Lecture Notes: Causal Inference Fall 2020

Claire Duquennois

7/28/2020

Contents

1 Regression and Causality: 2
1.1 Application: Regression for Prediction . . . 3
1.2 "No Causation Without Manipulation" . . . 7
1.3 The Rubin Causal Model . . . 8
1.4 The Conditional Independence Assumption . . . 9
1.5 Omitted Variable Bias . . . 11
1.6 Beware of the kitchen sink approach . . . 17
1.7 Selection on Observable Designs: How far does this get us? . . . 19
1.8 There is no Santa Claus: Arseneaux, Gerber and Green (2006) . . . 19

```r
library(stargazer)
```

```
##
## Please cite as:
## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library(here)
```

```
## Warning: package 'here' was built under R version 3.6.3
## here() starts at C:/Users/Claire/Dropbox/MQE_Causal_Inf/MQE_Causal_Inf
```

```r
# Set so that long lines in R will be wrapped:
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=TRUE)
getwd()
```

```
## [1] "C:/Users/Claire/Dropbox/MQE_Causal_Inf/MQE_Causal_Inf/Module_1_OVB"
```

```r
knitr::opts_knit$set(root.dir = here())
setwd('..')
a<-paste(setwd('..'))
knitr::opts_knit$set(root.dir = a)
b<-paste(setwd('..'))
knitr::opts_knit$set(root.dir = b)
getwd()
```

```
## [1] "C:/Users/Claire/Dropbox"
```

Welcome to Economic Inference from Data!

In this section of the quantitative methods sequence, you will learn the tools economists use to identify causal relationships and test hypotheses *in the wild*. This material is the bridge between theory and estimation. When

properly applied to the right dataset, approaches such as instrumental variables, differences-in-differences or regression discontinuity can allow us to test a hypothesis and make causal claims about the relationship between variables. The curriculum for this course will cover the main estimation strategies used today in empirical economics.

In addition to the official curriculum, there is an equally important hidden curriculum that will primarily be learned through individual and group homework assignments. This hidden curriculum involves acquiring the coding and work flow skills necessary to conduct causal analysis whether you are working solo or as a team. This will primarily involve learning how to work in R, using GitHub and learning how to present your findings clearly and effectively.

Before we get started, I would to highlight that these notes and materials draw heavily from pre-existing teaching materials. I would like to acknowlege Michael Anderson, Max Auffhammer, Sofia Villas-Boas, Jeremy Magruder and Fiona Burlig for their exceptional training and sharing their great lecture notes from which these borrow heavily. I would also like to thank Scott Cunningham, author of "Causal Inference: The Mixtape" and Joshua Angrist and Jorn-Steffen Pishke, authors of "Mostly Harmless Econometrics," whose examples and materials are also heavily featured in these notes. And xkcd, because nerds need comics too.

1 Regression and Causality:

As long as you satisfy certain trivial conditions, you can always run a linear regression. This is ok as long as you interpret the results appropriately.

Consider a dependent variable , y_i, and a vector of explanatory variables, x_i. We are interested in the relationship between the dependent variable and the explanatory variables. There are several possible reasons why we might be interested in this relationship, including:

1) Description– What is the observed relationship between y and x?
2) Prediciton–Can we use x to create a good forecast of y?
3) Causation– What happens to y if we experimentally manipulate x?

It is generally this last item where things get tricky. Before worrying about what regressions can't do, let's first focus on what regressions can do!

Obviously in the real world, particularly in the social sciences, few relationships are deterministic, and we are not going to be able to model our y outcomes perfectly. Recognizing this, we can focus on relationships that hold "on average," or "in expectation." Given our variables y and x, we may be interested in the *Conditional Expectation* of y given x. That is to say, given a particular value of x, where is the distribution of y centered? This relationship is given by the *Conditional Expectation Function*, or the CEF.

$$E[y_i|x_i] = h(x_i) \tag{1}$$

We can then define the CEF residual as:

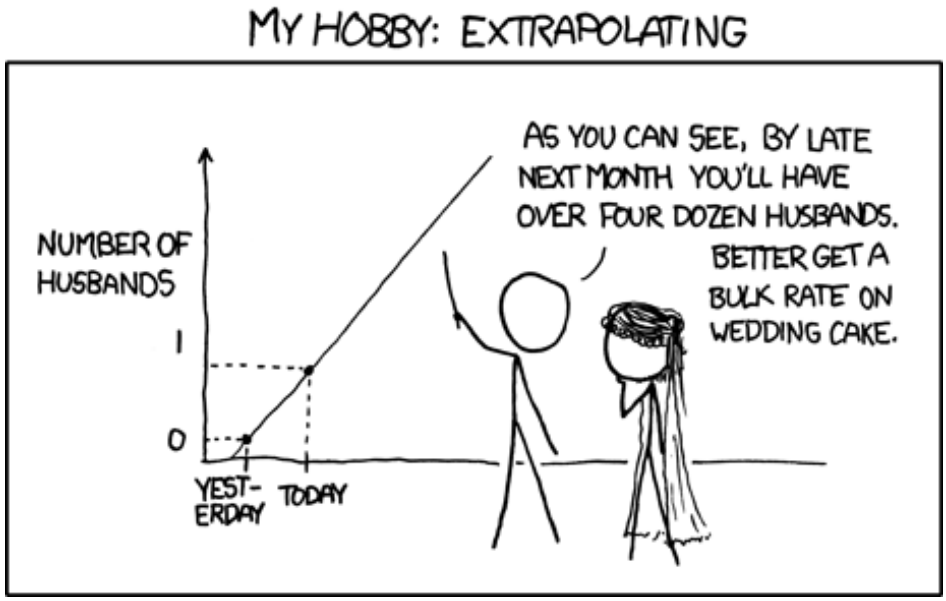
$$\begin{aligned} \epsilon_i =& y_i - h(x_i) \text{ where} \\ & E[\epsilon_i|x_i] = 0 \end{aligned} \tag{2}$$

Note that, because ϵ_i is the CEF residual, $E[\epsilon_i|x_i] = 0$ holds by definition– we do not require any exogeneity assumption regarding x_i.

If the CEF is linear, then a regression of y_i on x_i estimates the CEF. If the CEF is not linear, we still will often use linear regressions because:

1) they are computationally tractable
2) they have properties that we understand well,

3) they provide the best linear approximation to the CEF even when the CEF is non-linear (just don't try to extrapolate far beyond the support of x_i).



Thus, if what you are interested in is the CEF, linear regression is a good, and robust tool for estimating it.

You can estimate the CEF using the linear regression by doing the following. Run a linear regression of y_i on x_i, to generate estimates of β_0 and β_1 by doing the following estimation.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon \tag{3}$$

Then you can use the estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ to calculate $\hat{y}$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{4}$$

where $\hat{y}_i$ is the predicted value for y_i given x_i, ie. $\hat{y}_i = E[y_i|x_i]$, the CEF.

This is fine, as long as you interpret your results for what they are –an approximation of the conditional expectation function– and not what you might want them to be (an estimate of a causal relationship).

So if you are only intersted in decription or prediction, we can end the class here.

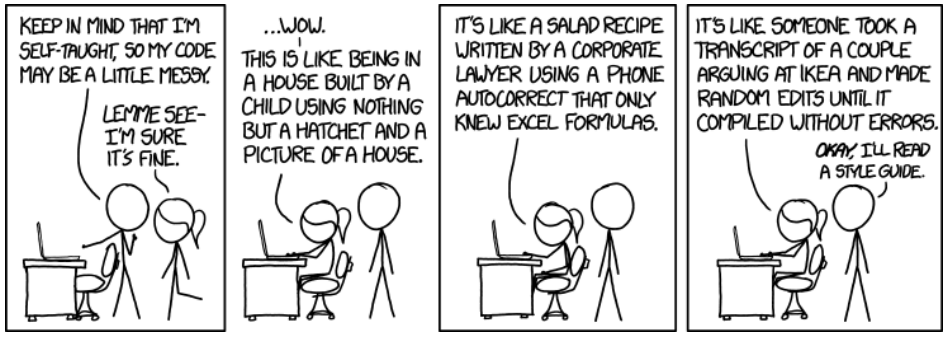
1.1 Application: Regression for Prediction

Suppose you are a bank and you are interested in predicting whether your customers will be able to repay their student loans. You have a subset of data from the CPS that includes individual earnings and the number of years they spent in education. You estimate the following on working age adults that are over 22 years old:

$$Income_i = \beta_0 + \beta_1 Schooling_i + \epsilon_i \tag{5}$$

to generate the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

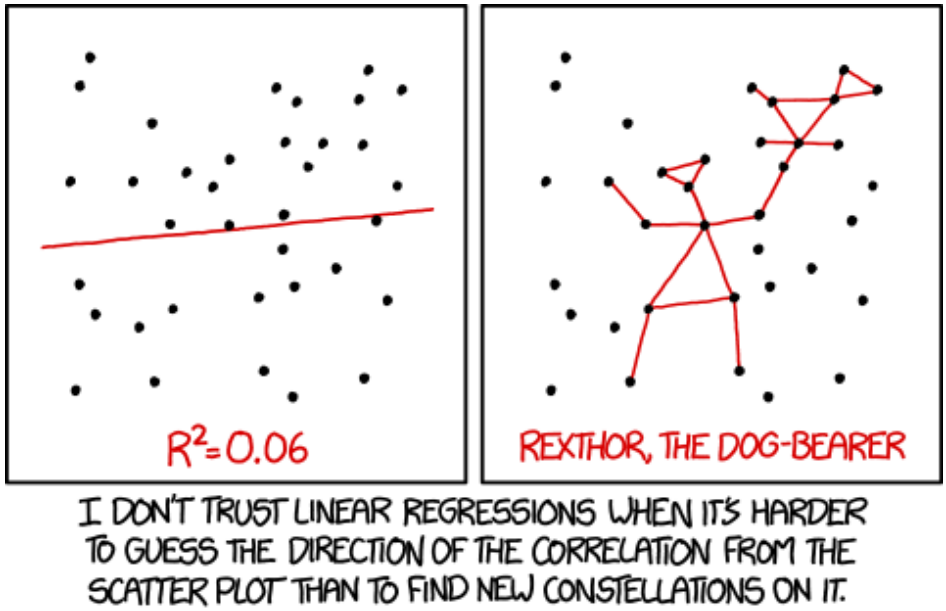
In R, you code:



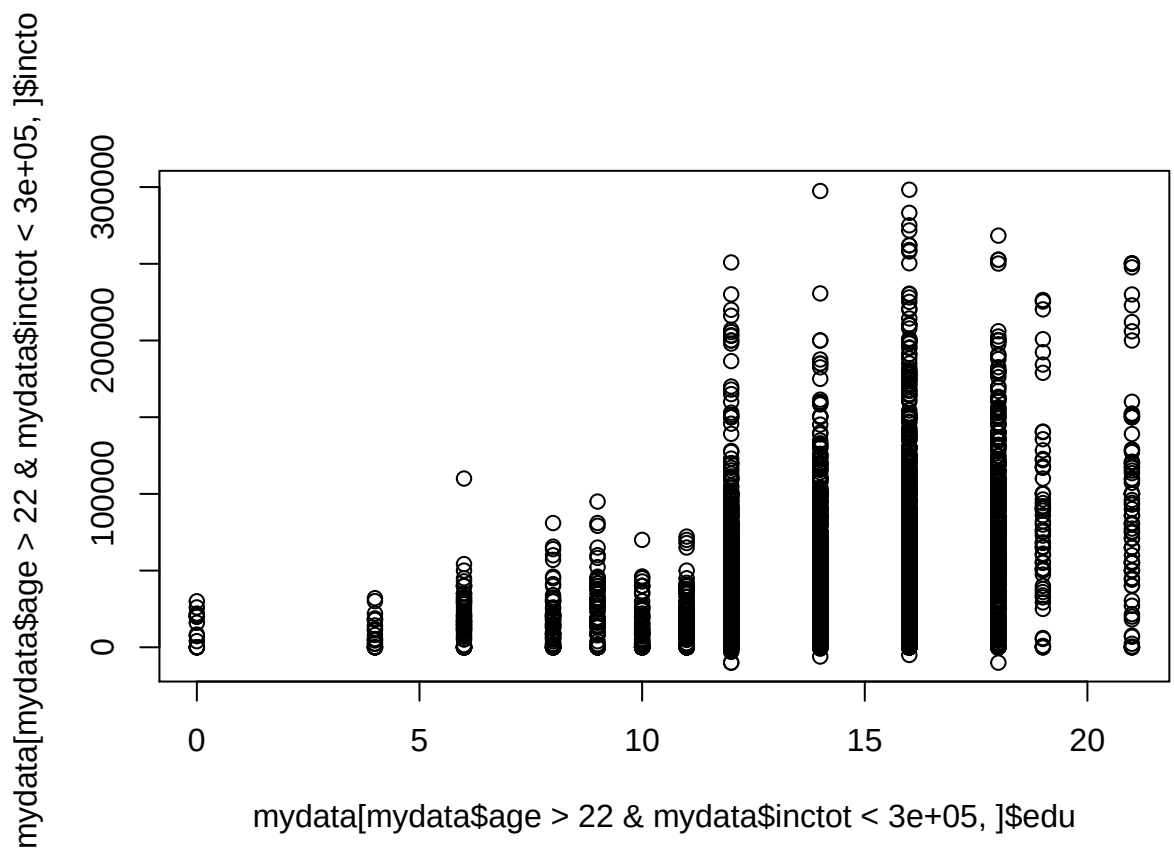

```r
mydata<-read.csv("data/data_M1_OVB/cps_clean.csv")

reg1<-lm(inctot~edu,mydata[mydata$age>22,])
summary(reg1)
```

```
##
## Call:
## lm(formula = inctot ~ edu, data = mydata[mydata$age > 22, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107200  -31055  -11015   13207 1069070
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -61933.9     5339.0  -11.60   <2e-16 ***
## edu           8054.0      375.3   21.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72970 on 4644 degrees of freedom
## Multiple R-squared:  0.09022,    Adjusted R-squared:  0.09003
## F-statistic: 460.6 on 1 and 4644 DF,  p-value: < 2.2e-16
```

Thus we have that $\hat{\beta}_0 = -61944$ and $\hat{\beta}_1 = 8054$, so an extra year of education predicts earnings that are \$8,115 higher.




```r
plot(mydata[mydata$age>22 & mydata$inctot<300000,]$edu,
     mydata[mydata$age>22 & mydata$inctot<300000,]$inctot)
```



Using these estimate we can predict the difference in annual income between a high school and college grad as

$$
\begin{aligned}
\widehat{Income}_{col} - \widehat{Income}_{hs} &= (\hat{\beta}_0 + \hat{\beta}_1 * 16) - (\hat{\beta}_0 + \hat{\beta}_1 * 12) \\
&= \hat{\beta}_1 * 4 \\
&= 8,054 * 4 = \$32,216.
\end{aligned}
$$

So we would predict annual returns of $32,216.

Alternatively, we could create an indicator variable set to 1 for individuals with college educations and estimate it on the subset of individuals who have at least 12 years of schooling:

$$
Income_i = \beta_0 + \beta_1 CollGrad_i + \epsilon_i \tag{6}
$$

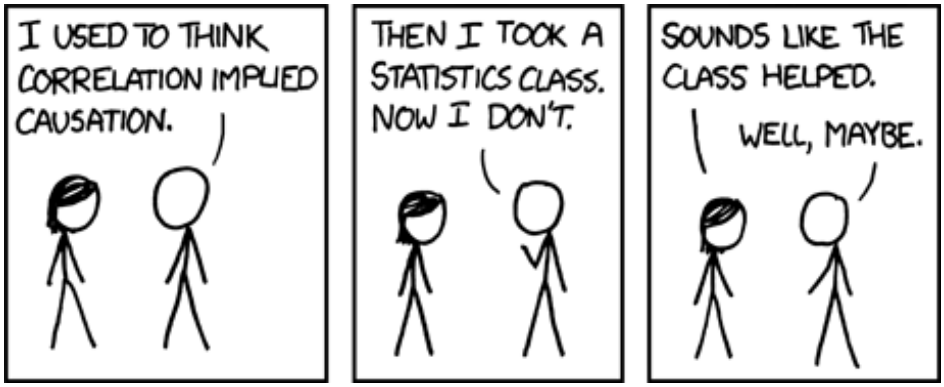
```
mydata$collgrad<-0
mydata$collgrad[mydata$edu>=16]<-1

reg2<-lm(inctot~collgrad,mydata[mydata$edu>=12 & mydata$age>22,])
summary(reg2)
```

```
##
## Call:
## lm(formula = inctot ~ collgrad, data = mydata[mydata$edu >= 12 &
##     mydata$age > 22, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -91324  -31425  -11433   13518 1054675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36483       1478   24.69   <2e-16 ***
## collgrad       44842       2409   18.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76010 on 4240 degrees of freedom
## Multiple R-squared:  0.07554,    Adjusted R-squared:  0.07532
## F-statistic: 346.4 on 1 and 4240 DF,  p-value: < 2.2e-16
```

Here we have that $\hat{\beta}_0 = 36483$ and $\hat{\beta}_1 = 44842$, so having a four year college degree predicts earnings that are $44,842 higher then the predicted income of $36,483 for individuals that finished highschool but did not get a four year degree.

The important thing to note here: we are not saying that the college degree **caused** higher earnings, but it does **predict** higher earnings, which for many applications, like approaving a loan, is sufficient.

To get at causation, we will need to do a lot more work.



1.2 "No Causation Without Manipulation"[1]

Economists (and other social scientists) tend to be interested in causal effects because we want to think about what the consequences will be if a particular policy is implemented or changed. Naive regressions without a carefully thought out research design are not enough to identify causal effects.

1.2.1 Application: The returns to schooling

It is easy to estimate the relationship between schooling and income. Using CPS data, I estimated the following linear regression:

$$Income_i = \beta_0 + \beta_1 Schooling_i + \epsilon \tag{7}$$

I then calculated the CEF, $E[Income_i|Schooling_i]$, the expected value of income conditional on years of schooling as

$$E[Income_i|Schooling_i] = \widehat{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 Schooling_i \tag{8}$$

These estimates, however, only tell us how income and schooling covary in the population. This DOES NOT reveal what would happen to an individuals income if there was an **"exogenous"** change to schooling. For instance if I raised the mandatory number of years students must stay in school or if I increased the schooling of a randomly selected group of students by 1 year. This is because the number of years a student spends in school is **"endogenously"** determined: the students who can afford to go to school, and who expect returns from additional schooling, stay in school. There is no reason that a regression coefficient that is estimated data on **endogenous** schooling choices should correspond to the effects of an **exogenous** change in schooling. In order to estimate the **causal** effect of schooling, we will need to identify some type of **"manipulation"** that created an **exogenous** change in schooling.

Note that it is not the case that one coefficient is "right" and the other is "'wrong" – the two coefficients are simply different and which one is "correct" depends on what question you are trying to answer. If you are interested in predicting income based on some knowledge of a person's education levels, using equation 22 is fine. \

It is **CORRECT** to interpret the estimates from equations 22 as

"We expect the income of a person with one additional year of schooling to be \$ $\hat{\beta}_1$ higher."

[1] This phrase comes from Holland (1986).

It is **INCORRECT** to interpret equations 22 as

"One additional year of schooling CAUSES earnings to increase by \$ $\hat{\beta}_1$."

The *Rubin Causal Model* discussed below offers a useful framework with which to understand common estimation techniques. More importantly, it is useful in framing and understanding what question you are trying to answer or what effect you are trying to estimate by using the idea that a unit has different *potential outcomes* depending on it's *treatment* status.

1.3 The Rubin Causal Model

Suppose that we have N units, $i = 1, ..., N$, drawn randomly from a large population. We are interested in the effect of some binairy treatment variable, D_i, on an outcome, Y_i. We postulate that each unit faces two potential outcomes,

$$\begin{aligned} Y_i &= \begin{cases} Y_i(1) & \text{if } D_i = 1 \text{ (\textit{the treatment condition})} \\ Y_i(0) & \text{if } D_i = 0 \text{ (\textit{the control condition})} \end{cases} \\ &= Y_i(0) + (Y_i(1) - Y_i(0))D_i \end{aligned} \tag{9}$$

The key challenge is that we will never observe both $Y_i(1)$ and $Y_i(0)$ though it must be theoretically possible that we could observe either $Y_i(1)$ or $Y_i(0)$. If you cannot conceptualize both $Y_i(1)$ and $Y_i(0)$ for the same unit, then D does not correspond to a treatment that is potentially manipulable and we cannot talk about the causal effect of manipulating D without further defining the problem.

Example:

Does going to college <u>cause</u> higher earnings? Let the treatment be going to college. We can imagine that each high school graduate faces two potential outcomes:

$$\textit{potential outcomes} = \begin{cases} earnings_{i,col} & \text{if } i \text{ goes to college (\textit{the treatment condition})} \\ earnings_{i,nocol} & \text{if } i \text{ does not go to college (\textit{the control condition})} \end{cases} \tag{10}$$

Though we can conceive of both $earnings_{i,col}$ and $earnings_{i,nocol}$, we will only ever observe one or the other, but we can imagine a policy or intervention that could make either of these values observable. Thus "college" is a treatment that is potentially manipulable.

Does being a woman <u>cause</u> lower earnings? Gender (or race) is not potentially manipulable. My gender has been in integral part of all the choices I have made throughout my life. It is not really possible for me to imagine some intervention that would reveal what my earnings outcomes would have been if was a man. Thus, though the relationships between, gender (or race) and earnings are extreamly interesting and worthy of study, and we know that beeing a woman <u>predicts</u> lower earnings, the question as posed is ill defined.

Using the notation above, we define the *causal effect* for tratment $D = 1$ on outcome Y for unit i as:

$$Y_i(1) - Y_i(0) = \tau_i \tag{11}$$

τ_i is the treatment effect for unit i. The treatment effect is **relative**. It tells us how the outcome for i with treatment compares to their outcome without treatment. How would i's earnings with a college degree compare to her earnings without a college degree? Note that this value need not be constant across different units, which is why the τ is indexed by i. Many (most) treatments have heterogeneous effects: a college degree may increase some people's earning potential a lot more than others, for example.

The imporant thing to realize here is that though $\tau_i = Y_i(1) - Y_i(0)$ we never actually observe both $Y_i(1)$ and $Y_i(0)$ for a given unit. This leads to the following theorem:

Fundamental Problem of Causal Inference: It is impossible to observe the value of $Y_i(1)$ and $Y_i(0)$ in the same unit i and, therfore, it is impossible to observe τ_i, the effect for unit i of the treatment on Y_i.[2]

The fundamental problem of causal inference would appear to rule out any precise estimation of τ_i and, at the unit level, it is true that we can never observe the exact treatment effect. However, all is not lost. We are generally interested in relationships that hold on "average," or in expectation. In this context, it is possible to estimate the quantities of interest. We define $\bar{\tau}$ as the *average causal effect* or *average treatment effect* (ATE) of the treatment relative to the control as the expected value of the difference $Y_i(1) - Y_i(0)$, or

$$\bar{\tau} = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]. \tag{12}$$

With the right research design, it is possible to estimate the ATE.

1.4 The Conditional Independence Assumption

Suppose the treatment effect is the same for everyone so that $Y_i(1) - Y_i(0) = \tau$, a constant. If this is the case, I can re-write equation 9 as

$$\begin{aligned} Y_i &= Y_i(0) + \tau D_i \\ Y_i &= E[Y_i(0)] + \tau D_i + Y_i(0) - E[Y_i(0)] \\ Y_i &= \alpha + \tau D_i + \eta_i \end{aligned} \tag{13}$$

where $\alpha = E[Y_i(0)]$, $\tau = Y_i(1) - Y_i(0)$, and η_i is the random part of $Y_i(0)$ since $\eta_i = Y_i(0) - E[Y_i(0)]$. We can then see that the expected outcomes for someone with treatment ($D_i = 1$), and without treatment ($D_i = 0$) are given by

$$\begin{aligned} E[Y_i(1)] &= \alpha + \tau + E[\eta_i | D_i = 1] \\ E[Y_i(0)] &= \alpha + E[\eta_i | D_i = 0] \end{aligned} \tag{14}$$

so that we can break down the difference between these outcomes as

$$E[Y_i(1)] - E[Y_i(0)] = \underbrace{\tau}_{\text{treatment effect}} + \underbrace{E[\eta_i | D_i = 1] - E[\eta_i | D_i = 0]}_{\text{selection bias}}. \tag{15}$$

It is thus clear that if we naively estimate the linear regression specified in equation 13, our estimate $\tilde{\tau}$, will not equal the true causal treatment effect τ, if there is selection bias such that $E[\eta_i | D_i = 1] \neq E[\eta_i | D_i = 0]$. This will happen if absent treatment, those who would select into treatment have a differnt expected outcome compared to those who would not select into treatment, such that $E[Y_i(0) | D_i = 1] \neq E[Y_i(0) | D_i = 0]$. This is because treatment is not random, formally if

$$\{Y_i(1), Y_i(0)\} \not\perp D_i, \tag{16}$$

that is if the outcome is not orthogonal (or independent) to treatment status.

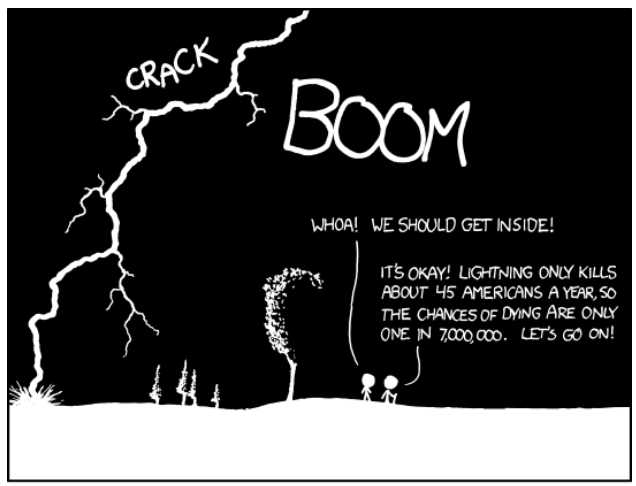
[2] Holland (1986).

Example:

Suppose I naively estimate the following regression on my observational data from the CPS, where $college_i$ is a dummy variable set to one if the person received a college degree, and generate estimated of α and τ,

$$earnings_i = \tilde{\alpha} + \tilde{\tau} college_i + \epsilon_i. \tag{17}$$

If I am interested in the causal treatment effect of a college degree on earnings, $\tilde{\tau}$ will not give an accurate estimate since it is biased by who selects into getting a college degree. If the people who receive college degrees would have had higher earnings even absent the degree then $E[\tilde{\tau}] \neq \tau$ since $E[\eta_i|D_i = 1] > E[\eta_i|D_i = 0]$.



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

The **Conditional Independence Assumption** asserts that conditional on observed characteristics, X_i, selection bias disappears. This means that once I control for X_i, treatment is as good as randomly assigned. Formally,

$$\{Y_i(1), Y_i(0)\} \perp D_i|X_i. \tag{18}$$

If this is the case, conditional-on-X_i comparisons have a causal interpretation. In other words,

$$E[Y_i(1)|X_i] - E[Y_i(0)|X_i] = E[Y_i(1) - Y_i(0)|X_i]. \tag{19}$$

Returning to equation 15, we saw that with selection bias, our estimate $\tilde{\tau}$ is not equal to the true causal treatment effect τ if there is selection bias. Suppose now that CIA holds given a vector of observed covariates, X_i. We can thus decompose the random term, η_i, into a linear function of observable characteristics X_i, and an error term ν_i, such that

$$\eta_i = X_i'\gamma + \nu_i, \tag{20}$$

where γ is a vector of population regression coefficients that is assumed to satisfy $E[\eta_i|X_i] = X_i'\gamma$. Since γ is defined by the regression of η_i on X_i, the residual ν_i and X_i are uncorrelated by construction. Moreover, by virtue of the CIA, we have

$$E[Y_i(D)|X_i] = \alpha + \tau D + X_i'\gamma. \tag{21}$$

The residual in the linear causal model

$$Y_i = \alpha + \tau D_i + X_i'\gamma + \nu_i \tag{22}$$

is therfore uncorrelated with the regressors, D_i and X_i and the estimated regression coefficient $\hat{\tau}$ is equal to the causal effect of interest, τ.

Example:

Returning to the example of earnings and college completion, we saw that

$$Income_i = \tilde{\alpha} + \tilde{\tau} college_i + \epsilon_i \tag{23}$$

generates a based estimate of τ due to selection bias since $E[\tilde{\tau}] \neq \tau$. Now suppose that CIA holds if I condition on a student's household income (*hhinc*). This means that once I control for student household income, which students complete college versus not is as good as randomly assigned.
If this assumption holds, then I generate the following estimates

$$earnings_i = \hat{\alpha} + \hat{\tau} college_i + \hat{\gamma}_i hhinc_i + \epsilon_i \tag{24}$$

and we have that $E[\hat{\tau}] = \tau$. Thus we can say that a college degree causes earnings to increase by $\$\hat{\tau}$.

It is important to note that the conditional independence assumption is a big assumption, and in many cases is not valid. Do we really believe that if I control for household income, who goes to college is as good as randomly assigned? Can you think of other important variables that would correlate with both completing college and earnings? If yes then you likely still have a biased estimate of the treatment effect.

1.5 Omitted Variable Bias

What happens to our estimate of the treatment effect τ when I introduce a new control variable?

Suppose I estimated the following model,

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{1i} + \epsilon_i \tag{25}$$

but I suspect that there is a relationship between my explanatory variable, x_{1i} and some other variable, x_{2i}, such that

$$x_{2i} = \rho_0 + \rho_1 x_{1i} + \varepsilon_i, \tag{26}$$

and that my outcome variable is also affected by x_i such that the true model is given by

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \nu_i. \tag{27}$$

where $\beta_2 \neq 0$.

Substituting equation 26 into 29 and rearanging gives

$$Y_i = \underbrace{(\beta_0 + \beta_2\rho_0)}_{\tilde{\beta}_0} + \underbrace{(\beta_1 + \beta_2\rho_2)}_{\tilde{\beta}_1} x_{1i} + \underbrace{(\beta_2\varepsilon_i + \nu_i)}_{\epsilon_i}. \tag{28}$$

So

$$\tilde{\beta}_1 = \underbrace{\beta_1}_{\text{treatment effect}} + \underbrace{\beta_2\rho_2}_{\text{bias}} \tag{29}$$

Thus we can see that the biased coefficient $\tilde{\beta}_1$ is equal to the true treatment effect, β_1, plus a bias term which depends on β_2 (the correlation between x_{2i} and Y_i), and ρ_2 (the correlation between x_{2i} and x_{1i}). On average, our regression is going to miss the true population parameter by $\beta_2\rho_2$.

We can use this regression to sign the bias. The sign of β_2 is obtained from thinking about how the dependent variable Y_i is correlated with the omitted variable ($cov(Y_i, x_2)$) and the sign ρ_2 is obtained by thinking about how the independent variable of interest, x_1, is correlated with the omitted variable ($cov(x_1, x_2)$). Multiplying the signs of these covariances can then establisht the sign of the bias, as seen in the table below,

	$Cov(x, x_{ov}) > 0$	$Cov(x, x_{ov}) < 0$
$Cov(y, x_{ov}) > 0$	Upward bias	Downward bias
$Cov(y, x_{ov}) < 0$	Downward bias	Upward bias

Example:

Suppose I am interested in thinking about how health relates to income. Using my CPS sample of working age adults, I begin by estimating the following model,

$$Income_i = \beta_0 + \beta_1 Health_i + \epsilon,$$

where heath is an rspondents subjective assesment of their health with 1 being very health and 5 being very unhealthy.

```
reghealth<-lm(inctot~health,mydata)
```

It then occurs to me that respondents age could be biasing this estimate since it is likely that $cov(health_i, age_i) > 0$ and $cov(income_i, age_i) > 0$ leading to an upward bias.

```
reghealth2<-lm(inctot~health+age ,mydata)
```

I then realize that schooling could also impact both health and income, as it is likely that $cov(health_i, schooling_i) < 0$ and $cov(income_i, schooling_i) > 0$ leading to a downward bias.

```
reghealth3<-lm(inctot~health+age+edu ,mydata)



stargazer(reghealth,reghealth2, reghealth3, type="latex", header=FALSE,
          title="Income and health", omit.stat=c("f", "ser"))
```

Table 1: Income and health

	Dependent variable:		
	inctot		
	(1)	(2)	(3)
health	−8,176.764***	−11,489.250***	−6,315.248***
	(991.929)	(1,011.858)	(1,003.357)
age		1,066.011***	953.415***
		(85.002)	(81.792)
edu			7,540.903***
			(365.735)
Constant	65,599.570***	28,903.240***	−82,442.910***
	(2,414.934)	(3,770.567)	(6,501.394)
Observations	5,000	5,000	5,000
R^2	0.013	0.044	0.119
Adjusted R^2	0.013	0.043	0.118
Note:			*p<0.1; **p<0.05; ***p<0.01

Any variable that is jointly correlated with my explanatory variable (*health*) and my outcome variable (*income*), will bias my estimate of β_1. *age* and *education* are certainly important ones but there are likely many more meaning that even our estimate from the third specification still cannot be interpreted as an unbiased causal estimate of the true β_1 and should be interpreted with caution.

Simulating Omitted Variable Bias in R:

To better understand how an omitted variable can bias our coefficient estimate, I will generate a simulated dataset. The advantage of working with simulated data is that with a simulation I can know the exact data generating process (DGP), and thus know exactly what the true coefficients should be. I can then run different estimations and see if the estimated coefficients are biased or not.

Suppose the data generating process is as follows. My outcome variable, Y depends on two variables, V_1 and V_2 such that

$$Y_i = \beta_0 + \beta_1 V_{1i} + \beta_2 V_{2i} + \epsilon_i$$

where V_1 and V_2 are correlated with $Cor(V_1, V_2) = 0.5$.

To simulate this DGP, I generate a data.frame of two correlated random variables pulled from standard normal distributions.

```
library(MASS)
library(ggplot2)
```

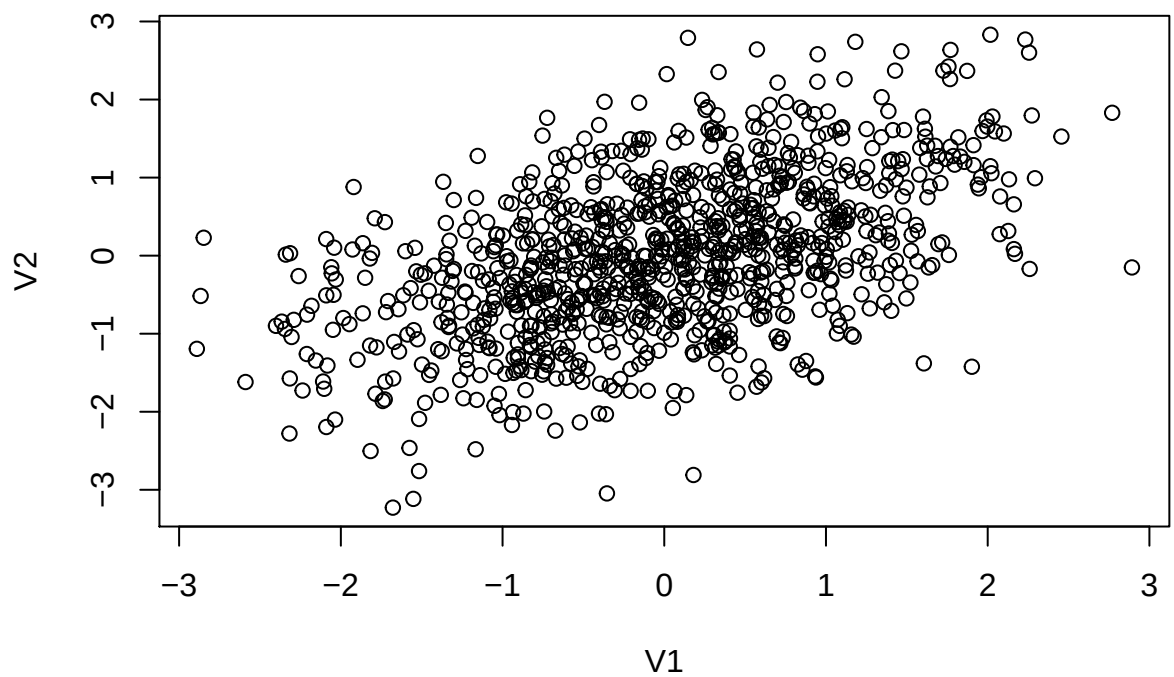
```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
set.seed(1999)
out <- as.data.frame(mvrnorm(1000, mu = c(0,0),
                     Sigma = matrix(c(1,0.5,0.5,1), ncol = 2),
```

```
                        empirical = TRUE))
cor(out)
```

```
##     V1  V2
## V1 1.0 0.5
## V2 0.5 1.0
```

```
plot(out)
```



I then add an error term for each observation, also drawn from a standard normal.

```
out$error<-rnorm(1000, mean=0, sd=1)
```

I can now simulate the data generating process. Suppose we want the true $\beta_1 = 5$ and $\beta_2 = 7$. I can simulate the outcome variable as follows:

```
#The data generating process
B1<-5
B2<-7

out$Y<-out$V1*B1+out$V2*B2+out$error
```

I can generate an unbiased estimate such that $E[\hat{\beta}_1] = \beta_1$ by estimating the correctly specified model

```
sim1<-lm(Y~V1+V2, data=out)
summary(sim1)
```

```
##
## Call:
```

```
## lm(formula = Y ~ V1 + V2, data = out)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85251 -0.66712 -0.04396  0.63625  2.78181
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01470    0.03102  -0.474    0.636
## V1           5.00724    0.03584 139.722   <2e-16 ***
## V2           7.02333    0.03584 195.979   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.981 on 997 degrees of freedom
## Multiple R-squared:  0.9913, Adjusted R-squared:  0.9913
## F-statistic: 5.687e+04 on 2 and 997 DF,  p-value: < 2.2e-16
```

As you can see, the estimated coefficients are very close to the true values.

However if I misspecify my model and fail to include V_2 then $E[\tilde{\beta}_1] \neq \beta_1$, as you can see below,

```
sim2<-lm(Y~V1, data=out)
summary(sim2)
```

```
##
## Call:
## lm(formula = Y ~ V1, data = out)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2747  -4.2230  -0.2513   4.2774  19.1838
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0147     0.1949  -0.075     0.94
## V1            8.5189     0.1950  43.683   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.164 on 998 degrees of freedom
## Multiple R-squared:  0.6566, Adjusted R-squared:  0.6563
## F-statistic:  1908 on 1 and 998 DF,  p-value: < 2.2e-16
```

Indeed, in this case, $\tilde{\beta}_1$ is upward biased since $Cor(V_1, V_2) > 0$ and $Cor(Y, V_2) > 0$.

So what does adding the control variable do to your regression? When you add a contol variable, you are effectively removing the variation in the outcome variable that is explained by that control variable, so that your estimates can be based on the variation due to the explanitory vaiable you are actually interested in. We can see this below.

I first generate a new variable, $adjY$ that "corrects" Y by removing the variation in Y that is explained by V_2, which I can do since I know the true β_2.

```
out$adjY<-out$Y-B2*out$V2

sim3<-lm(adjY~V1, data=out)
summary(sim3)
```

```
##
## Call:
## lm(formula = adjY ~ V1, data = out)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.86635 -0.65905 -0.04307  0.64875  2.78645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01470    0.03101  -0.474    0.636
## V1           5.01890    0.03103 161.759   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9807 on 998 degrees of freedom
## Multiple R-squared:  0.9633, Adjusted R-squared:  0.9632
## F-statistic: 2.617e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

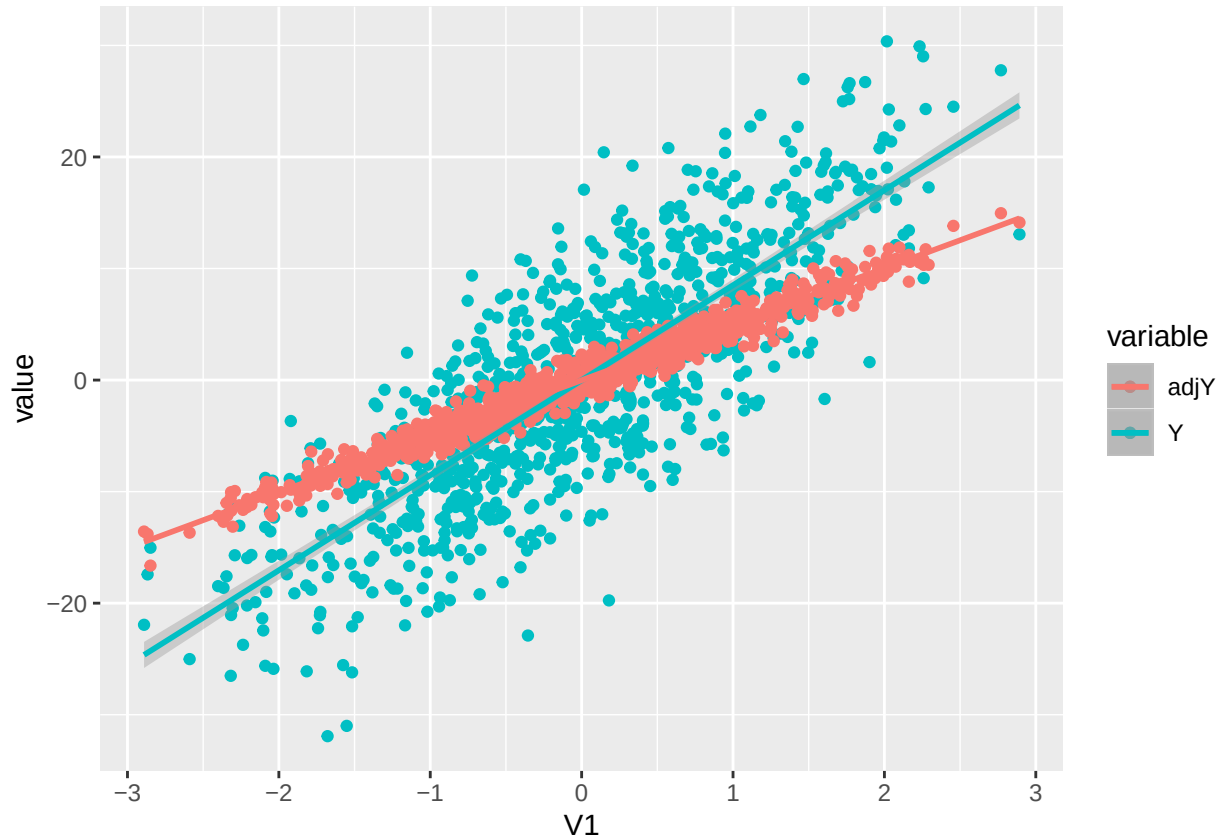
Having done this, I can now regress V_1 on $adjY$ to recover an unbiased estimate of the effect of V_1 on Y.

```
sim3<-lm(adjY~V1, data=out)
summary(sim3)
```

```
##
## Call:
## lm(formula = adjY ~ V1, data = out)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.86635 -0.65905 -0.04307  0.64875  2.78645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01470    0.03101  -0.474    0.636
## V1           5.01890    0.03103 161.759   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9807 on 998 degrees of freedom
## Multiple R-squared:  0.9633, Adjusted R-squared:  0.9632
## F-statistic: 2.617e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

We can visualize what we did in the plot below. In blue we see the "unadjusted" relationship between Y and V_1. In pink we see the "adjusted" relationship. Notice that we have $\tilde{\beta}_1 > \hat{\beta}_1$ because the adjustment is larger when the absolute value of V_1 is larger due to the positive correlation between V_1 and V_2.

```
plotted<-ggplot(out, aes(V1, y = value, color = variable)) +
    geom_point(aes(y = Y, col = "Y")) +
    geom_point(aes(y = adjY, col = "adjY"))+
    geom_smooth(method='lm', aes(y = Y, col = "Y"))+
    geom_smooth(method='lm', aes(y = adjY, col = "adjY"))

plotted
```



1.6 Beware of the kitchen sink approach

Given that adding omitted variables as controls corrects the bias generated by their omission, it can be tempting to go wild and add every possible control you can think of. In some datasets, you may have hundreds of potential observable covariates that you could add as control variables. Before embarking on this "Kitchen Sink Approach," it is important to realize that you should excercise caution in selecting your control variables. Furthermore, even in the ideal circumstances, unobservables will probably still be a problem for you if your aim is to identify a causal relationship.

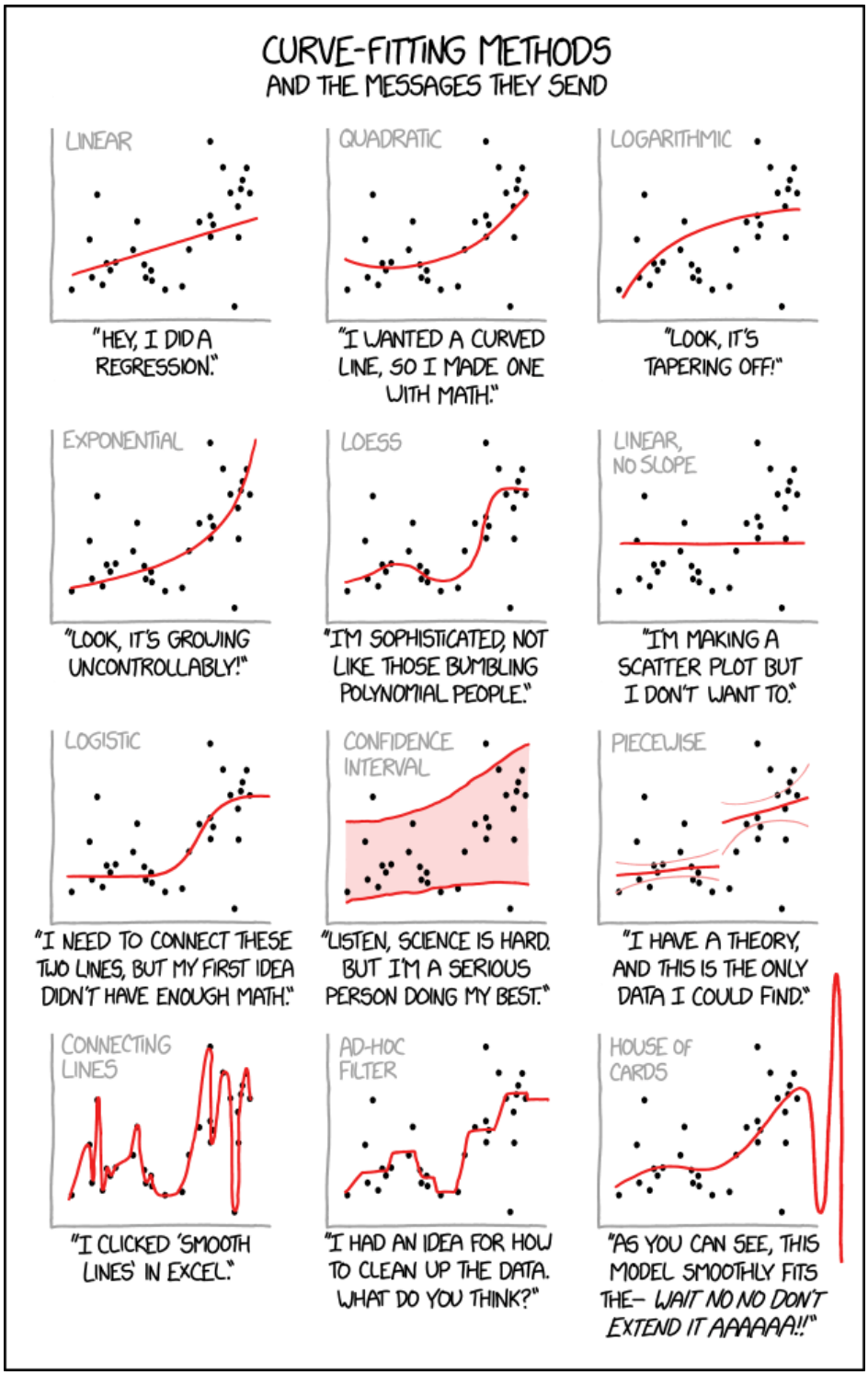
1.6.1 Caveat: Including Irrelevant Variables

What happens if we include a variable that doesn't have any effect on y (holding other variables constant)? In other words its population coefficient is zero, but we include it in our sample regression function anyways? For example, we estimate:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2$$

Even though the population regression function is $E[y|x_1] = \beta_0 + \beta_1 x_1$.

Including the irrelevant x_2 will have no effect in terms of the unbiasedness: $E[\tilde{\beta}_1] = \beta_1$. It will though affect the variance of our estimator, making it less precise: $Var(\tilde{\beta}_1) \geq Var(\hat{\beta}_1)$.



1.6.2 Caveat: Bad Controls

Even if you add a control and it does change the estimated coefficient, you may still want to carefully consider whether or not it belongs in your specification.

Bad controls are variables that are themselves outcomes of the treatment variable you are investigating. That is, bad controls might just as well be dependent variables too. Good controls are variables that we can

think of as being fixed at the time the regressor of interest was determined.

Example:

Suppose you are interested in estimating the effect of smoking during pregnancy on birthweight using observational data. You start by estimating the following regression,

$$Brthwgt_i = \beta_0 + \beta_1 cigday_i + \epsilon$$

but you are concerned that omitted variables are biasing your results. Fourtunately, your dataset includes the following additional variables: the mother's age, the mother's education level, the number of previous pregnancies, the number of prenatal doctor visits, mother's weight gain during pregnancy, and alcohol use during pregnancy.

Which of these control variables should you consider adding to your specification?

You will need to think carefuly about the causal channel that connects smoking during pregnancy to birth weight. Could smoking during pregnancy cause a difference in _____? If the answer is yes then you are probably looking at a bad control.

In this example, variables that are fixed at the start of the pregnancy are probably good controls. These would include the mother's age, the mother's education level and the number of previous pregnancies. The other variables listed could plausibly be affected by smoking. For example, women who smoke during pregnancy may be less likely to go to prenatal appointments because of shame or fear of being stygmatized by their doctor. We should thus think very carefully about whether we want to add this variable as a control.

1.7 Selection on Observable Designs: How far does this get us?

The key (untestable) assumption is that you have controled for everything that matters: you observe all the factors that affect treatment and that are correlated with the potential outcome. You are basically assuming that the treatment assignment is "as good as randomly assigned"- after you have conditioned on the controls. In other words, you are assuming that if there is any systematic selection into "treatment", this selection only depends on the observable variables you are controling for.

This is a big ask.

1.8 There is no Santa Claus: Arseneaux, Gerber and Green (2006)

Arseneaux, Gerber and Green use data from a large-scale voter "Get out the Vote" mobilization effort that randomly calls households and encourages them to vote.

Though the calling assignment is random, whether a household is actually contacted is not since many people don't answer their phones. Regressing a household's voting behavior on whether or not they were contacted can thus give biased estimates of the causal effect of encouragement on voting because of the selection into who is actually contacted. We can see this bias by comparing the estimates from a naive regression to corrected "experimental" estimates that uses the original randomization in the calling assignment to correct this bias using an instrumental variable approach that we will discuss later in this class.

Below, I replicate some of their key results using the replication files that are available on the harvard dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CTT87V

I start by loading the data (which is quite large), and estimating a simple regression with the most basic control variables.

```
library(haven)
```

```
## Warning: package 'haven' was built under R version 3.6.3
```

```r
library(lfe)
```

```
## Loading required package: Matrix
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:MASS':
##
##     select
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
agg_data<-read_dta("data/data_M1_OVB/IA_MI_merge040504.dta")
nrow(agg_data)
```

[1] 2474927

```r
##scalling the vote02 variable to remove excess 0's from tables
agg_data$vote02<-100*as.numeric(agg_data$vote02)



regols1<-felm(vote02~contact+state+comp_mi+comp_ia,agg_data)
regexp1<-felm(vote02~state+comp_mi+comp_ia|0|(contact~treat_real+state+comp_mi+comp_ia),agg_data)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```r
stargazer(regols1,regexp1, type='latex', se = list( regols1$rse, regexp1$rse))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Oct 02, 2020 - 11:08:32 AM

We see that the OLS estimates are quite different from the "experimental" estimates. The OLS estimates suggest that contacting a household raises the likelihood they vote by a statistically significant 6.2 percentage points! The "experimental" estimates however suggest that the effect is in fact 0.36 percentage points and not statistically significant... Things are not looking great for the OLS estimate. The effect it is detecting is really just the result of selection: the people that are contacted are the type of person who is more likely to vote already, even if they hadn't been contacted, thus $cor(Vote, Type) > 0$ and $cor(Contact, Type) > 0$ biasing our estimates upward.

Can OLS do better? Voter registration lists contain detailed information on voting history and demographic characteristics, allowing AAG to add a large number of controls to these regressions which I replicate below.

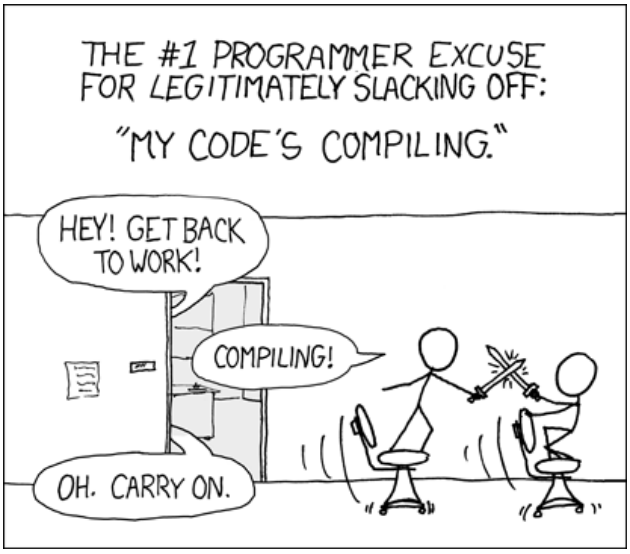
```r
regols2<-felm(vote02~contact+state+comp_mi+comp_ia+persons+age+female2+newreg+vote00+vote98+fem_miss
              |county+st_hse+st_sen,agg_data)

regexp2<-felm(vote02~state+comp_mi+comp_ia+persons+age+female2+newreg+vote00+vote98+fem_miss
              |county+st_hse+st_sen
```

Table 2:

	Dependent variable:	
	vote02	
	(1)	(2)
contact	6.207***	
	(0.306)	
state	6.671***	7.388***
	(0.347)	(0.350)
comp_mi	4.836***	4.911***
	(0.098)	(0.098)
comp_ia	6.353***	6.083***
	(0.177)	(0.178)
‘contact(fit)‘		0.360
		(0.498)
Constant	46.128***	46.081***
	(0.126)	(0.126)
Observations	1,905,320	1,905,320
R^2	0.012	0.012
Adjusted R^2	0.012	0.012
Residual Std. Error (df = 1905315)	49.486	49.491
Note:	*p<0.1; **p<0.05; ***p<0.01	

```
              |(contact~treat_real+state+comp_mi+comp_ia+persons+age+female2+newreg+vote00+vote98+fem_m
              ,agg_data)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```




```
stargazer(regols2, regexp2, type='latex', se = list(regols2$rse, regexp2$rse))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Oct 02, 2020 - 11:11:44 AM

Despite the addition of all these controls, OLS estimates are still biased. The OLS estimates suggest a statistically significant 2.7 percentage point effect as opposed to the statistically insignificant 0.5 percentage point effect we get from the experimental estimates. Despite all of these control variables, there still remains some unobservable characteristic that biases the OLS estimate.

Table 3:

	Dependent variable:	
	vote02	
	(1)	(2)
contact	2.688***	
	(0.260)	
state	2.364*	2.632**
	(1.296)	(1.296)
comp_mi	−1.793***	−1.769***
	(0.305)	(0.305)
comp_ia	−0.566	−0.667
	(0.685)	(0.686)
persons	7.001***	7.005***
	(0.064)	(0.064)
age	0.346***	0.346***
	(0.002)	(0.002)
female2	−1.174***	−1.173***
	(0.062)	(0.062)
newreg	5.456***	5.458***
	(0.111)	(0.111)
vote00	37.090***	37.092***
	(0.074)	(0.074)
vote98	21.657***	21.659***
	(0.082)	(0.082)
fem_miss	−32.082***	−32.113***
	(0.241)	(0.241)
‘contact(fit)‘		0.513
		(0.420)
Observations	1,905,320	1,905,320
R^2	0.288	0.288
Adjusted R^2	0.288	0.288
Residual Std. Error (df = 1905055)	42.001	42.002
Note:	*p<0.1; **p<0.05; ***p<0.01	