**POLSCI 630 Introduction to Regression Methods** Spring 2026

# Lecture 5 Causal Inference and Regression

*Jiawei Fu, Duke University* *Scribe: Jiawei Fu*

## 1 Overview

In previous lectures, we learned regression from the traditional econometrics perspective. Based on those techniques, in this lecture we examine regression from the causal inference paradigm. The estimand is not a regression coefficient; instead, we target the average causal effect defined by the potential outcomes. Throughout this lecture, we assume the SUTVA assumption holds.

## 2 Identification

In causal inference, our estimand is a causal effect, for example, the average treatment effect. We mentioned in Lecture 1 that ATE, $\mathbb{E}[Y_i(1) - Y_i(0)]$, is identified as $\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]$ under two key assumptions: $\{Y_i(1), Y_i(0)\} \perp Z_i$ and positivity $0 < Pr(Z_i = 1) < 1$. To better understand this identification assumption, let us see why the naive population difference-in-means does not identify the causal effect.

We decompose the population difference-in-means:

$$\begin{aligned}\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] &= \mathbb{E}[Y_i(1)|Z_i = 1] - \mathbb{E}[Y_i(0)|Z_i = 0] \\ &= \underbrace{(\mathbb{E}[Y_i(1)|Z_i = 1] - \mathbb{E}[Y_i(0)|Z_i = 1])}_{ATT} + \underbrace{(\mathbb{E}[Y_i(0)|Z_i = 1] - \mathbb{E}[Y_i(0)|Z_i = 0])}_{\text{Selection Bias}}\end{aligned}$$

The first term is the average treatment effect on the treated (ATT), $\mathbb{E}[Y_i(1) - Y_i(0)|Z_i = 1]$. It is a well-defined causal quantity. A causal quantity is defined on the same unit. ATT is more relevant in some applications; for example, it may not be of interest to know the causal effect of a policy on those who will never be targeted.

Without ignorability $\{Y_i(1), Y_i(0)\} \perp Z_i$, the population difference-in-means contains the second term, called selection bias. It will not be zero if treated units "select" into treatment. This selection implies that their counterfactual $Y_i(0)|Z_i = 1$ will typically be different from those who do not receive treatment, $Y_i(0)|Z_i = 0$.

With the identification assumption $\{Y_i(1), Y_i(0)\} \perp Z_i$ and $0 < Pr(Z_i = 1) < 1$, the selection bias is 0:

$$\mathbb{E}[Y_i(0)|Z_i = 1] - \mathbb{E}[Y_i(0)|Z_i = 0] = \mathbb{E}[Y_i(0)] - \mathbb{E}[Y_i(0)] = 0,$$

and ATT is equal to ATE:

$$\mathbb{E}[Y_i(1)|Z_i = 1] - \mathbb{E}[Y_i(0)|Z_i = 1] = \mathbb{E}[Y_i(1) - Y_i(0)]$$

# 3 Experiment

For any empirical study that aims to estimate a causal effect, the gold standard is an experiment. We will study how to identify and estimate causal effects with experiments first. Then, for observational studies, our main job is to try to "design" the study so as to mimic the experiment, so that we can identify causal effects as in an experiment.

In econometrics, we assume the population is infinite and the sample is i.i.d. from this infinite population. However, in causal inference, we have two frameworks: finite population and superpopulation. In the finite population framework, we adopt design-based inference. It is different from model-based inference in econometrics.

## 3.1 Experiment: Finite Population

We start with the finite population framework. We assume the population has $n$ units, and we are only interested in the ATE for those $n$ units. Therefore, the estimand is defined only on those $n$ units: $\tau = \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)]$. Potential outcomes for each unit are fixed, so they are not random. The only randomness comes from $Z$ via the experimental design. Therefore, the inference is called design-based inference.

The estimator is still the natural difference-in-means estimator, $\frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}(1-Z_i)Y_i$, where $n_1$ and $n_0$ are the number of units in the treatment and control groups. We will show it is unbiased under random assignment.

$$
\begin{aligned}
\mathbb{E}\left[\hat{\tau}\right] &= \mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i(1) - \frac{1}{n_0}(1-Z_i)Y_i(0)\right] \\
&= \frac{1}{n_1}\sum_{i=1}^{n}\mathbb{E}\left[Z_i\right]Y_i(1) - \frac{1}{n_0}\mathbb{E}\left[(1-Z_i)\right]Y_i(0) \\
&= \frac{1}{n_1}\sum_{i=1}^{n}\frac{n_1}{n}Y_i(1) - \frac{1}{n_0}\frac{n_0}{n}Y_i(0) \\
&= \frac{1}{n}\sum_{i=1}^{n}Y_i(1) - \frac{1}{n}\sum_{i=1}^{n}Y_i(0) \\
&= \tau
\end{aligned}
$$

Note that only $Z_i$ is random, and we use the fact that $Y_i = Z_i Y_i(1) + (1-Z_i)Y_i(0)$.

What about the variance? It is

$$
Var(\hat{\tau}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_1} - \frac{S^2(\tau)}{n}
$$

where $S^2(1)$ is the population variance, $S^2(1) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(1) - \overline{Y}_i(1))^2$, and $S^2(\tau)$ is the variance of the individual causal effect, $S^2(\tau) = \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \tau)^2$. We need to use the sample to estimate them.

For $S^2(1)$, we can use the sample analog, $\hat{S}^2(1) = \frac{1}{n_1-1}\sum_{i=1}^{n} Z_i(Y_i - \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i)^2$. This is unbiased. However, we never observe $\tau_i$, so we cannot estimate the variance of $\tau_i$. Fortunately, we

can ignore the last term and obtain the conservative estimator

$$\hat{V}(\hat{\tau}) = \frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}$$

This is called the "Neyman variance estimator." It is conservative in the sense that $\mathbb{E}\left[\hat{V}(\hat{\tau})\right] - V(\hat{\tau}) = \frac{S^2(\tau)}{n} \geq 0$. The bias is zero under a constant treatment effect.

**Remark 1.** *A variant of the finite population setup is to still have a finite population with $n$ units, but draw a random sample from this finite population. The resulting sample size is $N$. The estimand is still $\tau = \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)]$, but the estimator can only use information from the sample: $\frac{1}{N_1}\sum_{i=1}^{N} Z_i Y_i - \frac{1}{N_0}(1 - Z_i)Y_i$. The variance of the estimator still has a term that cannot be estimated.*

### 3.2 Experiment: Super-population

In the super-population framework, the $n$ observed units are viewed as a random sample from an essentially infinite population. In this framework, the estimand is $\tau_{PATE} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$. Now, potential outcomes are treated as random, as in traditional econometrics.

Again, under $\{Y_i(1), Y_i(0)\} \perp Z_i$, $\tau_{PATE} = \mathbb{E}\left[Y|Z=1\right] - E[Y|Z=0]$ is identified. The corresponding plug-in estimator is still the difference-in-means estimator

$$\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}(1 - Z_i)Y_i.$$

Conditional on the sample $O_i = \{Y_i(0), Y_i(1)\}_{i=1}^{n}$, the only randomness comes from $Z_i$. This reduces to the finite population framework. In the super-population framework, we now call the ATE on the sample the Sample ATE, SATE.

$$\mathbb{E}\left[\hat{\tau}|O_i\right] = \tau_{SATE}.$$

and SATE is an unbiased estimator of PATE:

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}\right] &= \mathbb{E}\left[\mathbb{E}\left[\hat{\tau}|O_i\right]\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} Y_i(0)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y_i(1)\right] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y_i(0)\right] \\
&= \tau_{PATE}
\end{aligned}$$

For the variance, one can apply the formula we used in the lecture on finite-sample properties of OLS, $Var(\hat{\tau}) = \mathbb{E}\left[Var(\hat{\tau}|O_i)\right] + Var[\mathbb{E}\left(\hat{\tau}|O_i\right)]$, to show that it is $Var(\hat{\tau}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_1}$. The last term disappears. Therefore, we can estimate it with $\hat{V}(\hat{\tau})$, and it is unbiased in the super-population framework.

As you may expect, we can also show the estimator is asymptotically normal. From this, we can conduct hypothesis testing, etc.

### 3.3 Experiment: Regression

Can we use regression to estimate ATE instead? Suppose we just run BLP $Y = \beta_0 + Z\beta_1 + e$. Let $p = P(Z_i = 1)$. In Homework 1, you worked out that the coefficient of $\beta_1$ is

$$\begin{aligned}
\beta_1 &= \frac{Cov(Y_i, Z_i)}{Var(Z_i)} \\
&= \frac{\mathbb{E}[Y_i Z_i] - \mathbb{E}[Y_i]\mathbb{E}[Z_i]}{p(1-p)} \\
&= \frac{\mathbb{E}[Y_i|Z_i = 1]p - (\mathbb{E}[Y_i|Z_i = 1]p + \mathbb{E}[Y_i|Z_i = 0](1-p))p}{p(1-p)} \\
&= \frac{p(1-p)(\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0])}{p(1-p)} \\
&= \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]
\end{aligned}$$

Great! It shows that $\beta_1 = \tau$ is the population difference-in-means. Under an experiment, $\{Y_i(1), Y_i(0)\} \perp Z_i$, we know ATE is identified as this population difference-in-means. Therefore, the regression parameter is equal to the causal parameter, ATE, for this binary treatment case. And, apparently, the OLS estimator is equivalent to the difference-in-means estimator: $\hat{\beta}_1 = \hat{\tau}$.

Another way to look at this is that regression can be interpreted as follows:

$$Y_i = \underbrace{\mathbb{E}[Y_i(0)]}_{\beta_0} + \underbrace{Z_i\mathbb{E}[Y_i(1) - Y_i(0)]}_{Z_i\beta_1} + \underbrace{Z_i(Y_i(1) - \mathbb{E}[Y_i(1)]) + (1 - Z_i)(Y_i(0) - \mathbb{E}[Y_i(0)])}_{e_i}$$

$$\text{or } Y_i = \underbrace{\mathbb{E}[Y_i(0)]}_{\beta_0} + \underbrace{Z_i\mathbb{E}[Y_i(1) - Y_i(0)]}_{Z_i\beta_1} + \underbrace{(Y_i(0) - \mathbb{E}[Y_i(0)]) + Z_i[(Y_i(1) - Y_i(0)) - (\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)])]}_{e_i}$$

Under random assignment, we can show $\mathbb{E}[e_i|Z_i] = 0$. Look at the first decomposition. You will work on the second decomposition in the homework.

$$\begin{aligned}
\mathbb{E}[e_i|Z_i = 1] &= \mathbb{E}[Y_i(1) - \mathbb{E}[Y_i(1)|Z_i = 1] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(1)] = 0 \\
\mathbb{E}[e_i|Z_i = 0] &= \mathbb{E}[Y_i(0) - \mathbb{E}[Y_i(0)|Z_i = 0] = 0
\end{aligned}$$

In the second equality, we use random assignment $\{Y_i(1), Y_i(0)\} \perp Z_i$ to remove the condition: $\mathbb{E}[Y_i(1)|Z_i = 1] = \mathbb{E}[Y_i(1)]$. Therefore, under random assignment, $\beta_1 = \tau$, and the OLS estimator $\hat{\beta}_1$ is unbiased and consistent. We conclude that the ATE identification assumption $\{Y_i(1), Y_i(0)\} \perp Z_i$ justifies CEF, BLP, and the structural model.

What about inference? It can be shown that

$$\begin{aligned}
\hat{V}^0 &= \frac{n(n_1 - 1)}{(n-2)n_1}\frac{\hat{S}^2(1)}{n_0} + \frac{n(n_0 - 1)}{(n-2)n_0}\frac{\hat{S}^2(0)}{n_1} \\
&\approx \frac{\hat{S}^2(1)}{n_0} + \frac{\hat{S}^2(0)}{n_1}
\end{aligned}$$

and

$$\hat{V}^{HC0} = \frac{n_1 - 1}{n_1}\frac{\hat{S}^2(1)}{n_1} + \frac{n_0 - 1}{n_0}\frac{\hat{S}^2(0)}{n_0}$$

and

$$\hat{V}^{HC0} = \frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}$$

by noting that $1 - h_{ii} = 1 - \frac{1}{n_1} = \frac{n_1 - 1}{n_1}$ when $Z_i = 1$ and $h_{ii} = \frac{1}{n_0} = \frac{n_0 - 1}{n_0}$ when $Z_i = 0$. We observe that the denominator in $\hat{V}^0$ is reversed. This suggests it is not a good estimator. $HC0$ is slightly smaller, while $HC2$ is exactly the same.

### 3.4 Covariate Adjustment

In regression, we often add covariates. For example, consider $Y_i = \beta_0 + \beta_1^F Z_i + \beta_2 X_i + e_i$.

Freedman (2008) documents some negative results. First, $\hat{\beta}_1$ is biased. Second, the asymptotic variance of $\hat{\beta}_1$ may be even larger than that of $\hat{\tau}$ when $n_1 \neq n_0$. Third, conventional OLS standard errors (homoskedastic and HC0) are invalid.

However, his student Lin (2013) shows the bias goes to zero asymptotically. Moreover, suppose $X_i$ are centered (so that $\sum_{i=1} X_i = 0$). If we run

$$Y_i = \beta_0 + \beta_1^L Z_i + \beta_2 X_i + \beta_3 Z_i X_i + e_i$$

instead, then adding covariates does not reduce asymptotic efficiency (the variance is smaller than DIM).

**Remark 2** (Interaction)**.** *Consider a CEF with quadratic and interaction terms:* $m(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$*. It is still a linear CEF because it is a linear function of the coefficients. At the same time, it has nonlinear effects because it is nonlinear in the underlying variables* $X_1$ *and* $X_2$*. To see this, for example,*

$$\frac{\partial}{\partial x_2} m(X_1, X_2) = \beta_2 + 2\beta_4 X_2 + \beta_5 X_1$$

*We call* $\beta_5$ *the interaction effect. The effect of* $X_2$ *on CE depends on* $X_1$*.*

In general, Lin (2013) shows

1. $\beta_1^L$ is consistent and asymptotically normal, as are $\beta_1^F$ and DIM.

2. Asymptotically, $\beta_1^L$ does not hurt asymptotic precision, even when the regression model is incorrect.

3. Sandwich variance estimator for $\beta_1^F$ and $\beta_1^L$ is greater than or equal to the true asymptotic variance of $\hat{\tau}$.

When the number of covariates is small compared with the sample size and the covariates do not contain outliers, the variants of the EHW standard error perform similarly to the original one. When the number of covariates is large compared with the sample size or the covariates contain outliers, the variants can outperform the original one. In those cases, we should use HC3. See more in Ding (2024).

### 3.5 Stratified/Block randomization

Completed randomized experiments may result in undesirable covariate imbalance. Such imbalance deteriorates the quality of the experiment, making it difficult to interpret the results since the difference in outcomes may be attributed to the treatment or the covariate imbalance.

One solution is stratified or block randomization. Instead of randomly assigning all observations at once, the researcher divides the observations into more homogeneous blocks and randomly allocates observations within each block. Blocking ensures that all variables used to create strata will be balanced and tends to improve the precision with which the ATE is estimated. For this reason, experimenters tend to follow the dictum "Block what you can, and randomize what you cannot".

With blocking, it is equivalent to conducting $K$ independent completed randomized experiments within the strata. Let $\tau_k$ be the ATE within stratum $k$, and let $\pi_k$ be the proportion of observations in stratum $k$. Now, intuitively,

$$\tau = \sum_{k=1}^{K} \pi_k \tau_k$$

## 4 Observational Study under CIA

For observational studies, we should always think about the ideal experiment that the observational study approximates. Then it becomes clear how to design the observational study. You will be clear about what the treatment is and whether it can be treated as if random. We will learn many such methods to "randomize" the treatment in observational studies later. They are often called natural or quasi-experiments.

Most of the time, we need to condition on some covariates so that (strong) ignorability holds: $(Y_i(0), Y_i(1)) \perp Z_i | X_i$ and $0 < \mathbb{P}[Z_i = 1 | X_i = x] < 1$. Then, ATE is identified as $\tau = \mathbb{E}[\mathbb{E}[Y_i | X_i, Z_i = 1]] - \mathbb{E}[\mathbb{E}[Y_i | X_i, Z_i = 0]]$, as we showed in Lecture 1.

**Example 1.** *Consider the data-generating process (DGP), $Y_i(1) = f_1(X_i, \epsilon_1)$, $Y_i(0) = f_1(X_i, \epsilon_0)$, $Z_i = I[g(X_i, \epsilon_i) > 0]$, where $\epsilon$ is random noise. In this DGP, ignorability holds conditional on $X_i$.*

**Example 2.** *Consider the data-generating process (DGP), $Y_i(1) = f_1(X_i, U_i, \epsilon_i)$, $Y_i(0) = f_1(X_i, U_i, \epsilon_0)$, $Z_i = I(g(X_i, U_i, \epsilon_i) > 0)$. Because the same random variable $U_i$ correlates $Z_i$ and potential outcomes, ignorability does not hold simply conditional on $X_i$.*

How do we understand conditioning on $X$? We cannot literally condition on $X$; we only have observed data. Suppose $X$ takes two values, 0 and 1. Conditioning means that we only look at strata defined by $X$. So, $(Y_i(0), Y_i(1)) \perp Z_i | X_i$ means that if we look only at the population with $X_i = 1$, it is a completed randomized experiment. Similarly, if we look only at the population with $X_i = 0$, there is an experiment.

### 4.1 Control Variables

Strong ignorability requires us to choose some covariates to condition on. There are many candidates. Which ones should we condition on? In this subsection, say we want to study the causal effect of treatment $X$ on outcome $Y$.

1. Common causes of the treatment and outcome should be controlled.

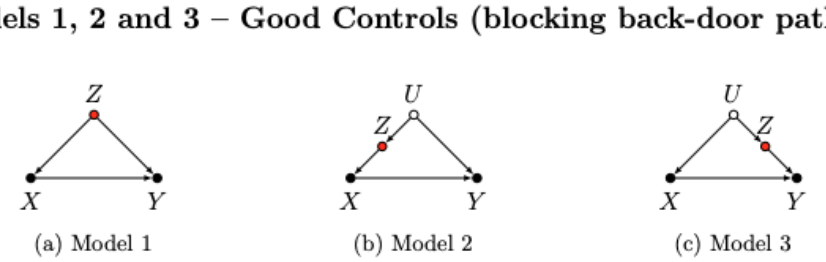


Figure 1: From Cinelli et al. (2024)

In model 1, $Z$ is the common cause; we call it a confounder. If we do not control it, then the relationship between treatment and outcome is confounded by $Z$. In models 2 and 3, we can cut the confounding effect from $U$ by controlling $Z$.

2. Mediator should not be controlled.

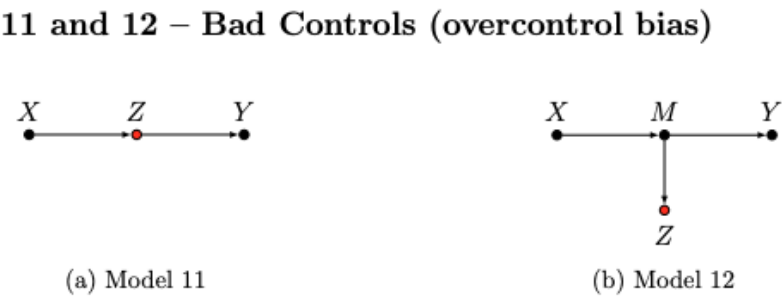


Figure 2: From Cinelli et al. (2024)

A mediator is affected by treatment $X$ and then affects outcome $Y$. We do not control the mediator because it captures some effect from treatment to outcome. If we control it, we cannot identify the "total" causal effect. In model 11, it is $Z$ and in model 12 it is $M$. Also, in model 12, we should not control $Z$ because controlling it is equivalent to controlling $M$.

3. A collider is a variable that is affected by two variables. If we control it, we introduce correlation among those two variables.

Consider $X$ as ability and $Y$ as effort, and $Z$ as whether the student is admitted into the PhD program, and suppose $X$ and $Y$ are independent. So $Z$ is the collider. If we condition on $Z = 1$ for those admitted, then if we find $X$ is low, it generally implies $Y$ is higher. Similarly, if we find $X$ is high, it generally implies $Y$ is low. Therefore, conditioning on a collider generates correlation between two otherwise independent variables.

In model 17, we do not want to control $Z$ because it will introduce another correlation between $X$ and $Y$ beyond the causal effect. Similarly, we do not want to control the collider $Z$ in model 16.

4. Neutral control: Controlling or not controlling does not affect identification, but may affect estimation precision. We know from OLS that it will decrease the variance of the OLS estimator.

**Models 16 and 17 – Bad Controls (selection bias)**


(a) Model 16 (b) Model 17

Figure 3: From Cinelli et al. (2024)

**Model 8 – Neutral Control (possibly good for precision)**

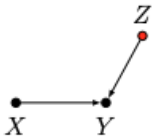


Figure 4: From Cinelli et al. (2024)

In model 8, we do want to control $Z$ because it decreases the variation within $Y$.

**Model 9 – Neutral Control (possibly bad for precision)**

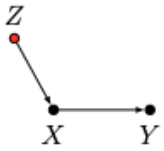


Figure 5: From Cinelli et al. (2024)

In model 9, we do not want to control $Z$ because it absorbs some variation in treatment. We know from OLS that it will increase the variance of the OLS estimator.

## 4.2 Regression

Let us focus on ATT now. $\tau_{att} = \mathbb{E}\left[Y_i(1) - Y_i(0) | Z_i = 1\right]$. It is identified under the weaker condition $Y_i(0) \perp Z_i | X_i$.

$$
\begin{aligned}
\tau_{att} &= \mathbb{E}\left[Y_i(1) - Y_i(0) | Z_i = 1\right] \\
&= \mathbb{E}_{X|Z=1}[\mathbb{E}\left[Y_i(1) - Y_i(0) | Z_i = 1, X_i\right]] \\
&= \mathbb{E}_{X|Z=1}[\mathbb{E}\left[Y_i(1) | Z_i = 1, X_i\right] - \mathbb{E}\left[Y_i(0) | Z_i = 1, X_i\right]] \\
&= \mathbb{E}_{X|Z=1}[\mathbb{E}\left[Y_i(1) | Z_i = 1, X_i\right] - \mathbb{E}\left[Y_i(0) | Z_i = 0, X_i\right]]
\end{aligned}
$$

The fourth line is because $Y_i(0) \perp Z_i | X_i$. Let $\delta_X = \mathbb{E}\left[Y_i | Z_i = 1, X_i\right] - \mathbb{E}\left[Y_i | Z_i = 0, X_i\right]$ denote the conditional difference-in-means. Then $\tau_{att} = \mathbb{E}_{X|Z=1}[\delta_X] = \mathbb{E}\left[\delta_X | Z_i = 1\right]$.

If $X$ is discrete and takes $k$ possible values, $\tau_{att} = \sum_x \delta_x \mathbb{P}[X_i = x | Z_i = 1]$. Because $\mathbb{P}[X_i = x | Z_i = 1] = \frac{\mathbb{P}[Z_i=1|X_i=x]\mathbb{P}[X_i=x]}{\mathbb{P}[Z_i=1]}$, $\tau_{att}$ can also be written as

$$
\tau_{att} = \frac{\sum_x \delta_x \mathbb{P}[Z_i = 1 | X_i = x]\mathbb{P}[X_i = x]}{\sum_x \mathbb{P}[Z_i = 1 | X_i = x]\mathbb{P}[X_i = x]}
$$

It is a weighted average of effects within each stratum, and weights are proportional to the probability of treatment at each value of the covariates. Note also that we can just as easily construct the unconditional average treatment effect,

$$
\tau_{ate} = \sum_x \delta_X \mathbb{P}[X_i = x]
$$

**Remark 3** (Saturated regression)**.** *When all regressors take a finite set of values, it turns out the CEF can be written as a linear function of regressors. This is great because we do not need to assume linearity.*

*Consider a simple case where $X_1$ is binary. In this case, CE takes only two distinct values $\mathbb{E}\left[Y | X_1 = 1\right] = \mu_1$ and $\mathbb{E}\left[Y | X_1 = 0\right] = \mu_0$. Given this, $\mathbb{E}\left[Y | X_1\right] = \beta_0 + \beta_1 X_1$, where $\beta_1 = \mu_1 - \mu_0$ and $\beta_0 = \mu_0$.*

*Now suppose we have two dummy variables $X_1$ and $X_2$. CE takes at most four possible values, $\mu_{00}, \mu_{01}, \mu_{10}, \mu_{11}$. In this case, we can write the CE as a linear function of $X_1$, $X_2$, and their product $X_1 X_2$:*

$$
\mathbb{E}\left[Y | X_1, X_2\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2
$$

*One can easily find that each coefficient can be written as a function of those four values. We can generalize this to more cases.*

Now, let us consider what happens if we run the following regression:

$$
Y_i = \beta_R Z_i + I[X_i = x_1]\beta_1 + ... + I[X_i = x_k]\beta_k + e_i
$$

There is no intercept to avoid multicollinearity. We want to ask whether $\beta_R = \tau_{att}$?

Angrist & Pischke (2009) shows that $\beta_R = \frac{\mathbb{E}\left[\sigma_Z^2(X_i)\right]\delta_X}{\mathbb{E}\left[\sigma_Z^2(X_i)\right]}$, where $\sigma_Z^2(X_i) = \mathbb{E}\left[(Z_i - \mathbb{E}\left[Z_i | X_i\right])^2 | X_i\right]$ is the conditional variance of $Z_i$ given $X_i$. Thus, the regression coefficient is a kind of weighted average of the strata averages $\delta_x$, but the weights depend on treatment variance rather than the proportion of the population in each stratum, $\mathbb{P}[X_i = x | Z_i = 1]$. Because $Z_i$ is binary, $\sigma_Z^2(X_i) = \mathbb{P}(Z_i = 1 | X_i)(1 - \mathbb{P}(Z_i = 1 | X_i))$, therefore,

$$
\beta_R = \frac{\sum_x \delta_x [\mathbb{P}(Z_i = 1 | X_i)(1 - \mathbb{P}(Z_i = 1 | X_i))]\mathbb{P}[X_i = x]}{\sum_x [\mathbb{P}(Z_i = 1 | X_i)(1 - \mathbb{P}(Z_i = 1 | X_i))]\mathbb{P}[X_i = x]}
$$

Figure 1: *On the left, the shading shows countries in the nominal sample for Jensen (2003)'s estimate of the effects of regime type on FDI. On the right, darker shading indicates that a country contributes more to the effective sample, based on the panel specification used in estimation.*

Figure 6: From Aronow & Samii (2016)

We can easily see that the weight is different from the weight in $\tau_{att}$. Aronow & Samii (2016) shows that for arbitrary $Z_i$ and $X_i$,

$$\hat{\beta}_R \to \frac{\mathbb{E}[w_i \tau_i]}{\mathbb{E} w_i}$$

where $w_i = (Z_i - \mathbb{E}[Z_i|X_i])^2$, in which case $\mathbb{E}[w_i|X_i] = Var[Z_i|X_i]$. This means that even with a representative sample, regression estimates may not aggregate effects in a representative manner. Regression estimates are "local" to an effective sample.

The point of this derivation is that the treatment-on-the-treated estimand puts the most weight on covariate cells containing those who are most likely to be treated. In contrast, regression puts the most weight on covariate cells where the conditional variance of treatment status is largest. As a rule, this variance is maximized when $\mathbb{P}[Z_i = 1|X_i = x] = \frac{1}{2}$, in other words, for cells where there are equal numbers of treated and control observations. A simple way to interpret the result is that more weight goes to units whose treatment values ($Z_i$) are not well explained by the covariates ($X_i$).

## 5 Final remarks

Causal inference does not need regression. However, regression is still useful for estimation. For example, to estimate $\tau = \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i = 1]] - \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i = 0]]$, there are many conditional expectations that can be estimated by a regression model. You will learn how to estimate the causal effect in my computational methods course, where we will learn modern estimators like de-biased/double machine learning and double-robust estimators.

## References


Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Aronow, P. M. & Samii, C. (2016). Does regression produce representative estimates of causal effects? *American Journal of Political Science*, 60(1), 250–267.

Cinelli, C., Forney, A., & Pearl, J. (2024). A crash course in good and bad controls. *Sociological Methods & Research*, 53(3), 1071–1104.

Ding, P. (2024). *A first course in causal inference*. Chapman and Hall/CRC.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2), 180–193.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, (pp. 295–318).