

Counterfactuals and Selection Bias

EC420 MSU Online

Justin Kirkpatrick

Last updated June 06, 2021

Lectures:

- (1) Motivation
- (2) Counterfactuals
- (3) Potential Outcomes Framework
 - ATE
- (4) Selection Bias
- (5) Randomization and Experiments

Motivation

top

What does it look like when $E[u|X] \neq 0$?

Counterfactuals

- Define counterfactual, treatment
- Introduce notation for counterfactuals
 - Potential outcomes framework

Selection Bias

- Define
- Examples

Course goals review

1. Understand core statistical concepts (OLS, t-tests and other tests)
2. Read, digest, and be critical of economic papers
3. Think critically about causality
4. Code in R with Rmarkdown
5. Learn current econometric models for causality
6. Design research questions

Counterfactuals

[top](#)

The Road Not Taken By Robert Frost

TWO roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

Make a difference in your life!

Image credit: [Book, Tea, and Sympathy](#)

Intuitive meaning

- The *counterfactual* is the outcome that *would be observed* had something else occurred.
 - Frost longs to know what *would have resulted* from choosing the other road
- Implies that a *counterfactual* is defined by:
 1. An outcome that has more than one potential value
 2. A choice or occurrence that may change the realization of the outcome

Let's call that "choice or occurrence" the **treatment**.

So, in the case of Robert Frost's poem:

- Outcome is, well, vague, which is the point of the poem.
 - But...we do know that he is looking at the difference of *some* outcome between roads he could have taken.
- The "choice or occurrence" that changed the (vague) outcome is the choosing of the "less traveled" road

Or, think of the two potential outcomes from a medical treatment - outcome with the drug, and outcome without.

The important point here is that **you do not get to observe both outcomes in reality**

- You can speculate, you can make comparisons, but you never get to see both (or all) counterfactuals.
- Ol' Bob Frost thinks he knows both.

In a way, when we run a regression, we **pretend** like we know both as well:

$$LifeExp = \beta_0 + \beta_1 cigarettes + \beta_2 1(exercises) + u$$

- If we plug in **0** for *exercises*, we get the $E[LifeExp | exercises == 0]$.
- If we plug in **1** for *exercises*, we get $E[LifeExp | exercises == 1]$
- β_2 is the difference in expected life expectancy associated with exercising
- We *identify* this in our regression by having data on people who do exercise, and data on people who don't exercise.
 - And people who smoke a lot, and people who don't smoke at all...
 - **And** we assume that all the other things in u follow $E[u | cigarettes, exercise] = 0$

What would it look like if *exercises* were in the error term?

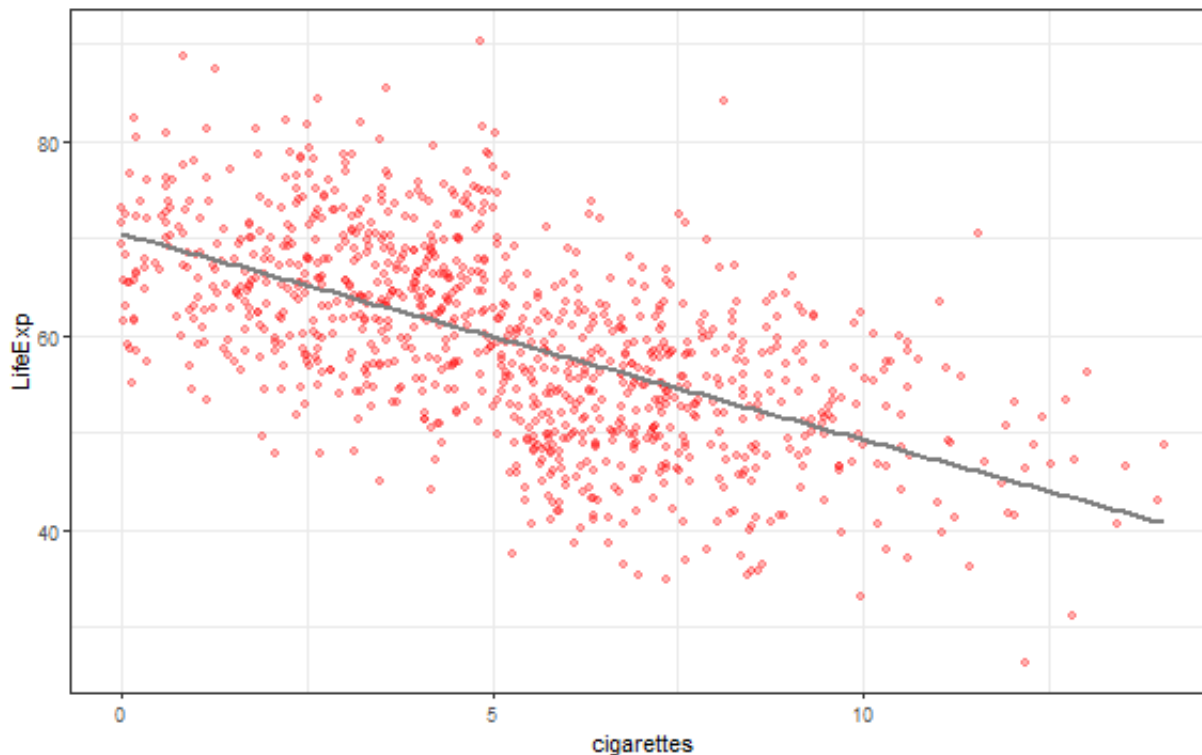
$$LifeExp = \tilde{\beta}_0 + \tilde{\beta}_1 cigarettes + \underbrace{\tilde{u}}_{\beta_2 1(exercises) + v}$$

Where v is actually a truly random error term ($E[v|X] = 0$).

Q: Do we think that *exercises* (in $E[\tilde{u}]$) is uncorrelated with *cigarettes* (X)?

Q: If not, what is the **relationship** between $E[\tilde{u}]$ and X ?

If we run the naive regression, we can fit the line:



This line is drawn assuming that $E[\tilde{u}|X] = 0$

- That is, assuming $E[\underbrace{\beta_2 1(exercises) + v}_{\tilde{u}} | cigarettes] = 0$

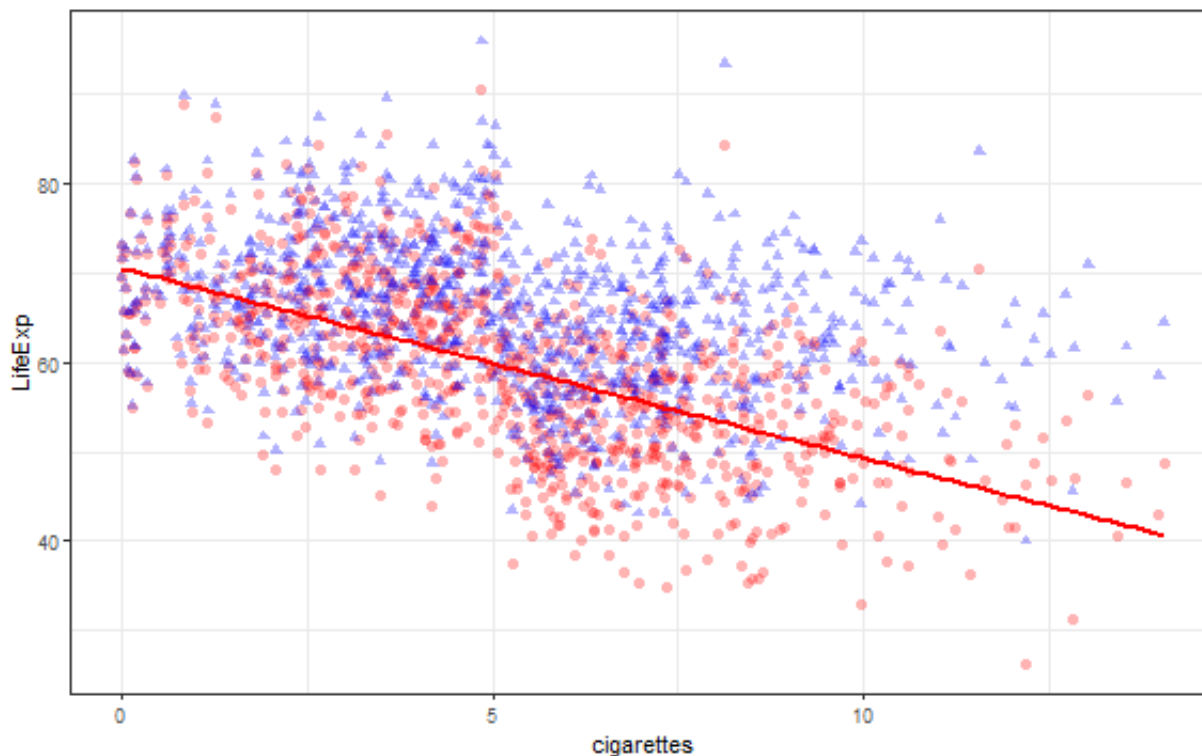
But what if we acknowledge that $E[\tilde{u}|X] \neq 0$

We would need to account for the $E[\tilde{u}] \neq 0$ when X is larger

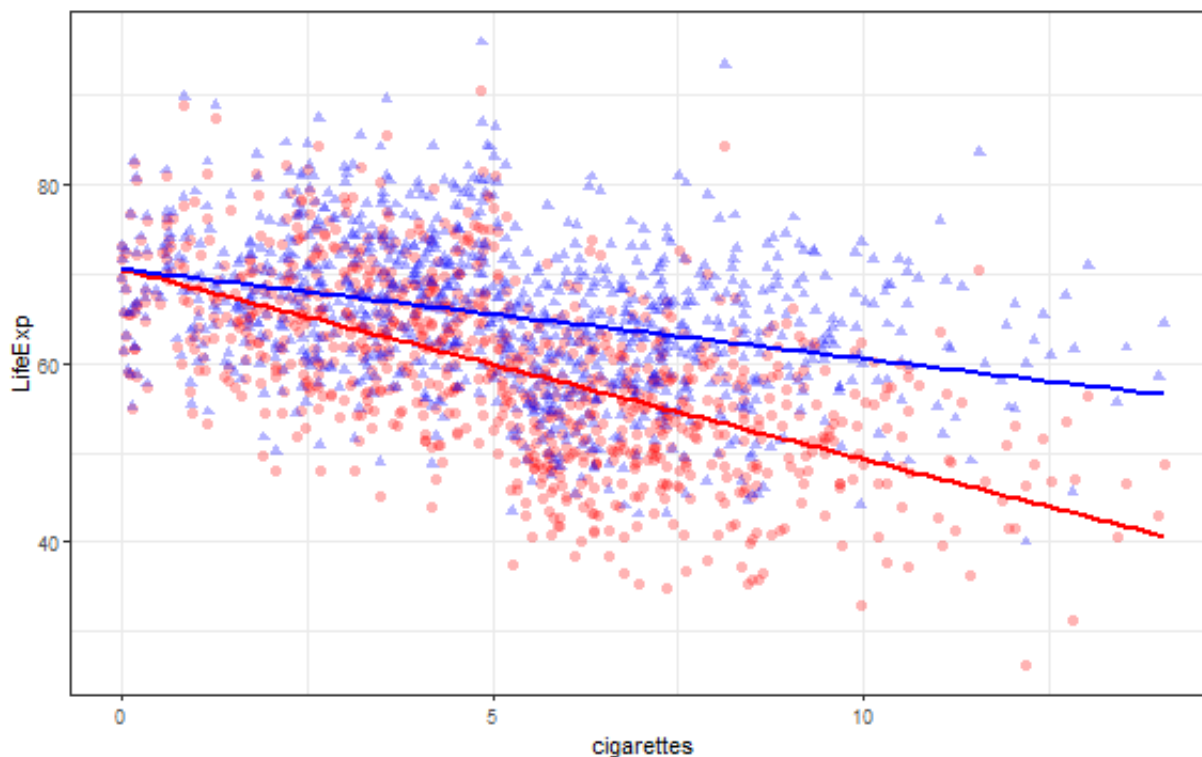
- $E[\tilde{u}]$ is "exercises" and X is cigarettes
- We would think $E[\tilde{u}]$ is negatively correlated with X

If we think about including that "wedge", we'd put out line of best fit in a different place.

- Of course, we would have to know $E[\tilde{u}|X]$ to actually correct
- So let's pretend for a moment

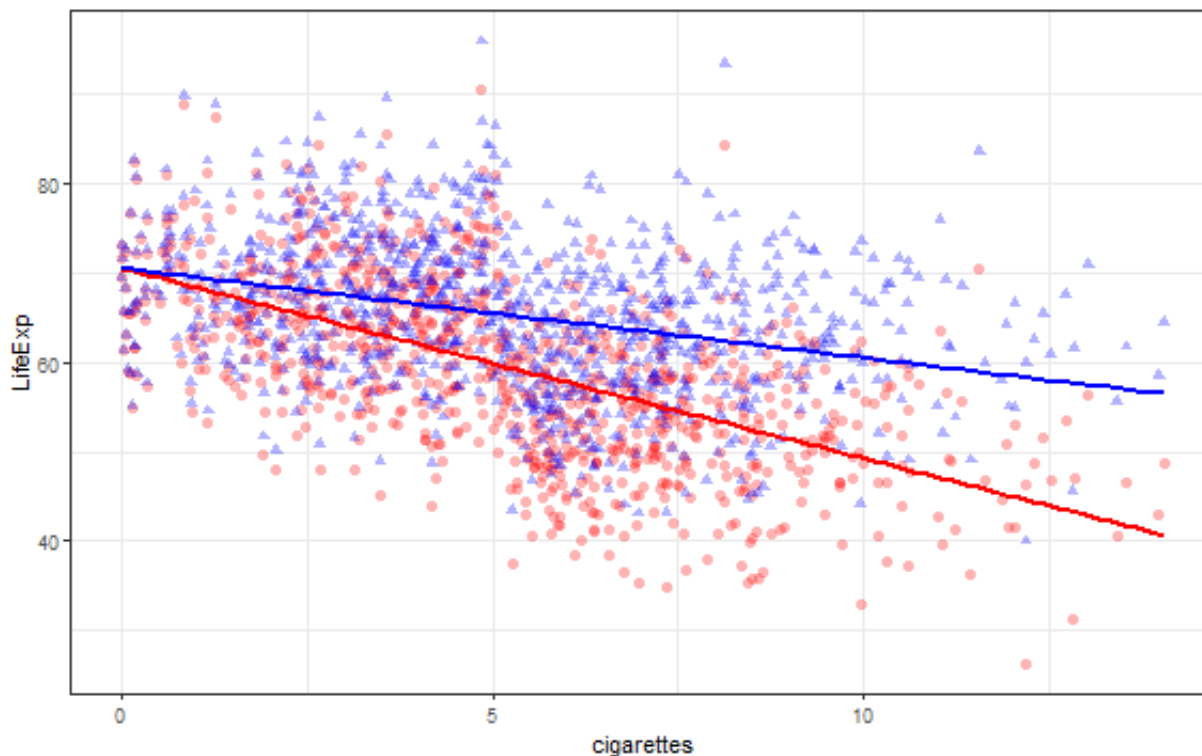


- The red dots are the given data, plotting $LifeExp \sim cigarettes$
- The blue triangles are where the points **would be if we could account for** $E[\tilde{u}|X] \neq 0$
- Here, $E[\tilde{u}|X] = -1.12 \times cigarettes$

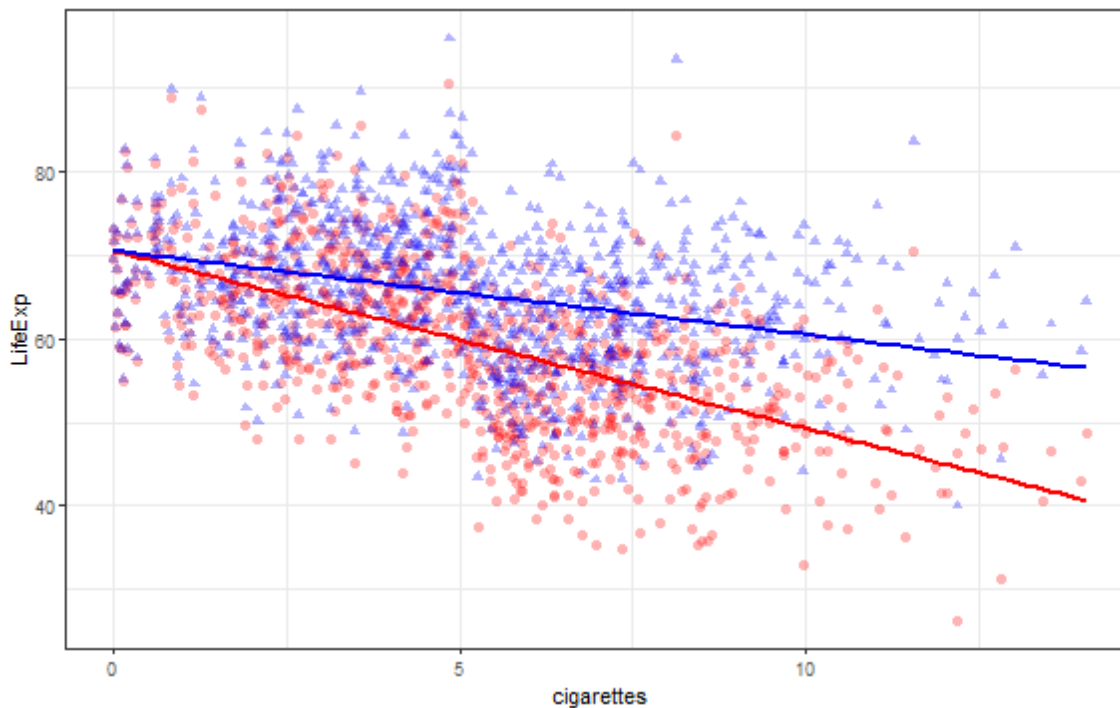


If we draw the line using the blue points which **account for the (unobserved) relationship between u and $cigarettes$** , then we get a different slope (the blue line).

The red line is the same as the blue line **if and only if** $E[\tilde{u}|X] = 0$.



So if we observe one person who smokes a lot, and one who smokes very little, in order to say that the effect of smoking is the difference between the two, we have to be able to say we've accounted for everything else.



If we had the same person (so all u are the same), but we could see that person "with" and "without" smoking, we wouldn't have to worry, right? *exercise* would be the **same**, so would everything else in \tilde{u}

So...

- Counterfactuals are closely related to our MLR.4 $E[u|X] = 0$
- Having a valid counterfactual is like knowing that everything in u is the same when the "treatment" is different

Potential Outcomes Framework

The "potential outcomes" notation (Rubin, 1974)

- Let Y be our outcome
 - Y_i is individual i 's outcome
- Let D be the **treatment** variable.
 - $D_i = 1$ if i is "treated"
 - $D_i = 0$ if i is "not treated"
 - Later, we'll talk about what happens if "treatment" is not 0/1 (binary)
- Let X be the observable characteristics (e.g. X_i)

Y has some potential outcomes:

We will index them (tell them apart) using a subscript

- Let Y_{1i} be the outcome *if* Y_i is treated (if $D_i = 1$)
- Let Y_{0i} be the outcome *if* Y_i is not treated (if $D_i = 0$)

Then, we can write:

$$Y_i = \begin{cases} Y_{1i}, & \text{if } D = 1 \\ Y_{0i}, & \text{if } D = 0 \end{cases}$$

What do we want to know?

What we'd like to know is $Y_{1i} - Y_{0i}$, individual i 's difference in outcomes with and without treatments, which is the *causal effect of the treatment* for i .

- Remember, *every* i has two potential outcomes.

For Robert Frost: $Y_{1i} - Y_{0i} = \text{"all"}$

But isn't there a problem here? How do we know Y_{1i} **and** Y_{0i} ?

We only see one, not both.

We could try looking beyond a single i , right? But that causes problems, as we saw in the previous example with *cigarettes* and *exercises*

Just as before, the function E is the *expectation* function.

$E[Y]$ is the expectation of Y

$E[Y] = \frac{1}{N_{pop}} \sum_{n=1}^{N_{pop}} Y_n$, which is the *population average*.

Of course, we don't usually observe the full population, so we use *sample average* instead.

When the *sample average* is used, MM will usually write $\hat{E}[Y]$ or $E_n[Y]$

Conditional expectations

- Let $E[Y]$ be the expected outcome
- Let $E[Y|D = 1]$ be the expected outcome if treated
 - Read this as "Expectation of Y *conditional on* D equaling 1"
- Let $E[Y|D = 0]$ be the expected outcome if untreated
 - Read this as "Expectation of Y *conditional on* D equaling 0"

This leads to some intuitive combinations:

- $E[Y_0|D = 0]$ and $E[Y_1|D = 1]$ are what we observe

And some not-so-intuitive combinations:

- $E[Y_1|D = 0]$ and $E[Y_0|D = 1]$ (these **are not observed**)
- Let's think about what these really mean, as they will come up again.

What do we really want to know

Unless we are i , we'd actually really like to know the population analog:
 $E[Y_1 - Y_0]$

This is the **Average Treatment Effect, or ATE**

- $ATE = E[Y_1 - Y_0]$

We could calculate this by taking the expectation of the causal effect of treatment:

$$E_i[Y_{1i} - Y_{0i}] = \frac{1}{N} \sum_{i=1}^{N_{pop}} Y_{1i} - Y_{0i}$$

But that means we need to know *both* Y_{1i} and Y_{0i} for each and every i . Getting the **ATE** has the same issue as getting the causal effect: **we don't observe the counterfactual**

Why would we want to know this?

Imagine that D is a drug. Then $E[Y_1 - Y_0]$ is the drug's **population** effect.

- If Y is the number of years surviving after a medical diagnosis, then the *ATE*, $E[Y_1 - Y_0]$ is the expected increase in years of survival.

Let's define another treatment effect: SATE

SATE is the *Selected average treatment effect*, the *ATE* for those that receive treatment.

$$SATE = E[Y_1 - Y_0 | D = 1]$$

This might not be quite as useful as the *ATE*, but under certain assumptions, they are the same. Specifically, under the *constant-effects assumption* (MM, p.10)

The **SATE**

The **SATE** also requires that we observe both Y_{1i} and Y_{0i} for each i .

How about we make an assumption:

- We observe some individual i where $D = 1$ and thus we observe Y_{1i}
- We observe some other individual j where $D = 0$ and thus we observe Y_{0j}
- Let's pretend i and j are interchangeable and we compare $E[Y_1] - E[Y_0]$!

We've got some Y_1 's and some Y_0 's, so we have:

- $E[Y|D = 1] = E[Y_1|D = 1]$
- $E[Y|D = 0] = E[Y_0|D = 0]$

Selection Bias

Selection bias occurs when the people who get treated (have $D = 1$) have **different values** for Y_{1i} and Y_{0i} from those who do not get treatment

Put formally: $(Y_{1i}, Y_{0i}) \perp D$

That is, the **potential** outcomes are **independent** of treatment

We observe $E[Y|D = 1]$ and $E[Y|D = 0]$

$$\begin{aligned}
 & E[Y|D = 1] - E[Y|D = 0] \\
 &= E[Y_1|D = 1] - E[Y_0|D = 0] \\
 &= E[Y_1|D = 1] - E[Y_0|D = 0] - E[Y_0|D = 1] + E[Y_0|D = 1] \\
 &= E[Y_1|D = 1] - E[Y_0|D = 1] + E[Y_0|D = 1] - E[Y_0|D = 0] \\
 &= \underbrace{E[(Y_1 - Y_0)|D = 1]}_{\text{SATE}} + \underbrace{E[Y_0|D = 1] - E[Y_0|D = 0]}_{\text{selection bias}}
 \end{aligned}$$

The first equality is based on the definition of what we observe.

The second is just subtracting and adding $E[Y_0|D = 1]$, which cancels out.

The fourth is using the definition of conditional - two expectations that are conditional on the same thing can be combined.

The last is from the definition of SATE (and introduces the definition of **"selection bias"**).

What is bias in general?

In this context, bias can be thought of as anything that distorts a statistic of interest like the *ATE*, *SATE*, or population mean.

$$E[W] \neq \theta$$

What is selection bias?

Selection bias is the distortion that is included when we try to infer a population causal effect from differences in observed outcomes (e.g. what we did before).

Selection bias is a "bias" because it biases our estimate of the thing we want: population *SATE*.

$$\underbrace{E[Y|D = 1] - E[Y|D = 0]}_{\text{Observed Comparison of Means}} \neq \underbrace{E[Y_1 - Y_0|D = 1]}_{\text{SATE, the thing we want}}$$

What are the common sources of selection bias?

This can occur if:

- Treatment is assigned in some way that is associated with potential outcomes
 - Treatment is given to people with higher Y_{1i}
 - Optimizing behavior can lead to this
- Treatment is self-selected
 - If people are allowed to "select in" to treatment
 - Esp. if they know their potential outcomes
 - E.g. the "smokers at the hospital" example

Note that traditionally, "selection bias" refers to non-random sampling (e.g. you survey people outside of a senior citizens home on their age, and use that to estimate a population average age - it'll be biased!)

Some examples of selection bias

If we are interested in the effect of college on earnings...

- We would worry that the sort of people who select into college do so because they know they have a high Y_{1i} , where the treatment is "attending college".

If we are interested in the effect of a drug on expected survival time...

- We would worry that people who select to take a drug might have very bad (low) values of Y_{0i} , the outcome without the drug, and thus "select in" to treatment. Without the drug, they would have a much worse (lower) survival expectancy than those who do not select in to treatment. If we compare the average survival of the treated to the untreated, we will not get a useful measure.

If we are interested in the effect of health insurance on some health outcome...

- We would worry that people who have insurance might have very different potential health outcomes. For one, people who purchase insurance tend to have higher income, and higher income earners tend to have better health regardless of insurance. $Y_{0i} > Y_{0j}$ when $income_i > income_j$.

Some examples of selection bias

If we are interested in the effect of smoking on probability of stroke (and we ignore age)...

- What would the treatment and control groups be defined by?
- Would there be any reason people would "select" into a group?
- Would there be any reason to think there would be a selection bias?

Formally, selection bias occurs when

$$E[Y_0|D = 1] - E[Y_0|D = 0] \neq 0$$

Which is when

$$E[Y_0|D = 1] \neq E[Y_0|D = 0]$$

The potential outcome associated with no treatment is different for those who are treated relative to those who are otherwise untreated.

Another way of saying this is that treatment D is **not** independent of the values of (Y_{1i}, Y_{0i})

- $(Y_{1i}, Y_{0i}) \not\perp D$

And, $E[Y_0|D = 1]$ is never observed, so we cannot just subtract off the selection bias from our naive comparison of means: $E[Y|D = 1] - E[Y|D = 0]$.

To summarize, when we do a comparison of means:

$$E[Y|D = 1] - E[Y|D = 0]$$

we get the SATE + Selection Bias:

$$E[Y_1 - Y_0|D = 1] + E[Y_0|D = 1] - E[Y_0|D = 0]$$

And when we regress $y = \beta_0 + \beta_1 D + u$, we are doing a "comparison of means"

Another example:

Imagine we have an energy conservation program where a homeowner has someone come to their home and show them their energy consumption, as well as ways they can save electricity. We would want to know the treatment effect of this program - maybe policymakers want to know how effective it is. Or maybe we have a theory that costly information keeps homeowners from consuming a more efficient amount of energy.

Our sample is four people.

Name	Use Without Tmt	Use With Tmt
Allison	4	3
Bertrand	5	4
Carmine	3	2
Dalia	4	3

Unrealistic data (we observe both potential outcomes)

- Here we get to *unrealistically* observe both outcomes Y_{i0} and Y_{i1}
- Define the **Average Treatment Effect**?
- Calculate the Average Treatment Effect here.

Our sample is four people.

Name	Treated	Use Without Tmt	Use With Tmt
Allison	TRUE	4	3
Bertrand	TRUE	5	4
Carmine	FALSE	3	2
Dalia	FALSE	4	3

Unrealistic data (we observe both potential outcomes)

- Allison and Bertrand are "treated"; we observe their "Use With" values
- Carmine and Dalia are "untreated"; we observe only their "Use Without" values

First, let's be naive. Let's look at the average use **of those who received the treatment** and compare it to the average use **of those who did not**,
 $E[Y|D = 1] - E[Y|D = 0]$

Name	Treated	Observed
Allison	TRUE	3
Bertrand	TRUE	4
Carmine	FALSE	3
Dalia	FALSE	4

Why might the naive estimate not be the same as the ATE (or the SATE)

Why does our naive comparison of means not yield the (S)ATE?

Let's draw this on the board

Now, we've seen **selection bias** in action, and we showed that it is part of what we get from a naive comparison of means **using algebra on conditional expectations**

How would we test to see if our "treated" group is different from our "untreated" group?

- If we observed both before and after for each individual, we could **still** use a naive comparison of means **if** the before and after consumption were just about equal.
- Is it just about equal in the example?
- What if we didn't get to observe the "before" for each i ?

Covariate balance

All the other things we observe along with the outcome are our *covariates*.

We might hope that we would have "just about equal" values for the (unobserved) before if all those covariates were "just about equal" too.

TABLE 1.3
Demographic characteristics and baseline health in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible – catastrophic (2)	Coinsurance – catastrophic (3)	Free – catastrophic (4)	Any insurance – catastrophic (5)
A. Demographic characteristics					
Female	.560 [12.9]	–.023 (.016)	–.025 (.015)	–.038 (.015)	–.030 (.013)
Nonwhite	.172	–.019 (.027)	–.027 (.025)	–.028 (.025)	–.025 (.022)
Age	32.4 [12.9]	.56 (.68)	.97 (.65)	.43 (.61)	.64 (.54)
Education	12.1 [2.9]	–.16 (.19)	–.06 (.19)	–.26 (.18)	–.17 (.16)
Family income	31,603 [18,148]	–2,104 (1,384)	970 (1,389)	–976 (1,345)	–654 (1,181)
Hospitalized last year	.115	.004 (.016)	–.002 (.015)	.001 (.015)	.001 (.013)
B. Baseline health variables					
General health index	70.9 [14.9]	–1.44 (.95)	.21 (.92)	–1.31 (.87)	–.93 (.77)
Cholesterol (mg/dl)	207 [40]	–1.42 (2.99)	–1.93 (2.76)	–5.25 (2.70)	–3.19 (2.29)
Systolic blood pressure (mm Hg)	122 [17]	2.32 (1.15)	.91 (1.08)	1.12 (1.01)	1.39 (.90)
Mental health index	73.8 [14.3]	–.12 (.82)	1.19 (.81)	.89 (.77)	.71 (.68)
Number enrolled	759	881	1,022	1,295	3,198

Notes: This table describes the demographic characteristics and baseline health of subjects in the RAND Health Insurance Experiment (HIE). Column (1) shows the average for the group assigned catastrophic coverage. Columns (2)–(5) compare averages in the deductible, cost-sharing, free care, and any insurance groups with the average in column (1). Standard errors are reported in parentheses in columns (2)–(5); standard deviations are reported in brackets in column (1).

From Mastering Metrics: The Path from Cause to Effect. © 2015 Princeton University Press. Used by permission. All rights reserved.

Unfortunately, covariate balance makes a weak case about the balance of the unobserved factors (like the "before" consumption).

A lack of balance does make a strong case *against* balance in the unobserved factors.

- "Necessary but not sufficient"

We do have tools to let us *control* for differences.

The main takeaways from the last few lectures:

1. Every i has two potential outcomes
2. We don't get to see both of them
3. A counterfactual is the "what would have been"
4. The ATE is a population parameter
5. It is usually what we're after
6. We can use the $E[\cdot \cdot \cdot]$ notation to formalize our estimates
7. Doing so lets us see where selection bias may occur
8. Lack of balance in covariates may indicate presence of a selection bias
9. Balance in covariates does not indicate lack of selection bias

Randomization and Experiments

[top](#)

You are going to randomly assign treatment to the four people on the **next slide**:

- First, randomly decide to which group each member belongs.
 - Use `sample(c(T,F), size = 4, replace=T)` to randomly draw 4 treatment assignments in a vector
 - `TRUE` means that person is treated and we'll use their "with treatment" value, `FALSE` means the opposite
- Second, build a new observed dataset with the randomly realized observed values
 - Use the "with" value for that person if the person is randomly placed in the with-treatment
 - Use "without" value for that person if the person is randomly placed in the without-treatment
- Third, calculate the group means and the difference between the group means.
 - Repeat 15 times.

We want to see how this comparison of means between the with-treatment and without-treatment groups does in estimating the *ATE*.

Name	Use Without Tmt	Use With Tmt
Allison	4	3
Bertrand	5	4
Carmine	3	2
Dalia	4	3

Report the 15 trial results: the mean with, the mean without, and the difference between means

Calculate the average over all the randomized trials. What is the average with- and without-treatment value? What is the average difference?

What was our previous result comparing the means?

We know that the Average Treatment Effect is 1. Which one is closer to the true value of 1?

Randomization (of treatment) eliminates selection bias

Let's see if we can show this *formally*....

Recall that in a previous lecture, we used some algebra on conditional expectations to *decompose* the naive difference in means into two parts:

$$\begin{aligned} & E[Y|D = 1] - E[Y|D = 0] \\ &= E[Y_1|D = 1] - E[Y_0|D = 0] \\ &= E[Y_1|D = 1] - E[Y_0|D = 0] - E[Y_0|D = 1] + E[Y_0|D = 1] \\ &= E[Y_1|D = 1] - E[Y_0|D = 1] + E[Y_0|D = 1] - E[Y_0|D = 0] \\ &= \underbrace{E[(Y_1 - Y_0)|D = 1]}_{\text{SATE}} + \underbrace{E[Y_0|D = 1] - E[Y_0|D = 0]}_{\text{selection bias}} \end{aligned}$$

$$\begin{aligned}
 & E[Y|D = 1] - E[Y|D = 0] \\
 &= E[Y_1|D = 1] - E[Y_0|D = 0] \\
 &= E[Y_1|D = 1] - E[Y_0|D = 0] - E[Y_0|D = 1] + E[Y_0|D = 1] \\
 &= E[Y_1|D = 1] - E[Y_0|D = 1] + E[Y_0|D = 1] - E[Y_0|D = 0] \\
 &= \underbrace{E[(Y_1 - Y_0)|D = 1]}_{\text{SATE}} + \underbrace{E[Y_0|D = 1] - E[Y_0|D = 0]}_{\text{selection bias}}
 \end{aligned}$$

Now, we saw that randomization got us closer to the correct answer. It might even be unbiased! But what does randomization do?

- It randomly assigns treatment
- So what would $E[Y_0|D = 1]$ be if treatment, D , is random? It would be:

$$E[Y_0|D = 1] = E[Y_0|D = 0] = E[Y_0] \quad \text{under randomization}$$

So what happens to *selection bias*?

Under randomization, where treatment is randomly assigned, selection bias is equal to zero in expectation

Remember the definition of mean independence: $E[Y|X] = E[Y]$ for RV's Y, X

Applying the same thing to **potential outcomes** of Y and D :

- Randomization gives us $(Y_1, Y_0) \perp D$
- $(Y_1, Y_0) \perp D \Rightarrow E[Y_0|D] = E[Y_0]$

A comparison of means is then an *unbiased estimate* of the *SATE*

How would we randomize to answer the following with a simple cross-sectional comparison of means

- Effect of energy consumption information on future energy consumption
- Effect of health insurance on health outcomes
- Effect of college education on future earnings

Randomization, in it's simplest form, is quite straightforward. If the econometrician^{*} can reliably and with certainty assign treatment to individuals, then great!

But even in closely-managed operation, this won't always work.

- People are good at finding ways around assignment
- Even drug trials, which are about as controlled as you can get, are subject to shenanigans.
- *Endogenous treatment* is the term frequently used for this
 - It means that treatment is determined "within the system" (by an individual affected by the treatment).
 - You will be sick of the term "endogenous" by the end of this semester.

^{*} that's you

Relating SATE to ATE

- We seem to be able to get to SATE, but what about the ATE?
 - $SATE = E[Y_1 - Y_0 | D = 1]$
 - $ATE = E[Y_1 - Y_0 | D = 1]$
- If we assume that $E[(Y_1 - Y_0) | D = 1] = E[(Y_1 - Y_0) | D = 0]$, then the $SATE$ is the same as the ATE
- That is, the effect of treatment on the treated is the same as the effect of treatment on the untreated

This is the *constant effects assumption*

....OK, back to randomization

Randomized Control Trial

When the econometrician is able to randomly decide treatment, we call the study a *randomized control trial* or *RCT*.

- These are *very* common in development economics as they can be done at low-cost in many places.

Often, when doing a RCT, we want to make sure we have randomly selected a similar sample into each group (treated, untreated).

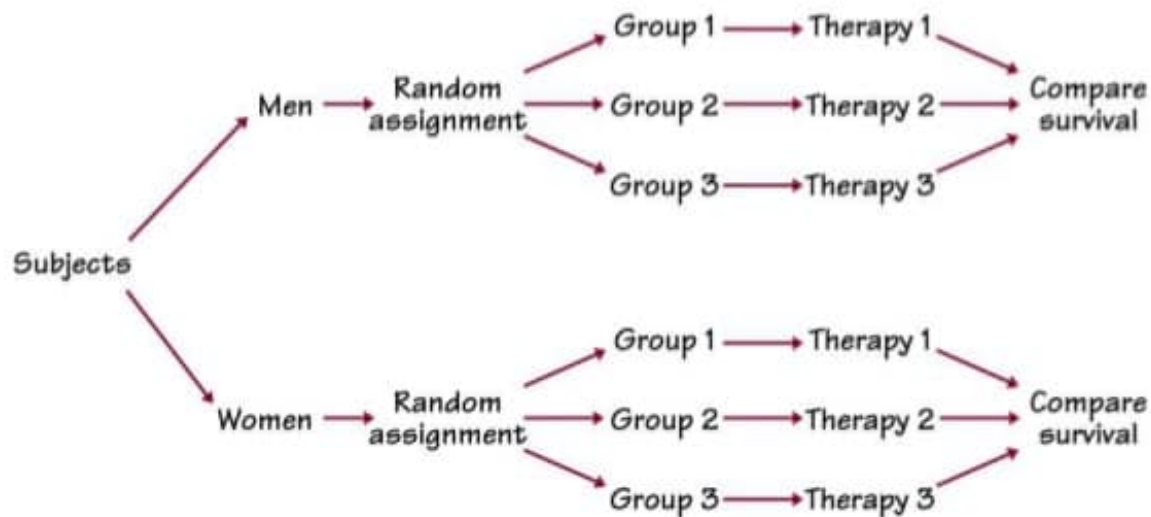
- Imagine if you randomly selected people to take part in a smoking campaign, but your "treated" group were mostly older people. If your outcome of interest is "mortality rate", what would happen?

Blocking (stratification)

A *blocked* or *stratified* experiment means we have randomly assigned treatment inside of observable subgroups to ensure balance.

- For example, if our population is 40% college graduates and 60% non-college graduates, and we are assigning treatment to 50% of the participants, we would treat 50% of all college graduates and 50% of all non-college graduates. We would *not* randomly select 50% of all participants.
- "Blocking" because you group observables into "blocks", then select
 - You can block on more than one characteristic: "college graduates over 40", etc..
- This can help ensure balance.
- Of course, this requires being able to assign treatment.

Block Design (Example Diagram)



Source: Bravo Students

1. We talked more about selection bias and examples of it.
2. We found that violation MLR.4 led to some pretty big problems.
3. We talked about randomization and how it can alleviate selection bias.
4. We also showed in notation how randomization alleviates selection bias in comparison of means.
 - Because $E[Y_0|D = 1] = E[Y_0|D = 0]$
5. We defined unbiased and consistent estimators (sample statistics).