ECON 1190: Applied Econometrics 2: Module 1: Omitted Variable Bias

Claire Duquennois

Module 1: Regressions,causality and bias

- Regression and causality
- No Causation Without Manipulation
- The Rubin Causal Model
- The Conditional independence assumption
- Omitted variable bias
- The kitchen sink approach
- How far does this get us? AGG(2006)

Regression and Causality

As long as certain trivial conditions are satisfied, you can always run a linear regression. This is fine as long as you interpret the results appropriately. We may be interested in the relationship between x and y for the purposes of:

- Description-What is the relationship between x and y?
- Prediction-Can we use x to create a good forecast of y?
- Causation-What happens to y if we manipulate x?

Causation... this is where things get tricky...

But First: What Regressions can do!

In the social sciences, we tend to focus on relationships that hold "on average," or "in expectation."

The *Conditional Expectation Function*: Given a particular value of x, where is the distribution of y centered?

$$E[y_i|x_i] = h(x_i)$$

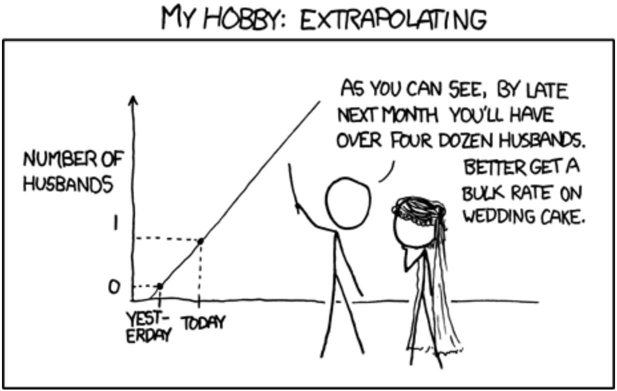
Linear Regression

If the CEF is linear, regressing y_i on x_i estimates the CEF.

If the CEF is not linear, we still often use linear regression because:

- Computationally tractable
- Well understood and desirable properties
- Provide the best linear approximation of the CEF even when it is non-linear (just don't try to extrapolate far beyond the support of x_i)
- can often adjust variables with logs and quadratics to linearize the relationship

Linear Regression



Estimating the CEF

Let

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

- Run a linear regression of y_i on x_i
- Get estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the true population β_0 and β_1
- Calculate $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, the predicted value for y_i given x_i, such that

$$\hat{y}_i = E[y_i|x_i], \text{ the CEF.}$$

If you are interested in description or prediction, this is fine and we can end the class here!

Application: Regression for Prediction

Who might be interested in using regressions for prediction?

Suppose you are a bank interested in predicting customer's ability to repay student loans. You have a subset of CPS data on earnings and the number of years spent in education.

You estimate the following on working age adults (22+):

$$Income_i = \beta_0 + \beta_1 Schooling_i + \epsilon_i$$

Application: Regression for Prediction

```
mydata<-read.csv("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
```

Interpret your results.

Application: Regression for Prediction

```
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
```

So an extra year of education **predicts** earnings that are 8,054 USD higher (since $\hat{\beta}_1 = 8054$).

Application: Regression for Prediction

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.

Application: Regression for Prediction

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:

$$Income_i = \beta_0 + \beta_1 CollGrad_i + \epsilon_i \tag{1}$$

Application: Regression for Prediction

```r
mydata$collgrad<-NA
mydata$collgrad[mydata$edu<16]<-0
mydata$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
```

Interpret your results.

Application: Regression for Prediction

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

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

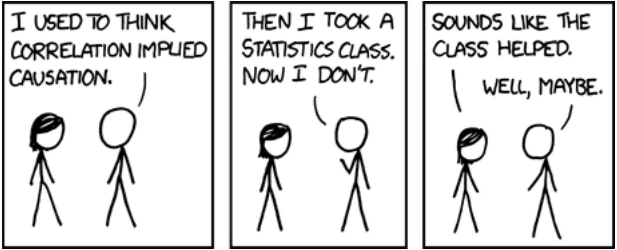
```
##
## Call:
## lm(formula = inctot ~ collgrad, data = mydata[mydata$edu >= 12 &
##     mydata$age > 22, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -91324  -31425  -11433   13518 1054675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    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
```

$\hat{\beta}_1 = 44,842$, so having a four year college degree **predicts** earnings that are $44,842 higher.

Application: Regression for Prediction

The key point: we are not saying that the college degree **caused** higher earnings, but it does **predict** higher earnings. For many applications, prediction is enough.

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



"No Causation Without Manipulation"

What if we are interested in causal effects?

It was easy to estimate the relationship between income and schooling. As illustrated in the application, I estimated

$$Income_i = \beta_0 + \beta_i Schooling_i + \epsilon$$

and was able to recover the conditional expectation function

$$E[Income_i|Schooling_i] = Inc\hat{o}me_i = \hat{\beta}_0 + \hat{\beta}_1 Schooling_i$$

BUT this only tells us how income and schooling co-vary. This **DOES NOT** tell us what would happen to income if there was an **"exogenous"** change in schooling.

What is the difference?

Here, schooling is **“endogenously”** determined.

Who is most likely to select into schooling?

What is the difference?

Here, schooling is **"endogenously"** determined. For example:

- those who expect to benefit the most select into schooling.
- those with the highest family incomes select into schooling.

A regression coefficient estimated using data on **endogenous** schooling choices will not correspond to the effects of an **exogenous** change in schooling.

To estimate the **causal** effect, we will need to identify some type of **manipulation** that created an **exogenous** change in schooling.

A note on interpretation

It in NOT the case that the endogenous estimate is *wrong* and the exogenous estimate is *right*. They are simply measuring different things and should be interpreted accordingly.

Regarding our estimates using the endogenous CPS data:

CORRECT:

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

INCORRECT:

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

The Rubin Causal Model

Two roads diverged in a yellow wood,
And sorry I could not travel both
–Robert Frost

To understand causal inference, it is helpful to think about how a unit has different potential outcomes depending on it's treatment status.

The Rubin Causal Model

Let D_i be a binary treatment variable that could affect Y_i which is the outcome Y of observation i. Each unit faces two potential outcomes:

$$Y_i = \begin{cases} Y_i(1) & \text{if } D_i = 1 \ (\textit{i is in the treatment group}) \\ Y_i(0) & \text{if } D_i = 0 \ (\textit{i is in the control group}) \end{cases}$$

The problem: Unobserved **counterfactuals**. We will never observe both $Y_i(1)$ and $Y_i(0)$.

Example: Does going to college cause higher earnings?

Let the treatment, D_i be going to college. Each high school graduate faces two potential outcomes:

$$i\text{'s potential outcomes} = \begin{cases} earn_{i,col} & \text{if } i \text{ goes to college } (\textit{treatment}) \\ earn_{i,nocol} & \text{if } i \text{ no college } (\textit{control}) \end{cases}$$

We can conceive of both $earn_{i,col}$ and $earn_{i,nocol}$ (but will only ever observe one or the other).

The treatment is potentially manipulable: we can imagine a policy or intervention that could make either of these values observable.

“No Causation without Manipulation” (2)

Can you conceptualize both $Y_i(1)$ and $Y_i(0)$ for the same unit?

If no: D does not correspond to a potentially manipulable treatment.

- We need to further define the problem.

Example: Does being a woman cause lower earnings?

"No Causation without Manipulation" (2)

Can you conceptualize both $Y_i(1)$ and $Y_i(0)$ for the same unit?

If no: D does not correspond to a potentially manipulable treatment.

- We need to further define the problem.

Example: Does being a woman cause lower earnings?

It is not possible for me to imagine some intervention that would reveal what my earnings outcome would have been if I was a man.

We know that being a woman ***predicts*** lower earnings, but the causal question as posed is ill defined.

Causal Effects

The difference in i's observed outcome and the counterfactual is τ_i, the causal effect of treatment D on the outcome Y for unit i. So

$$Y_i(1) - Y_i(0) = \tau_i$$

Note: the treatment effect is relative and specific to observation i.

But how can we identify τ_i if we never observe both $Y_i(1)$ and $Y_i(0)$, the realized outcome and the counterfactual, for a given unit?

The 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, therefore, it is impossible to observe τ_i, the effect for unit i of the treatment on it's outcome, Y_i. (Holland 1986)

So, are we doomed?

The 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, therefore, it is impossible to observe τ_i, the effect for unit i of the treatment on it's outcome, Y_i. (Holland 1986)

So, are we doomed?

No! Though we can't identify τ_i at the unit level, we can identify the *Average Causal Treatment Effect* (ATE)

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

with the right research design, we can recover $\bar{\tau}$.

Getting to an econometric specification

Each observation faces two potential outcomes:

- $D_i = 1$ if i is treated and i's outcome is then $Y_i(1)$
- $D_i = 0$ if i is not treated and i's outcome is then $Y_i(0)$

We can summarize this as

$$Y_i = Y_i(0) + (Y_i(1) - Y_i(0))D_i$$

Getting to an econometric specification

Suppose a constant causal treatment effect such that $\tau = Y_i(1) - Y_i(0)$: The treatment effect is the same for all observations. Then

$$\begin{aligned} Y_i &= Y_i(0) + (Y_i(1) - Y_i(0))D_i \\ &= Y_i(0) + \tau D_i \end{aligned}$$

- if $D_i = 0$, then $Y_i = Y_i(0)$
- if $D_i = 1$, the $Y_i = Y_i(0) + \tau$

Getting to an econometric specification

Adding and subtracting the average outcome for untreated observations, $E[Y_i(0)]$, and reorganizing gives the following:

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

where $\alpha = E[Y_i(0)]$, and η_i is the random part of $Y_i(0)$ since $\eta_i = Y_i(0) - E[Y_i(0)]$.

Causal identification

The expected outcomes of someone who is treated is

$$E[Y_i(1)] = \alpha + \tau + E[\eta_i | D_i = 1]$$

and someone without treatment is

$$E[Y_i(0)] = \alpha + E[\eta_i | D_i = 0]$$

so that the difference between these outcomes can be broken down into

$$E[Y_i(1)] - E[Y_i(0)] = \underbrace{\tau}_{\text{causal treatment effect}} + \underbrace{E[\eta_i | D_i = 1] - E[\eta_i | D_i = 0]}_{?}$$

What is the second term? $\Rightarrow$ Top Hat

The Conditional Independence Assumption

The expected outcomes of someone with and someone without treatment is then given by

$$E[Y_i(1)] = \alpha + \tau + E[\eta_i | D_i = 1]$$
$$E[Y_i(0)] = \alpha + E[\eta_i | D_i = 0]$$

so that the difference between these outcomes can be broken down into

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

Selection Bias

So if $E[\eta_i|D_i = 1] \neq E[\eta_i|D_i = 0]$ and I run

$$Y_i = \alpha + \tilde{\tau} D_i + \eta_i$$

the estimated $\tilde{\tau} \neq \tau$ because of selection bias since

$$\tilde{\tau} = \tau + E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]$$

Selection Bias

Selection bias will occur if treatment is not random.

If treatment is not random, those who would select into treatment have a different expected outcome, even absent treatment, than those who would not select into treatment so

$$E[Y_i(0)|D_i = 1] \neq E[Y_i(0)|D_i = 0],$$

In English:

- even absent treatment (since we are comparing the untreated outcomes $Y_i(0)$) the average for those who would select into treatment (so who have $D_i = 1$) is not the same as for those who do not select into treatment (so who have $D_i = 0$)

Example

I naively use my observational CPS data and estimate

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

If I want to estimate τ, the **causal** effect of a college degree on earnings, this estimate, $\tilde{\tau}$ will be biased: $E[\tilde{\tau}] \neq \tau$.

Why?

Example

I naively use my observational CPS data and estimate

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

If I want to estimate τ, the **causal** effect of a college degree on earnings, this estimate, $\tilde{\tau}$ will be biased: $E[\tilde{\tau}] \neq \tau$.

Why?

<u>Selection bias:</u> If people who receive college degrees would have had higher earnings even without the degree,

$$E[Y_i(0)|D_i = 1] > E[Y_i(0)|D_i = 0]$$

so

$$E[\eta_i|D_i = 1] > E[\eta_i|D_i = 0]$$

and

$$\tilde{\tau} = \tau + E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]$$

The Conditional Independence Assumption

The Conditional Independence Assumption: conditional on observed characteristics, X_i, selection bias disappears.

- If CIA holds, once I control for X_i, treatment is as good as randomly assigned:

$$E[Y_i(0)|X_i, D_i = 1] = E[Y_i(0)|X_i, D_i = 0]$$

and our comparisons have a causal interpretation.

- In other words, if I can perfectly control for **all** the X_i characteristics that generate selection, I can recover an unbiased estimate of the causal effect τ.

CIA in Regressions

I run $Y_i = \alpha + \tilde{\tau} D_i + \eta_i$, but $E[\tilde{\tau}] \neq \tau$ due to selection bias.

I add observed covariates X_i' to the regression:

- I can decompose η_i: $\eta_i = X_i'\gamma + \nu_i$ into
  - $X_i'\gamma$: a part explained by the covariates X_i'
  - an unexplained error ν_i
- **If the CIA assumption holds given the added covariates** X_i then $E[\nu_i | D_i = 1] = E[\nu_i | D_i = 0]$
- The difference between the treated and untreated group is

$$
\begin{aligned}
&E[Y_i(1)|X_i] - E[Y_i(0)|X_i] \\
&= (\alpha + \tau + X_i'\gamma + E[\nu_i|D_i = 1]) - (\alpha + X_i'\gamma + E[\nu_i|D_i = 0]) \\
&= \tau + E[\nu_i|D_i = 1] - E[\nu_i|D_i = 0] \\
&= \tau
\end{aligned}
$$

and we can interpret $\hat{\tau}$ as the causal effect of interest.

Example:

If I estimate $Earnings_i = \tilde{\alpha} + \tilde{\tau} college_i + \epsilon_i$ we saw that $E[\tilde{\tau}] \neq \tau$ due to selection bias.

Suppose CIA holds if I condition on a student's household income. (ie. if I control for student household income, which students complete college is as good as randomly assigned) .

Then

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

and $E[\hat{\tau}] = \tau$: a college degree causes earnings to increase by $\hat{\tau}$ USD.

WARNING: This is a big assumption, that often does not hold. Do you think it holds in this example?

Omitted Variable Bias

Suppose the true model is given by

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

but I failed to include x_{2i} and instead estimated

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

If there is a relationship between x_{1i} and x_{2i} such that

$$x_{2i} = \rho_1 + \rho_2 x_{1i} + \varepsilon_i$$

we can substitute this into the first equation and by rearranging,

$$Y_i = \underbrace{(\beta_0 + \beta_2\rho_1)}_{\tilde{\beta}_0} + \underbrace{(\beta_1 + \beta_2\rho_2)}_{\tilde{\beta}_1} x_{1i} + \underbrace{(\beta_2\varepsilon_i + \nu_i)}_{\epsilon_i},$$

show that

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

Omitted variable bias

$$\tilde{\beta}_1 - \beta_1 = \underbrace{\beta_2 \rho_2}_{\text{bias}}$$

We can thus sign the bias by signing β_2, the covariance between x_{2i} and Y_i, and signing ρ_2, the covariance between x_{2i} and x_{1i}.

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

Example: Using CPS data

I am interested in how health relates to income. Using my CPS sample of working age adults I estimate

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

where heath is a respondents subjective assessment of their health with 1 being very healthy and 5 being very unhealthy.

Example: Using CPS data

```r
mydata <- mydata %>% rename(bad_health = health)
reghealth<-lm(inctot~bad_health,mydata)
summary(reghealth)
```

```
##
## Call:
## lm(formula = inctot ~ bad_health, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -59413  -32893  -15716   11198 1103107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65599.6     2414.9  27.164   <2e-16 ***
## bad_health   -8176.8      991.9  -8.243   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73960 on 4998 degrees of freedom
## Multiple R-squared:  0.01341,    Adjusted R-squared:  0.01322
## F-statistic: 67.95 on 1 and 4998 DF,  p-value: < 2.2e-16
```

Interpret.

Example: Using CPS data

How might the omission of age be biasing these estimates?

$\Rightarrow$ Top Hat

Example: Using CPS data

How might the omission of age be biasing these estimates?

- $cov(badhealth_i, age_i) > 0$
- $cov(income_i, age_i) > 0$
- $\Rightarrow$ upward bias.

Example: Using CPS data

```r
reghealth2<-lm(inctot~bad_health+age ,mydata)
summary(reghealth2)
```

```
##
## Call:
## lm(formula = inctot ~ bad_health + age, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -84573  -31695  -12625   11680 1101885
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    28903       3771   7.665 2.13e-14 ***
## bad_health    -11489       1012 -11.355  < 2e-16 ***
## age             1066         85  12.541  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72830 on 4997 degrees of freedom
## Multiple R-squared:  0.04352,    Adjusted R-squared:  0.04314
## F-statistic: 113.7 on 2 and 4997 DF,  p-value: < 2.2e-16
```

Example: Using CPS data

How might the omission of schooling be biasing these estimates?

$\Rightarrow$ Top Hat

Example: Using CPS data

How might the omission of schooling be biasing these estimates?

- $cov(badhealth_i, schooling_i) < 0$
- $cov(income_i, schooling_i) > 0$
- $\Rightarrow$ downward bias.

Example: Using CPS data

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

```
##
## Call:
## lm(formula = inctot ~ bad_health + age + edu, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -117272  -28392   -9807   12671 1077286
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -82442.91    6501.39 -12.681  < 2e-16 ***
## bad_health   -6315.25    1003.36  -6.294 3.36e-10 ***
## age            953.42      81.79  11.657  < 2e-16 ***
## edu           7540.90     365.73  20.619  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69920 on 4996 degrees of freedom
## Multiple R-squared:  0.1185, Adjusted R-squared:  0.118
## F-statistic: 223.9 on 3 and 4996 DF,  p-value: < 2.2e-16
```

Example: Presenting results

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

Table 2: Income and bad health

	Dependent variable:		
	inctot		
	(1)	(2)	(3)
bad_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

Check out jakeruss.com/cheatsheets/stargazer/

Example: Using CPS data

So is our estimate of β_1 in column 3 the causal effect of bad health on income?

Does the CIA hold?

Conditional on age and schooling, is subjective health as good as randomly assigned?

Building intuition: A simulation

Suppose the data generating process (DGP) 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 $Cor(V_1, V_2) = 0.5$.

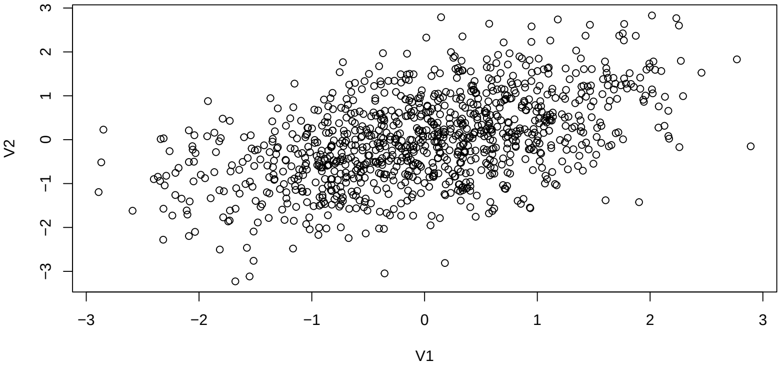
```
library(MASS)
library(ggplot2)

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
```

Building intuition: A simulation

```
plot(out)
```



Building intuition: A simulation

I add an error term for each observation and then simulate the true DGP with $\beta_1 = 5$ and $\beta_2 = 7$.

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

B1<-5
B2<-7

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

I can now estimate the correct model and an under-specified model:

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

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

Building intuition: A simulation

```
stargazer(sim1,sim2, type="latex", header=FALSE,
          title="Omitted Variable Bias Simulation", omit.stat=c("f", "ser"))
```

Table 3: Omitted Variable Bias Simulation

	Dependent variable:	
	Y	
	(1)	(2)
V1	5.007***	8.519***
	(0.036)	(0.195)
V2	7.023***	
	(0.036)	
Constant	−0.015	−0.015
	(0.031)	(0.195)
Observations	1,000	1,000
R^2	0.991	0.657
Adjusted R^2	0.991	0.656
Note:	*p<0.1; **p<0.05; ***p<0.01	

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

Building intuition: A simulation

What is adding the V_2 control doing? How does it change the V_1 coefficient?

- Adding V_2 in the regression removes the variation in the outcome variable that is explained by that control variable.
- The estimates can now be based on the variation due to the explanatory variable you are actually interested in.

To see this, I generate *adjY* that "corrects" Y by removing the variation in Y that is explained by V_2. (I can do this since I know the true β_2.)

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

Building intuition: A simulation

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

stargazer(sim1,sim2,sim3, type="latex", header=FALSE,
          title="Omitted Variable Bias Simulation 2", omit.stat=c("f", "ser"))
```

Table 4: Omitted Variable Bias Simulation 2

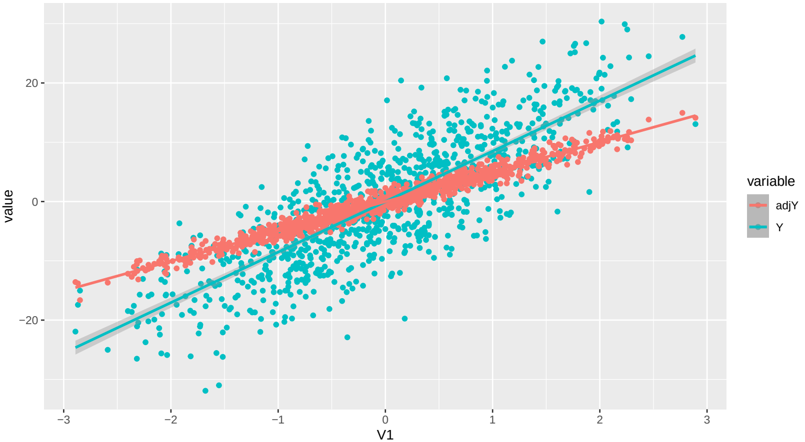
	Dependent variable:		
	Y		adjY
	(1)	(2)	(3)
V1	5.007***	8.519***	5.019***
	(0.036)	(0.195)	(0.031)
V2	7.023***		
	(0.036)		
Constant	−0.015	−0.015	−0.015
	(0.031)	(0.195)	(0.031)
Observations	1,000	1,000	1,000
R^2	0.991	0.657	0.963
Adjusted R^2	0.991	0.656	0.963
Note:	*p<0.1; **p<0.05; ***p<0.01		

Building intuition: A simulation

```
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
```

```
## `geom_smooth()` using formula = 'y ~ x'`geom_smooth()` using formula = 'y ~ x'
```



The kitchen sink

Adding more controls is not always better.

- Irrelevant variables
- Bad controls

Moreover, without a carefully thought out research design, omitted variable bias will still be a problem.

Caveat: Including irrelevant variables

Suppose I estimate

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

even though the true model is actually

$$E[y|x_1] = \beta_0 + \beta_1 x_1$$

- Including x_2 will not bias our estimation: $E[\tilde{\beta}_1] = \beta_1$.
- The variance of our estimator will be less precise: $Var(\tilde{\beta}_1) \geq Var(\hat{\beta}_1)$.

Caveat: Bad Controls

Some control variables could themselves be outcomes of the treatment you are evaluating.

Good controls are variables that were fixed at the time treatment was determined.

Bad Controls: Example

$$Wages = \beta_0 + \beta_1 Schooling$$

Assume Schooling is randomly assigned and $\beta_1 > 0$.

Then someone estimates

$$Wages = \beta_0 + \beta_1 Schooling + \beta_2 Occupation$$

and now $\beta_1 = 0$.

So does schooling have an effect or not?

Bad Controls: Example

$$Wages = \beta_0 + \beta_1 Schooling$$

Assume Schooling is randomly assigned and $\beta_1 > 0$.

Then someone estimates

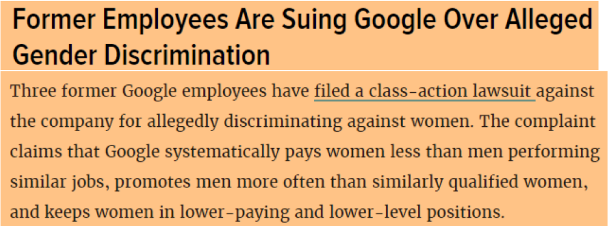
$$Wages = \beta_0 + \beta_1 Schooling + \beta_2 Occupation$$

and now $\beta_1 = 0$.

So does schooling have an effect or not?

Yes! Most of the effect of schooling may very well come from occupation! It is a bad control.

Bad Controls: Example



Former Employees Are Suing Google Over Alleged Gender Discrimination

Three former Google employees have filed a class-action lawsuit against the company for allegedly discriminating against women. The complaint claims that Google systematically pays women less than men performing similar jobs, promotes men more often than similarly qualified women, and keeps women in lower-paying and lower-level positions.

Bad Controls: Example

After the *New York Times* detailed the employee spreadsheets on Friday, Google spokesperson Gina Scigliano told Gizmodo that its own data shows, when you take "location, tenure, job role, level and performance" into account, that "women are paid 99.7% of what men are paid at Google." Scigliano described the *Times* story as "extremely flawed."

Bad Controls: Example

Including occupation dummies in a wage equation is LITERALLY the textbook example of what not to do (Angrist and Pischke, Mastering 'Metrics: The Path from Cause to Effect p 215)

Beware Bad Control

Perhaps more controls are the answer. Why not control for occupation for example? Many data sets that report earnings also classify workers' jobs, such as whether they're employed as a manager or laborer. Surely occupations are strong predictors of both schooling and earnings, possibly capturing traits that distinguish Mick and Johan from more average Joes. By the logic of OVB, therefore, we should control for occupation, a matter easily accomplished by including dummy variables to indicate the types of jobs held.

In spite of the fact that occupation is strongly correlated with both schooling and wages, occupation dummies are *bad controls* in a wage equation. The fact that Master Joshway works today as a professor and not a busboy (as he once did) is in part a reward for his extravagant schooling. It's a mistake to eliminate this benefit from our calculation by comparing only professors or busboys when attempting to quantify the economic value of schooling. Even in a world where all professors earn a uniform one million dollars a year (may it soon come to pass) and all busboys earn a uniform $10,000, an experiment that randomly assigns schooling would show that schooling raises wages. The channel by which wages are increased in this notional experiment is the shift from lowly busboyhood to elevated professorness.

Bad Controls: Top Hat Example

You are interested in smoking's effect on birth-weight. You estimate

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

but are concerned there may be important omitted variables.

You data includes information on the following: 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?

$\Rightarrow$ Top Hat

How far do controls get us?

The key (untestable) assumption is that you have controlled for everything that matters.

You are assuming that treatment assignment is "as good as randomly assigned"- after you have conditioned on the controls.

You are assuming that if there is any systematic selection into "treatment", it only depends on the observable variables you are controlling for.

These are VERY STRONG assumptions (that often do not hold).

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

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

Evaluate a "Get out the Vote" mobilization:

- Who gets called ($Call_i$) is random
- Who answers the call ($Contact_i$) is not

Will the following approach give us an unbiased estimate of the causal effect of being contacted on voting?

$$Vote_i = \alpha + \tau Contact_i + \epsilon_i$$

There is no Santa Claus: AGG (2006)

```
library(haven)
```

```
## Warning: package 'haven' was built under R version 4.1.2
```

```
library(here)
```

```
## Warning: package 'here' was built under R version 4.1.2
```

```
## here() starts at C:/Users/johnh/Dropbox/ECON1190_Causal/econ1190
```

```
library(lfe)
library(dplyr)
```

There is no Santa Claus: AGG (2006)

Replicating results of columns 1 of p.49 and p.50:

```
agg_data<-read_dta("IA_MI_merge040504.dta")
nrow(agg_data)
```

[1] 2474927

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

#note: basic controls are included since the randomization happened at the state level
#and to distinguish between competitive and un-competitive races in each state.
regols1<-felm(vote02~contact+state+comp_mi+comp_ia,agg_data)

#Getting an unbiased estimate using insturumental variables approach
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
```

There is no Santa Claus: AGG (2006)

Replicating results of columns 1 of p.49 and p.50:

```
stargazer(regols1,regexp1, type='latex', se = list(regols1$rse,regexp1$rse),
          header=FALSE,  title="AGG replication 1",omit.stat=c("f", "ser"), single.row = TRUE)
```

Table 5: AGG replication 1

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

There is no Santa Claus: AGG (2006)

Our OLS estimator it not doing so good: $\tilde{\tau} > \tau$.

Why?

There is no Santa Claus: AGG (2006)

Our OLS estimator it not doing so good: $\tilde{\tau} > \tau$.

Why?

- the people that are contacted are the type of person who is more likely to vote already
- $cor(Vote, Type) > 0$ and $cor(Contact, Type) > 0$ biasing our estimates upward.

Can OLS do better? AGG have lots of controls in their data.

There is no Santa Claus: AGG (2006)

Replicating results of columns 2 of p.49 and p.50:

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

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

There is no Santa Claus: AGG (2006)

Replicating results of columns 2 of p.49 and p.50:

```
stargazer(regols2,regexp2, type='latex', se = list(regols2$rse,regexp2$rse),
          header=FALSE,  title="AGG replication 2",omit.stat=c("f", "ser"), single.row = TRUE)
```

Table 6: AGG replication 2

	Dependent variable:	
	vote02	
	(1)	(2)
contact	2.688*** (0.260)	
state	2.364* (1.296)	2.632** (1.296)
comp_mi	−1.793*** (0.305)	−1.769*** (0.305)
comp_ia	−0.566 (0.685)	−0.667 (0.686)
persons	7.001*** (0.064)	7.005*** (0.064)
age	0.346*** (0.002)	0.346*** (0.002)
female2	−1.174*** (0.062)	−1.173*** (0.062)
newreg	5.456*** (0.111)	5.458*** (0.111)
vote00	37.090*** (0.074)	37.092*** (0.074)
vote98	21.657*** (0.082)	21.659*** (0.082)
fem_miss	−32.082*** (0.241)	−32.113*** (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
Note:		*p<0.1; **p<0.05; ***p<0.01

There is no Santa Claus: AGG (2006)

Our OLS estimates are still biased. Even with all these controls, $\tilde{\tau} > \tau$.

Unless you had a variable that told you if the person is the type to answer and talk to an unknown caller about voting, the kitchen sink approach will not solve the OVB problem.

Discussion questions for Washington (2008)

- What is her research question? Is it interesting? Are the effects she detects meaningful?
- Does she have an experimental design? How is she getting exogenous variation with which to identify causal effects?
- Do you believe that the CIA assumption holds? (once she controls for key observables, her explanatory variable is as good as randomly assigned) What are some concerns she has and discusses in the paper?
- Are you convinced she has identified a causal relationship?