $Z_1, Z_2, ..$is a sequence of random variables that *converges in probability* to b if for any $\varepsilon > 0$:

$$lim_{n\to\infty} P(|Z_n - b| < \varepsilon) = 1$$

  Write as $plimZ_n = b$ or $Z_n \to^p b$

- *Weak Law of Large Numbers (WLLN).* Suppose that $\{y_1, y_2, ..., y_n\}$ is a random sample from a pop with mean μ, variance σ^2(both finite). Then the sample mean $\bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$ (a r.v. itself) converges in probability to the population mean

$$plim\overline{y}_n = \mu \text{ or } \overline{y}_n \to^p \mu$$

  in other words, the sample mean is a **consistent** estimator for the population mean.

- *Continuous Mapping Theorem*: Suppose g is a continuous function defined on the same metric space as the random variable $Z_n's$. If $Z_n \to^p b$ and g is continuous at b, we have

$$g(Z_n) \to^p g(b)$$

Weak Law of Large Numbers

To prove WLLN $plim\bar{y}_n = \mu$, we first introduce two important results: Markov inequality; Chebyshev inequality:

1. *Markov*: if
   X is a positively-valued r.v., with $P(X > 0) = 1$, then for any $t > 0$:

$$P(X \geq t) \leq \frac{E[X]}{t} \tag{1}$$

   Proof:

$$E[X] = \int_0^\infty xf(x)dx = \underbrace{\int_0^t xf(x)dx}_{\geq 0} + \int_t^\infty xf(x)dx$$

$$\geq \int_t^\infty tf(x)dx = t \times Pr(X \geq t)$$

Weak Law of Large Numbers

2. *Chebychev*: If X is a random variable s.t. $Var[X] \in (0, \infty)$, then for any $t > 0$:

$$P(|X - E[X]| \geq t) \leq \frac{Var[X]}{t^2} \tag{2}$$

Proof (via Markov): consider r.v. $Y = (X - E[X])^2$. Note $E[Y] = Var[X]$. Using Markov, $\forall\, \tau > 0$,

$$P(Y \geq \tau) \leq \frac{E[Y]}{\tau}$$

$$\text{equiv. to } P(|X - E[X]| \geq \sqrt{\tau}) \leq \frac{E[Y]}{\tau} = \frac{var(X)}{\tau}$$

letting $\tau = t^2$, we have shown (2).

Weak Law of Large Numbers

- Now we are ready to prove WLLN: suppose the population distribution has finite (μ, σ^2), $plim \overline{y}_n = \mu$. That is,

$$lim_{n \to \infty}(|\overline{y}_n - \mu| < \epsilon) = 1$$

  Proof. Applying Chebychev to $\overline{y}_n$: for any $\epsilon > 0$,

$$P(|\overline{y}_n - \mu| \geq \varepsilon) \leq \frac{Var[\overline{y}_n]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$
$$\Rightarrow P(|\overline{y}_n - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}, \text{ where } \frac{\sigma^2}{n\varepsilon^2} \to 0 \text{ as } n \to \infty$$

  so $lim_{n \to \infty} P(|\overline{y}_n - \mu| < \varepsilon) = 1$.
- WLLN says that the distribution of the sample mean "collapses" to the point μ as the sample size gets bigger.

Convergence in Distribution & Central Limit Theorem

- *Convergence in Distribution*: $Z_1, Z_2, ..$is a sequence of real-valued random variables with cumulative distribution functions $F_1, F_2...$ (cdf $F_i(z) = P(Z_i \leq z)$). It is said to *converge in distribution* to a random variable Z with cdf F if

$$lim_{n \to \infty} F_n(z) = F(z)$$

for all $z \in \mathbb{R}$ at which F is continuous. Write as $Z_n \to^d Z$.

- The *Central Limit Theorem (CLT)*: Let $\{y_1, ...y_n\}$ be a random sample a population with mean μ, variance σ^2. The distribution of sample mean $\overline{y}_n$ collapses to a *normal distribution* at the rate $\sqrt{n}$:

$$lim_{n \to \infty} P\left(\frac{\sqrt{n}(\overline{y}_n - \mu)}{\sigma} \leq z\right) = \Phi(z),$$

where Φ is cdf for the standard normal $N(0,1)$. Write as $\sqrt{n}(\overline{y}_n - \mu)/\sigma \to^d N(0,1)$.

  - informally, write $\overline{y}_n \approx N(\mu, \sigma^2/n)$

Convergence in Distribution & Central Limit Theorem

- CLT says that the sample mean is "asymptotically normal", **regardless of the underlying population distribution** (as long as μ and σ^2 are finite).
- A key idea of statistics is that for a given n we can step back from the limit and still be "approximately" OK.
  - CLT remains true if, instead of scaling by population σ, we scale by sample estimate s_n:

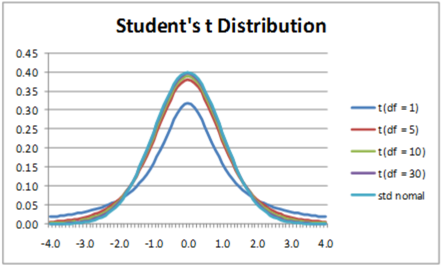
$$\frac{\sqrt{n}(\overline{y}_n - \mu)}{s_n} \text{ approximately } N(0,1)$$

  - The distribution when we use $s_n$instead of σ(unknown) to scale is known as "t-distribution":

$$\frac{\sqrt{n}(\overline{y}_n - \mu)}{s_n} \sim^d t_{n-1}$$

    where $t_{n-1}$is the t-distribution with $n-1$ degrees of freedom. For large n the t is very close to the standard normal ($t_n \to^d N(0,1)$). For smaller n the t distribution has fatter tails.

Convergence in Distribution & Central Limit Theorem



Slutsky Theorem

Slutsky Theorem has 2 parts. We have seen part 1 - contraction mapping.

1. *(Continuous Mapping)*: Suppose g is a continuous function defined on the same metric space as the random variable $Z_n's$. If $Z_n \rightarrow^p b$ and g is continuous at b, we have

$$g(Z_n) \rightarrow^p g(b)$$

2. Given a sequence of random variables $Z_n \rightarrow^d N(0, \sigma^2)$ and another sequence of r.v. $A_n \rightarrow^p \alpha$, we have

$$Z_n A_n \rightarrow^d \alpha \times N(0, \sigma^2) = N(0, \alpha^2 \sigma^2)$$

Confidence intervals.

- Suppose $Z \sim N(0, 1)$. Then we know Z is symmetrically distributed around 0 with a "bell curve" distribution.
- Define $z_p > 0$ as the real number such that $\Phi(z_p) = 1 - p$ (for $p < .5$). This is the point such that $P(Z > z_p) = p$.
- What is the symmetric interval (around 0) such that a standard normal falls in the interval with probability $1 - \alpha$? This is the interval $(-z_{\alpha/2}, z_{\alpha/2})$. Why? Because the probability of falling *above* $z_{\alpha/2}$ is $\alpha/2$, and by symmetry the probability of falling *below* $-z_{\alpha/2}$ is also $\alpha/2$. So with probability of being outside the interval is α.

Inference - Confidence Interval

- For $Z \sim N(0,1)$ $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$. Suppose we have obtained a random sample of some $Y's$ and formed the estimated mean and standard deviation. By the CLT $\sqrt{n}(\overline{y}_n - \mu)/s_n \approx N(0,1)$, so (approximately):

$$P(-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{y}_n - \mu)}{s_n} \leq z_{\alpha/2}) = 1 - \alpha$$

$$\Rightarrow P(\overline{y}_n - \frac{s_n z_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \overline{y}_n + \frac{s_n z_{\alpha/2}}{\sqrt{n}}) = 1 - \alpha$$

- This is interpreted as: if we kept repeating a sample of size n, $(1-\alpha)$ percent of the time the interval $\overline{Y}_n \pm \frac{s_n z_{\alpha/2}}{\sqrt{n}}$ would "capture" the true mean μ. This is called the $(1-\alpha)$ "confidence interval".
- Frequentist view: μ is a population parameter that is **fixed**. Can we interpret the CI at with with $1-\alpha$ probability μ takes a value between $\overline{Y}_n \pm \frac{s_n z_{\alpha/2}}{\sqrt{n}}$?

Univariate Regression

- Consider a normal linear model with a single independent variable:

$$Y = \beta_0 + \beta_1 X + U,\ U \sim N(0, \sigma^2) \tag{3}$$

  - In this model $U \sim N(0, \sigma^2)$ is orthogonal to $(1, X)$: $E[U] = E[UX] = 0$ (note $cov(X, U) = E[UX] - E[U]E[X] = 0$). In fact, the model assumes independence $E[U|X] = 0$, which is stronger than $u_i \perp x_i$!
  - Show (population regression):

$$\beta_0 = E[Y] - \beta_1 E[X] \tag{4}$$

$$\beta_1 = \frac{cov(Y, X)}{var(X)}$$

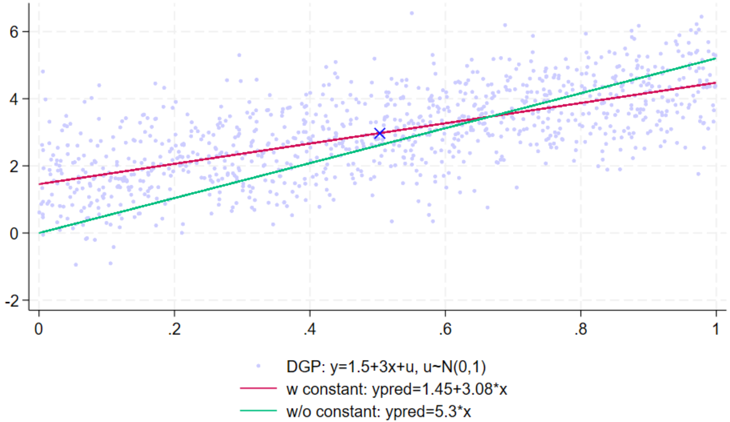
Univariate Regression

- Given a random sample $\{y_i, x_i\}_{i=1}^n$, we estimate an OLS regression of y_i on 1 and x_i:

$$
\begin{aligned}
(\hat{\beta}_0, \hat{\beta}_1) =& argmin_b \frac{1}{n} \sum_i (y_i - b_0 - b_1 x_i)^2 \qquad (5) \\
\rightarrow \hat{\beta}_0 =& \bar{y}_n - \hat{\beta}_1 \bar{x}_n \\
\hat{\beta}_1 =& \frac{\sum_i (y_i - \bar{y}) x_i}{\sum_i (x_i - \bar{x}) x_i} = \frac{\hat{cov}(x_i, y_i)}{\hat{var}(x_i)}
\end{aligned}
$$

- Discuss: what's the role of the constant? What if we a run a regression without a constant?

Univariate Regression



Inference - Consistency

We want to do inference on $(\hat{\beta}_0, \hat{\beta}_1)$: how are the estimates compared to the true parameter (β_0, β_1)?
Plug $y_i = \beta_0 + \beta_1 x_i + u_i$ into $\hat{\beta}_1$ and let $\bar{u} := \bar{y} - \beta_0 - \beta_1 \bar{x}$:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_i (y_i - \bar{y}) x_i}{\sum_i (x_i - \bar{x}) x_i} = \frac{\sum_i (\beta_0 + \beta_1 x_i + u_i - \beta_0 - \beta_1 \bar{x} - \bar{u}) x_i}{\sum_i (x_i - \bar{x}) x_i} \\
&= \beta_1 + \frac{\sum_i x_i (u_i - \bar{u})}{\sum_i (x_i - \bar{x}) x_i}
\end{aligned}
$$

- First, show consistency: $\hat{\beta}_0 \rightarrow^p \beta_0$, $\hat{\beta}_1 \rightarrow^p \beta_1$. By Weak LLN,

$$\frac{1}{n-1} \sum_i x_i (u_i - \bar{u}) \rightarrow^p cov(X, U) = 0$$

$$\frac{1}{n-1} \sum_i (x_i - \bar{x}) x_i \rightarrow^p var(X)$$

by CMP, $\hat{\beta}_1 = \beta_1 + \dfrac{\frac{1}{n-1} \sum_i x_i (u_i - \bar{u})}{\frac{1}{n-1} \sum_i (x_i - \bar{x}) x_i} \rightarrow^p \beta_1 + \dfrac{cov(X, U)}{var(X)} = \beta_1$

and $\hat{\beta}_0 \rightarrow^p E[Y] - \beta_1 E[X] = \beta_0$

Inference - Asymptotic Distribution

Second, derive the asymptotic distribution of $\hat{\beta}_1$ and find the 95% confidence interval for β_1.
Note for large sample, $n-1 \approx n$.

- By the Central Limit ThM,

$$\frac{1}{\sqrt{n}}\sum_i x_i(u_i-\bar{u}) = \frac{1}{\sqrt{n}}\sum_i (x_i-\bar{x})(u_i-\bar{u})$$

$$\rightarrow^d N(0, E[(X-E[X])U^2])\underbrace{= N(0,\sigma^2 var(X)))}_{(*)}$$

  (*) comes from the law of iterated expectations and $E[U^2|X]=\sigma^2$ under the assumption that $U \sim N(0,\sigma^2)$.

  - By WLLN, the denominator $\frac{1}{n-1}\sum_i (x_i-\bar{x})x_i \rightarrow^p var(X)$

- We can apply Slutsky Theorem:

$$\sqrt{n}(\hat{\beta}_1-\beta_1) = \frac{\frac{1}{\sqrt{n}}\sum_i x_i(u_i-\bar{u})}{\hat{var}(X)} \rightarrow^d N(0,\frac{\sigma^2}{var(X)})$$

OLS in Vector Notation

Suppose we are interested in the OLS regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \tag{6}$$

where $i = 1...N$ indexes elements of a sample. Here (x_{1i}, x_{2i}, y_i) are observed values of two *covariates* and our *outcome of interest* (y) for unit i. We can define the 3-row vectors x_i and β :

$$x_i = \begin{pmatrix} 1 \\ x_{1i} \\ x_{2i} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Using these vectors we can write the model in vector notation:

$$y_i = x_i' \beta + u_i \tag{7}$$

and the OLS estimator:

$$\hat{\beta} = argmin_b \sum_i (y_i - x_i' b)^2$$

Differentiate the dot product $x_i' b$ w.r.t. b?

OLS in Vector Notation

$$\frac{\partial(x_i'b)}{\partial b} = \begin{bmatrix} \frac{\partial(x_i'b)}{\partial b_1} \\ \frac{\partial(x_i'\beta)}{\partial b_2} \\ \cdots \\ \frac{\partial(x_i'\beta)}{\partial b_K} \end{bmatrix} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \cdots \\ x_{iK} \end{bmatrix} = x_i$$

So

$$\frac{\partial \sum_i (y_i - x_i'b)^2}{\partial b} = -2\sum_i x_i\,(y_i - x_i'b) = 0$$

$$\Rightarrow \hat{\beta} = \left(\sum_i x_i x_i'\right)^{-1} \left(\sum_i x_i y_i\right)$$