Week 2: Causal Inference

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2020

Outline

- Correlation and causation
- Potential outcomes and counterfactuals
- Defining causal effects
- The fundamental problem of causal inference
- Solving the fundamental problem of causal inference:
    a) Randomization
    b) Statistical adjustment
    c) Other methods
- The ignorable treatment assignment assumption
- Stable Unit Treatment Value Assumption (SUTVA)
- Assignment mechanism

Big picture

- We are going to review the basic framework for understanding **causal inference**
- This is a fairly new area of research, although some of the statistical methods have been used for over a century
- The "new" part is the development of a mathematical notation and framework to understand and **define** causal effects
- This new framework has many advantages over the traditional way of understanding causality. For me, the biggest advantage is that we can talk about causal effects without having to use a specific statistical model (design versus estimation)
- In econometrics, the usual way of understanding causal inference is linked to the linear model (OLS, "general" linear model) – the **zero conditional mean** assumption in Wooldridge: whether the additive error term in the linear model, ϵ_i, is correlated with the variable of interest (treatment). The problem is that when you move to circumstances in which a linear model is not the best tool (say, outcome is 1/0 or a count), things are confusing because there is no additive error term in these models

Big Picture II - Readings

- As we will see, there are reasons to even prefer a **nonparametric** model rather than a linear/OLS model to estimate a conditional expectation function $E[y_i|\boldsymbol{x}_i]$
- The main ideas of this new framework have been refined recently (last 15 years?), but there is still no common notation or language
- **For the economists in the room**: please read Imbens and Wooldridge (2009) "Recent developments in the econometrics of program evaluation" for a road map. Wooldridge (2010) book Chapter 21 has an interesting take. That chapter is the framework of Stata's new(ish) commands `teffects` and `eteffects`
- Angrist and Pischke (2009) also provide a bridge from the statistical literature on causal effects to econometrics, so we will use their notation, although they focus too much on the linear/OLS model. Imbens and Rubin (2015) is a better introduction to these topics (on Canvas)
- Note that the economics examples are mostly from labor economics. Labor economics is the field where econ PhD students end up if they want to focus on empirical methods

Big picture III

- You have all heard that **correlation (association) does not imply causation**
- Causal inference is about understanding under which circumstances correlation (association) **does imply** causation
- It's obviously a fundamental question since we want to understand causal effects when doing research and when using statistical models
- It's fundamental in **health services research** and **health economics**
- But remember that we also use statistical models as **descriptive tools** and for **prediction**. Make sure you understand the difference

Basics concepts

- Causality is linked to a **manipulation** (treatment, intervention, action, strategy) that is applied to a **unit**
- A unit could be a person, firm, hospital, country, county, classroom, etc
- Think of it as the **"thing" that received the action** or was manipulated
- The unit could have been exposed to an **alternative** action
- For **simplicity**, only two possibilities: receiving or not receiving the action or treatment (**active** versus **control** treatment in Imbens and Rubin, 2015)
- This is the usual simplification to understand key concepts, but there could be multiple treatment levels that could even be continuous like medication doses or monetary incentives
- A unit either receiving or not receiving a treatment is linked to a **potential outcome**

Potential outcomes for units

- The "potential" part refers to the idea that **only one outcome is realized** after the intervention; the other is, well, potential
- (Dictionary definition: Potential: having or showing the capacity to become or develop into something in the future)
- **Before** the intervention, there are **two potential outcomes**. **Only one is realized** after the action is conducted
- Example: a person may or may not receive a job training program if unemployed. The outcome could be income one year after the program – income if the person participated in the program and the income if the person didn't participate in the program. The outcome could be binary: getting a job or not
- **Jargon alert**: economists like to use a priori, a posteriori, ex ante, ex post instead or before and after

Notation

- The **observed** or **realized** outcome for a unit i is Y_i
- We will denote the **treatment** received by a unit i as D_i, which could be 0 or 1, so $D_i \in \{0, 1\}$
- The other common notation is for treatment to be W; Stata manual uses t for treatment
- We have then two **potential** outcomes: Y_{1i} if the unit received treatment ($D_i = 1$) and Y_{0i} if not ($D_i = 0$)
- Other common notation for *potential* outcomes [control]: Y_i^0 or $Y_i(0)$; Stata uses y_0. **Don't get confused with notation**
- It's also helpful to keep the index i. We will extend this notation when the structure of the data changes. Y_i is different than, say, Y_{its}. The latter could be the outcome for person i at time t in state s

Definition of causal effects

- The **causal effect** effect of receiving treatment for a unit i is a **comparison of potential outcomes**
- We could compare $Y_{1i} - Y_{0i}$, the difference between outcomes when units are treated versus not
- Or we could measure the effect as a relative measure: $\frac{Y_{1i}}{Y_{0i}}$. Or any other way, but the idea is that the causal effect of the treatment a unit received D_i is a **comparison of potential outcomes**
- Note that the definition of causal effects **does not depend on the actual treatment or action taken**

Causal effects

- This idea seems a bit confusing at first but it's a deep concept. You just need to get used to it
- A person i could receive an antidepressant or not. The outcome Y_i could be being depression free at two months (a binary outcome)
- The causal effect of receiving the antidepressant is defined as the comparison of the outcome for the person with the antidepressant and without the antidepressant, potential outcomes that do not depend on what treatment the person gets
- We could define the outcome as the probability of being depression free at two months, so the causal effect could be $P(Y_{1i}) - P(Y_{0i})$
- Again, we could use other ways of measuring effects. For example, the odds-ratio (which is hard to interpret anyway)

$$\frac{\frac{P(Y_{1i})}{1-P(Y_{1i})}}{\frac{P(Y_{0i})}{1-P(Y_{0i})}}$$

Causal effects

- This way of thinking about causal effects **matches everyday thinking**
- The causal effect of the next election is the comparison of an outcome, say, the status of the pandemic next year if the country reelects the current president or not
- The causal effect of studying hard for this class is a comparison of two potential outcomes: your grade if you study hard and your grade if you don't. **Only one will be observed** because you will study hard or not
- In the movie a It's a Wonderful Life, Clarence the angel shows George Bailey what would have happened had he not been born. In essence showing Bailey the causal effect of him being alive
- These ideas could be extended to **multi-value treatments**. Rather than study hard or not, we could have multiple levels of effort that could be measured continuously

The fundamental problem of causal inference

- We now have a **definition of causal** effects but we also have a **BIG PROBLEM**
- The challenge in causal inference is that **we do not observe both potential outcomes**; we only observe one. (In the stats literature this is called the "fundamental problem of causal inference." In the economics literature, it's called the fundamental problem of *program evaluation*)
- Note that in this framework, the *same unit receiving a treatment at a different time is a different unit*
- The non-observable or not-realized outocome is called the **counterfactual** (Dictionary: relating to or expressing what has not happened or is not the case)
- Said another way, in real life, **the problem is that there is no Clarence the angel to tell us the counterfactual**
- This way of understanding causal effects (potential outcomes, counterfactuals) is now called the **Rubin causal model** (see Imbens and Wooldridge, 2009 for some history, also Imbens and Rubin, 2015, Chapter 2)

Counterfactuals

- Going back to our notation, we can write the **observed outcome** Y_i in terms of the **potential outcome** as

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

- If $D_i = 1$ then the observed outcome is $Y_i = Y_{1i}$. If $D_i = 0$ then the observed outcome is $Y_i = Y_{0i}$
- A unit either receives the treatment or not, never both
- So if a unit receives the treatment the **observed** outcome is Y_{1i} and **counterfactual** outcome is Y_{0i}
- **Important (!!)**: We can think of causal inference as a **PREDICTION problem**. How could we predict the counterfactual given that we never observe it?

Digression for home study

- **You have to read the assigned readings for this week**. This is an area that requires a lot of home study
- In the words of my Jesuit priest high school director: To be a good student you need two things: a big behind and a head. The big behind is more important so you can sit down and study for hours. Intelligence is overrated
- When you do so, the key to not get lost in the derivations that follow is to understand the equation that allows you to go from observed to counterfactuals
- For the **treated** group, Y_{0i} is a counterfactual while Y_{1i} is observed. So for the treated group $Y_{1i} = Y_i$
- For the **control** group, Y_{1i} is a counterfactual while Y_{0i} is observed. So for the control group $Y_{0i} = Y_i$
- Remember, Y_i is the observed outcome, and units are either treated or controls, never both

Multiple units to solve the causal inference problem

- We solve the causal inference problem or the **prediction problem** with multiple units
- We use the information from other units to make predictions. We do this in everyday reasoning:
  1. Each of us at different times is a different unit. That's how we figure out what works for us
  2. Different people getting different treatments are a source of comparisons
- We often compare different people doing different things (i.e. getting different "treatments"): Crossfit or yoga for building muscle? What does Tom Brady and Halle Berry do that they don't seem to age at all? If I do X, what would happen?
- **But**... there is still something missing
- Intuitively, **to correctly predict the counterfactual**, we want to compare **similar units**

Comparison and bias

- Suppose that we have multiple units. Some of them received a binary treatment. We observe a continuous outcome Y_i
- We could compare the average **observed outcomes** or **conditional expectation**: $E[Y_i|D_1 = 1] - E[Y_i|D_1 = 0]$
- **Small digression**: We often compare averages (first moments), but we could compare other things like the median, the variance (second moment), or the distribution or any other moment. Or probabilities
- The above expression can be rewritten as

  $E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$
- There is a lot to unpack there. Notice that $E[Y_i|D_1 = 1] = E[Y_{1i}|D_1 = 1]$. In words, the mean of the observed outcome for the treated is the same as the mean potential outcome for the treated when treated (duh?)
- In the same way, $E[Y_i|D_1 = 0] = E[Y_{0i}|D_1 = 0]$. We are just **rewriting observed expectations in terms of potential outcome expectations**

Comparison and bias

- We then squeezed in a term that equals zero: $E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 1]$
- Rewriting the mean observed difference $E[Y_i|D_1 = 1] - E[Y_i|D_1 = 0]$ into two components provides some intuition
- $E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]$ is the **definition** of the average treatment effect on the treated; a comparison of the observed outcome for the treated minus what would have happened if the treated had not been treated
- The second part is what Angrist and Pischke call the **selection bias**: $E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$. Can be written as $E[Y_{0i}|D_i = 1] - E[Y_i|D_i = 0]$ (that's different than in the textbook)
- The **selection bias** is the difference in the average Y_{0i} between those who received the treatment and those who did not

Association is causation

- To interpret the comparison between observed outcomes $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$ as causal effects, we want the **selection bias to be zero**, so $E[Y_{0i}|D_i = 1] = E[Y_i|D_i = 0]$
- Said in words, if the outcome for the treated when they are not treated [counterfactual] is the same as the [observed] outcome for the control units when they are not treated, then a simple comparison of average outcomes can provide an estimate of average treatment effects
- When is this going to be true? When the control group outcome provides an unbiased **prediction** as to what would have happened to the treated group if not treated and vice versa
- Note the **vice versa** part. We could have rewritten in a different way to make the selection bias the average difference in tersm of the treated group
- **Bottom line is this**: For causal effects, **treatment assignment D_i must be independent of potential outcomes**

Independence

- This is a good time for a quick refresher on **independence**. Two random variables are independent if and only if: $f_{X,Y}(x, y) = f_X(x)f_Y(y)$. For discrete random variables: $P(X = x, Y = y) = P(X = x)P(Y = y)$
- In terms of events: $P(A \cap B) = P(A)P(B)$. These definitions are not that intuitive but: What is the conditional probability if two events are independent?
- $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$
- So the probability of of A given than B occurs is just $P(A)$. In words, B happening does not affect $P(A)$ (and vice versa)
- **Better**: knowing one doesn't tell you anything about the other event chances of happening

Conditional independence

- Conditional independence is an important concept and closely related to regression models and the conditional independence *assumption*
- Events A and B are **conditionally independent** if $P(A \cap B|Z) = P(A|Z)P(B|Z)$
- More useful: If A and B are conditional independent given Z, then $P(A|B, Z) = P(A|Z)$
- In words, knowing B doesn't tells us anything about $P(A)$ **once we know Z**

Conditional independence...

- Usual example: probability of that two people are late when commuting back home from work. Once we know where they live, wen can tell if the probability of being late is independent or not
- Another example: Suppose you randomly assign patients to an experimental treatment T based on severity of illness S. People with more severe illness are given more chances of receiving the treatment. Furthermore, age A is positively related to severity of illness
- Therefore, T and A are **NOT independent** but T and A are **conditionally independent** given I:
- $P(T|A, I) = P(T|I)$

Ways to solve the fundamental problem of causal inference (a)

- **a) Randomization**: Randomly dividing units into treatment and control groups before the action or intervention takes place makes groups of units **comparable**
- Don't underestimate how deep (and mysterious) randomization truly is
- After an action is applied to one group, the **other group provides the counterfactual** or the best prediction
- Another way of saying this is that the potential outcomes do not depend on the particular group assignment, they are independent
- In the new epi causal inference literature they call this **exchangeability**: the groups are so similar that they could be exchanged; **it does not matter which group receives the intervention**

Randomization and independence

- Going back to our notation to go from observed to counterfactual outcomes. The difference in average expected outcomes is

  $E[Y_i|D_1 = 1] - E[Y_i|D_1 = 0] = E[Y_1 i|D_1 = 1] - E[Y_0 i|D_1 = 0]$
- But randomization ensures that $E[Y_0 i|D_1 = 1] = E[Y_0 i|D_1 = 0]$ so we have:

  $E[Y_i|D_1 = 1] - E[Y_i|D_1 = 0] = E[Y_1 i|D_1 = 1] - E[Y_0 i|D_1 = 1]$
- Same as $E[Y_i|D_1 = 1] - E[Y_i|D_1 = 0] = E[Y_i|D_1 = 1] - E[Y_0 i|D_1 = 1]$
- In words, with randomization, comparing observed outcomes between treated and control groups is the same as comparing the observed outcome for the treated group with the counterfactual (not observed) outcome for the control group
- If you **frame causal inference as a prediction problem**, randomization solves the problem because what happens to the control provides a good prediction of what would happened to the treated group if they had not received treatment

Average treatment effects versus individual effects

- Note that we solved the causal inference problem with multiple units and obtained average treatment effects
- **We cannot estimate individual effects**
- In a clinical trial of medication, individuals are randomized into treatment and control. The effects are average treatment effects
- It doesn't mean that every single individual would have the same effect. Effects could be **heterogeneous** (not the same)
- We only know that on average there was (or not) a treatment effect. The treatment effect could be the same for all (**homogeneous**) but we don't know

Treatment effects: ATE, ATET, LATE

- The equation

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 1] =$$
$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

- can be described in words as:

  Observed average treatment effect = Average treatment effect on the treated + selection bias
- So there is more than one treatment effect. Most common: the **average treatment effect** (ATE) and the **average treatment effect on the treated** (ATET)
- Under randomization, ATE = ATET because the selection bias is zero
- We will come back to this and the notion of another type of treatment effect: the Local Average Treatment Effects or LATE when we discuss empirical methods

Regression to compare observed expected values

- In our review of regression analysis we saw that the linear/OLS model is a **conditional expectation function**, so we can use a regression model. Suppose we estimate the **population** model

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

- The expected value of the outcome conditional on treatment is: $E[Y_i|D_i = 1] = \beta_0 + \beta_1 + E[\epsilon_i|D_i = 1]$
- For the control: $E[Y_i|D_i = 0] = \beta_0 + E[\epsilon_i|D_i = 0]$
- The **difference** (contrast) between treatment and control is:

$$\beta_1 + E[\epsilon_i|D_i = 1] - E[\epsilon_i|D_i = 0]$$

- So β_1 is a measure of treatment effect provided

$$E[\epsilon_i|D_i = 1] = E[\epsilon_i|D_i = 0]$$

- (Next class we'll see that testing $H_0 : \beta_1 = 0$ is the same as a t-test of independent samples)

Regression

- When is $E[\epsilon_i|D_i = 1] - E[\epsilon_i|D_i = 0]$ equal to zero?
- When treatment assignment D is **mean independent** of the error term; both are just $E[\epsilon_i]$
- **Definition**: Y is mean independent of X iff $E(Y|X = x) = E(Y)$ for all x such that the probability that $X = x$ is not zero. In words, if the conditional expectation is the same as the unconditional expectation
- Mean independence is a milder assumption than the usual assumption. An alternative and stronger assumption is to say that $E[\epsilon_i|D_i = 1] = E[\epsilon_i|D_i = 0]$ if ϵ is independent from D
- Independence implies that D and ϵ are *uncorrelated* in any functional form. On the other hand, mean independence doesn't imply zero correlation. But independence does imply mean independence
- **Confusion alert**: I can't emphasize enough how easily you can get confused with this line of reasoning. The key is that we are talking here about a **population** model. Something we do not observe. To ascertain if $D \perp \epsilon$, we must argue based on subject knowledge. In this sense, this assumption is a **exclusion restriction**. In **any** sample, the residual $\hat{\epsilon} \perp D$

Regression

- Randomization ensures $\epsilon \perp D$
- You probably have seen this before: We can use a regression when the **regression model includes all confounders**
- If we leave a confounder out, it's part of the error term ϵ, and by definition of confounder, it's correlated with both D and Y
- Notice that I said confounders: they matter only if the variable(s) left out are correlated with **BOTH**, D and Y
- Make sure you understand the definition of confounder

Reminder: confounder

- From Wikipedia (with some edits):

  "In statistics, a confounding variable (also confounding factor, a confound, a lurking variable or a confounder) is a variable in a statistical model that correlates (directly or inversely) with **both** the dependent variable and an independent variable"
- Another way: a confounder predicts both a covariate and outcome
- More often than not we talk about a confounder in relationship to a treatment
- Antidepressants and the risk of suicide: severity of depression is a confounder. It's correlated to both the probability of taking antidepressants (the treatment) and the probability of suicide (the outcome)

Ways to solve the fundamental problem of causal inference (b)

- **b) Statistical adjustment**: The previous discussion provides the intuition for the other way in which we can obtain causal effects: statistical adjustment with (some form of) regression
- Consider **conditional randomization**. In conditional randomization randomization depends on the value of one or more variables
- For example, we want to confirm the efficacy of a new treatment that a pilot study showed to be effective. For ethical reasons, we want the patients that are in worse condition to have more chances of getting the treatment. We classify patients by severity level (for simplicity, 0 or 1 variable but it could be continuous). If severity (Z) equals 1, the patient has 80% chances of getting the treatment. If the patient is not that sick ($Z = 0$), then she only has 20% chances of getting the procedure. Assume that Z is associated with Y (correlated in any from)

Conditional randomization

- Note the consequence of this type of conditional randomization. On average, the treated group has sicker patients than the control group
- In other words, the selection bias is not zero: $E[Y_{0i}|D_i = 1] \neq E[Y_i|D_i = 0]$
- In the population model $Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$, severity Z is part of ϵ and therefore there is no independence or mean independence ($E[D|Z = 1] \neq E[D|Z = 0]$)
- The expected value of the outcome for the control groups is NOT the same as the expected value of the counterfactual for the treated group. Another way: the control group cannot provide a counterfactual for what would have happened to the treated group (and vice versa)
- But there is an easy fix. We can control for severity in our regression model: $Y_i = \gamma_0 + \gamma_1 D_i + \gamma_3 Z_i + \eta_i$
- In this sense, **statistical adjustment is the oldest causal inference method we have**. Now η is *conditional* independence from D: $\eta_i \perp D_i|Z_i$

The ignorable treatment assignment assumption

- We can now state the **fundamental** assumption to justify some type of regression adjustment (not necessarily linear/OLS)
- This assumption says that conditional on observable covariates X_n, the assignment of units to experimental groups is **independent of potential outcomes**:

  $(Y_{0i}, Y_{1i}) \perp D_i | X_{ni}$, where Y_{0i}, Y_{1i} are the potential outcomes for unit i, D is treatment assignment and X_n are a set of n **observable** covariates ($\perp$ is "perpendicular", "orthogonal")
- In the conditional randomization example, we satisfied this assumption by conditioning on Z
- The ignorability of treatment assignment says that if you can't control for confounders, your **statistical model is showing a correlation and not causation**
- **Jargon, jargon, jargon**: This assumption comes in many names, the most common perhaps is **no unmeasured confounders**. Other names: selection on observables, exogeneity, conditional independence assumption (CIA), ignorability

Refining assumptions

- We saw that to interpret the comparison between observed outcomes $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$ as causal effects, we want the **selection bias to be zero**, so:

$$E[Y_{0i}|D_i = 1] = E[Y_i|D_i = 0]$$

- If we know and observe covariates that affect selection intro treatment, we can extend this condition to be:

$$E[Y_{0i}|\boldsymbol{X}_i, D_i = 1] = E[Y_i|\boldsymbol{X}_i, D_i = 0]$$

- Notice that there is some nuance here. Going back to the example of randomization based on severity. Suppose that there was another factor used to randomize patients: gown color. If gown color is not related to the outcome Y, does it have to be part of the covariates in a model in order to obtain causal effects? NO. We want to control for **confounders**
- Although not obvious, we can in fact estimate a global average treatment effect or an average treatment for a **subpopulation defined by** covariates $\boldsymbol{X}$

Digression: Language matters a lot

- How we name things matters a lot and may create confusion: taxing lucky people who inherit a lot of money sounds different than "death tax," even if the end result is the same. A lie somehow is not the same as a "false statement" (dysphemism and euphemism, respectively)
- The CIA or ignorability assumption is also called **selection on observables**, perhaps the most common name and one I use a lot. It's helpful to convey the main result to justify regression adjustment: if the factors that determine treatment assignment are observable, we can control for them to obtain causal effects
- But assumption $(Y_{0i}, Y_{1i}) \perp D_i | X_{ni}$ leaves room for D to be correlated (in any form) with unobservable variables that determine treatment assignment. The key is that these unobserved variables must NOT be associated with Y_{0i}, Y_{1i} (the gowns example above)
- So **unconfoundness** or "**no unmeasured confounder**" would be a better word
- Big Wooldrige (2010) Chapter 21 has an interesting discussion about this (and presents the milder assuption of "Ignorability in Mean")

Heterogeneous treatment effects

- Estimating treatment effects using the model above adds a structure:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_3 Z_i + \epsilon_i$$

- The treatment effect we obtain will average across severity levels Z
- We could estimate two models: one for $Z = 1$ and one for $Z = 0$ (stratified models). We would get two estimates of treatment effects, say $\hat{\beta}_1$ if $Z = 1$ and $\hat{\gamma}_1$ if $Z = 0$. They may be different if there is effect heterogeneity, but we could combine them weighting them to obtain ATE. Using he sample of each stratum would be an intuitive choice (this insight goes back to Rubin, 1977)
- With heterogeneous treatment effects, the parametric model would be an interacted model:

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_3 Z_i + \gamma_4 D_i \times Z_i + \eta_i$$

- One message: we can extend the definition of causal effects to a subpopulation defined by covariates, in this case Z, severity level. Another message: average treatment effects mask heterogeneous treatment effects

Check

- The idea that the interacted model is weighting the estimates from the stratified analysis is important to consider
- With the fully interacted model, we will average over the heterogenity using the margins command (come back to this after the class on marginal effects)
- A better way of weighting would be using variances

```
bcuse bwght, clear
gen     smoked = 0
replace smoked = 1 if cigs ==0
qui reg bwght i.smoked##i.white
margins, dydx(smoked)
Average marginal effects                        Number of obs     =      1,388
Model VCE    : OLS
Expression   : Linear prediction, predict()
dy/dx w.r.t. : 1.smoked
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.smoked |   8.889065   1.488571     5.97   0.000     5.968966    11.80917
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
quietly {
   reg bwght i.smoked if white ==1
   scalar beta1 = _b[1.smoked]
   scalar N1 = e(N)

   reg bwght i.smoked if white ==0
   scalar beta2 = _b[1.smoked]
   scalar N2 = e(N)
}
di (N1*beta1 + N2*beta2)/(N1+N2)
8.8890654
```

Caution

- Angrist and Pischke used the linear/OLS model to discuss these issues **but statistical adjustment could be done with other models** and the average treatment effect is not the only measure of outcome
- The linear/OLS makes functional form assumptions that could be relaxed. Instead of estimating $E[Y_i|\boldsymbol{X}_i, D_i]$ with a linear/OLS model we could do so **nonparametrically** (local polynomial, lowess, kernel regression)
- Nonparametric methods are the basis of more "modern" approaches so we will cover them
- Also, the outcome could be a 0/1 variable; therefore, the outcome is a probability. We do not need the expectation to make sense of causal inference, but we are often interested in averages. Remember, causal effects are comparisons of potential outcomes
- The key issue is: **can the outcome of the control group be a good counterfactual for the treated group (and vice versa)?**
- The answer hinges on understanding the **assignment mechanism**. We know that with randomization the answer is yes. Without randomization, we need to make sure that we control for all factors that determine treatment and also affect outcomes (confounders)

Regression adjustment requires another assumption: **overlap**

- We actually need another assumption –often overlooked– to obtain causal effects using regression adjustment
- Intuition: Can we use the control group to make predictions about what would have happened to the treated had they not been treated if, say, the control group is made of people younger than 50 but most of the treated group is older than 75? Maybe, but we are relying on **extrapolation**
- For now, we can state the assumption. **Overlap**: For all $\boldsymbol{X} \in \varphi$, where φ is the support (domain) of the covariates $\boldsymbol{X}$, $0 < P(D = 1|\boldsymbol{X}) < 1$
- In words, for all the covariate values, there is a chance of seeing units in both treatment and control groups. Notice that randomization ensures overlap is satisfied
- We will see this issue when we cover propensity scores as propensity scores help diagnose this problem. Propensity scores and matching estimators offer a solution too
- Actually, $P(D_i = 1|\boldsymbol{X}_i)$ is the *definition* of the propensity score

Ways to solve the fundamental problem of causal inference (c)

- c) other ways to solve the problem of causal inference
- **Randomization** is not always possible, feasible, ethical, or easy to do
- When we observe all confounders – that is, ignorability, selection on observables, and son on hold– we can use **regression** methods and/or propensity scores/matching
- But we will also cover other methods to find causal effects: difference-in-difference, instrumental variables, regression discontinuity
- All of them have some version of LATE in the sense that we are able to find a type of treatment effect that may or may not be ATE or ATET
- All of them have assumptions that can be verified with data and **exclusion restrictions** that cannot be verified with data

Assignment mechanism

- If the key is to control for variables associated with both treatment D and outcome Y, it follows that the key to understand if we can interpret models as providing causal effects is to understand how units ended up receiving treatment (or not)
- **Assignment mechanism**: What determines which units receive (or not) the treatment?
- Imbens and Rubin (2015) define several classes of assignment mechanisms:
  1. **Classical randomized experiments**: Interventions under the control of the investigators. With good randomization, causality is not problematic (includes **conditional randomization**). There could be non-compliance or other issues, though
  2. **Regular assignment mechanisms**: Interventions not under the control or not known to the researcher – think observational studies. Several subtypes based on how much we know about assignment
- Most of health services research question are related to the second type

Assignment mechanism II

- If the units were not randomized, how some units ended up receiving treatment? What factors influenced that decision? **Can we observe these factors**? **This is perhaps the most important consideration of a study design**
- Example: taking antidepressants and the risk of suicide. You have observational data; patients were not randomized to take antidepressants
- Why some took antidepressants? Severity of illness, access to care, family history, preferences
- If you can't control for these factors, you do not have conditional independence between antidepressant use (treatment) and suicide (outcome)
- In other words, your study is showing a correlation, not a causation
- Others: Medicaid expansion and hospital closures. Flu vaccines and hospitalizations
- **You need good subject knowledge to analyze data!!**

Stable Unit Treatment Value Assumption (SUTVA)

- I mentioned that we need more assumptions even with randomization. SUTVA is one. SUTVA holds when:
  1. The **potential outcomes** for any unit do not vary with the treatment assigned to other units (**no interference**)
  2. For each unit, there are no different versions of each treatment level (**no hidden variation of treatments**)
- Another way of explaining SUTVA: SUTVA requires that the outcome of a particular unit depends only on the treatment to which the unit was assigned, not the treatments of other units (spillovers)
- This is an example of **exclusion restrictions**: assumptions that rely on theory or substantive knowledge to rule out (or in) the existence of a causal effect
- Assumptions about how the world works that allow us to "exclude" alternatives or define mechanisms that you can't test with data
- In instrumental variables, for example, we need to rule out the possibility that the instrument is related to the outcome (conditional on other factors)

On forests and trees

- This week's lectures in one slide:
  1. We defined causal effects as comparisons of potential outcomes
  2. We encountered a big problem: we only get to see one potential outcome for each treatment level
  3. We framed the problem as a prediction problem: we need a way to predict the counterfactual
  4. We used multiple units to find those predictions
  5. We saw that we solve the causal inference problem with a) randomization , b) statistical adjustment, c) other methods
  6. We still can't make causal statements for single units. We can find average treatment effects
  7. We saw that we can define several types of average treatment effects: ATE, ATET, LATE
  8. We saw that b) statistical adjustment requires 1) unconfoundness and 2) overlap
  9. We saw that for a), b), and c) we also need SUTVA
  10. We learned that we need to understand the assignment mechanism well to be able to argue that association implies causation
- The rest are details and proofs

How economists talk (or not?) to each other

- The model is **endogenous**: They mean that the ignorability assumption is not satisfied. There are unobservable variables not controlled for. The model has unmeasured confounders
- The model is not **identified**: It means that whatever model you are estimating does not represent the true model. You cannot learn the "true" causal value of the parameters from your model
- **Identification strategy**: the method used for finding causal effects, as in "My identification strategy is to use XYZ as an instrument for..."
- **Selection on observables**: The economist version of no unmeasured confounders, ignorability or conditional independence assumption (CIA). Meaning, people (it's usually people) selected into treatment based on factors that you can measure (and control for). So your treatment is endogenous

A framework for understanding the importance of potential outcome

- It's helpful to follow Heckman and Vytlacil (2007a) discussion on causal effects as three separate steps (paraphrasing to match our language)
  1. **Define** causal effects using potential outcomes
  2. **Identify** causal effects from a hypothetical population data
  3. **Estimate** parameters from observed samples
- An effect is identifiable if the counterfactual estimates are the same as the observable data
- We saw that with randomization causal effects are identified because we can show that $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$ or written in a different way: $E[Y_{0i}|D_i = 1] = E[Y_i|D_i = 0]$
- **Clones**: In every homework somebody mentions clones. Clones would be like identical twins born many years apart (I think). But twins are extremely helpful for causal inference for issues that we suspect are completely determine by biology. In a sense, randomization is creating clone groups

Summary

- The counterfactual framework offers a way of thinking about causal inference
- It has resulted in a lot of progress and new causal inference field in statistics; it has clarified the causal inference approach in econometrics even though some methods have been used for decades
- Get used to the language since we are going to use it for the rest of the class
- Remember, we are trying to learn if our models are descriptive/correlational or if they can have a causal interpretation
- But remember too that we often run models as descriptive or predictive tools. It's not that a model is automatically wrong if it doesn't have a causal interpretation. It depends on what you do with the model and how you interpret it