Political Science 15
Introduction to Research in Political Science
Lecture 6a: Fighting Endogeneity with Multivariate Regression

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Announcements

Midterm complete – good job!

- Midterm review video will be posted on Monday.

Plan for the Lecture 6 on Multivariate Regression

1. Multiple predictors
2. Causality as a problem
3. Why use multivariate regression?
4. Interpreting multivariate regression
5. Visualizing multivariate regression
6. Prediction in multivariate regression
7. Goodness of fit and irrelevant variables
8. Model specification

Housekeeping

Progress:

- ✓ Problem Set 1
- ✓ Problem Set 2
- ✓ Problem Set 3
- ✓ Midterm
- □ Problem Set 4
- □ Problem Set 5
- □ Final

Looking forward to the rest of the course:

- Multivariate regression
- Multivariate regression in real research
- Types of data and variables
- Experiments
- Data science

Bivariate regression

We talked about the bivariate regression model:

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

Or, in terms of our "best guess" or predicted value $\hat{Y}_i$:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Multiple predictors

But often we think there are more factors affecting our outcome variable. We need more independent (AKA predictor, explanatory) variables, e.g.:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$$

Why?

- to better guess/predict Y_i
- to "control for" or "remove the effect of" one variable (say Z_i) when interpreting the coefficient on the other

Usually we like to call the independent variables X_1, X_2, X_3, etc. That is:

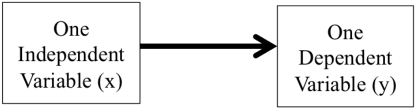
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

Surpisingly, we can calculate the β's that make this model fit best in the least-squares sense, i.e.

$$\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\beta}_3$$

though the mathematics for this is beyond this course.

Single-Variable Regression:

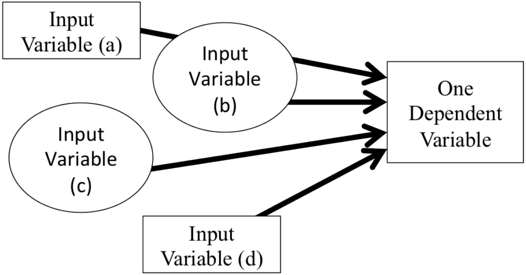


One Independent Variable (x)

One Dependent Variable (y)

Output: y = *f*(x)

Multiple Regression:



Input Variable (a)

Input Variable (b)

Input Variable (c)

Input Variable (d)

One Dependent Variable

Output: y = *f*(a, b, c, d)

The causal regression interpretation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

This is a multivariate regression model. We will still think of this as a "prediction machine" that tells us $\hat{Y}_i$ for some choice of X_1, X_2, X_3.

But we can see how this differs from a causal interpretation of the β's by pretending for a moment.

Let's do so, first with just one covariate:

- *Suppose* we really did want to think of β_1 as causal in:

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

- In order to causally interpret $\hat{\beta}_1$, we would have to insist that the exogeneity assumption holds. Check: what does that mean?
  - we'd have to insist that things influencing Y_i other than X_i (i.e. in the error term) are not correlated with X_i, or
  - put differently, nothing is both correlated with X_i and influences Y_i
  - put differently, no confounders

The causal regression interpretation

Our model was:

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

If we want to make the claim the β meaningfully estimates the causal effect of X on Y, then the **exogeneity assumption** needs to hold. Check: what is another way of writing the exogeneity assumption?

$$cor(X, \epsilon) = 0$$

(Or equivalently that $cov(X, \epsilon) = 0$.)

A tricky but critical point:

- You cannot *check* whether $cor(\epsilon, X) = 0$ – it's an **assumption**!
- Because the $\hat{\epsilon}$ you get is estimated from the data.
- You don't get to observe what is *really* in the error term ϵ_i.
- So how do you achieve $cor(\epsilon, X) = 0$? Research design!
  - You can randomize the assignment of X_i.
  - (More advanced.) Use a "natural experiment" to argue that X_i is "as good as" if it was randomly assigned.

Recap of key terms

Endogeneity

- We say that X_i is endogenous if it is partially related to something that influences the outcome.

Exogeneity

- By contrast, if X_i is uncorrelated with the error term ($cor(X, \epsilon) = 0$), we say that it is exogenous.

Confounder

- A variable that is correlated both with the explanatory variable X_i and the outcome variable Y_i.

Omitted variable bias

- The bias that results from not including a confounder in our model.

Why multivariate regression?

Key idea behind multivariate regression

If we can include variables that would have been confounders in the regression, we take them out of the ϵ, and avoid omitted variable bias.

Suppose we are interested in the effect of X_1 on economic performance. So we might run:

$$growth = \beta_0 + \beta_1 X_1 + \boxed{\epsilon}$$

But some X_2 correlated with X_1 is also influencing Y (i.e. X_2 is a confounder). So we instead run:

$$growth = \beta_0 + \beta_1 X_1 + \boxed{\beta_2 X_2 + \eta}$$

Structure of multivariate regression model

Bivariate model

X_1 is the *(only)* **independent variable** affecting our outcome

β_1 is the **slope coefficient** on X_1

ϵ is the **error term**, or everything that we haven't captured in our model

$$growth = \beta_0 + \beta_1 X_1 + \epsilon$$

Multivariate model

X_1 and X_2 are independent variables

β_2 is the **slope coefficient** on X_2

η is the **error term**, which again represents everything we haven't captured in this model

$$growth = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \eta$$

Why multivariate regression?

$$growth = \beta_0 + \beta_1 X_1 + \boxed{\epsilon} \qquad \textbf{(bivariate model)}$$

$$growth = \beta_0 + \beta_1 X_1 + \boxed{\beta_2 X_2 + \eta} \qquad \textbf{(multivariate model)}$$

- X_2 is a potential confounder, so including it in the model allows β_2 to "soak up" the effect of X_2.
- Had we not, our ϵ in the first model would include a function of X_2, and we said that X_2 correlated to X_1. You can think about it like X_2 was hanging out in the error term before we put it in the model, and that was a problem.
- As a result β_1 in our first model is wrong for the causal effect: it soaks up some of the effect that should have been attributed to X_2.
- Including both X_1 and X_2 in the model, we figure out what part of Y is due to X_1 and X_2, and hope nothing correlated with X is left in the new error term η.

Attempting to avoid Omitted Variable Bias

$$growth = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \eta$$

How can we understand this regression?

We say that:

- this “controls for” the role of X_2
- this “soaks up the effect of X_2” so that β_1 more accurately captures the effect of X_1
- X_1 and X_2 “compete” to explain *growth*, so the one which is the better predictor will “win out”
- β_1 tells us how X_1 relates to *growth* if we could “hold X_2 constant”

How do I interpret multivariate regression?

- In a bivariate regression, we read "a 1-unit shift in X is associated with a $\hat{\beta}_1$ shift in Y".
- In a multivariate regression, the interpretation is not that different. (See textbook p. 131, digital version p. 197.)
- We can make the same interpretation, except we must add:
  - "controlling for the other covariates (Xs or IVs)", or
  - "holding other factors constant", or
  - (if you want to be fancy) "*ceteris paribus*" (which means "everything else being equal")
- In a sense, "controlling for" is implying you're holding these other variables at their means. But you could also set them to specific values.

Political Science 15
Introduction to Research in Political Science
Lecture 6b: Example of Multivariate Regression

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Ethnic fractionalization and growth

Does “ethnic fractionalization” (`ethfrac`) slow economic growth (`growth`)?

Warning: This phrasing makes “slow” a causal word, so we will be treading carefully and come back around to a safe interpretation.

Let’s hypothesize that `ethfrac` does slow `growth`:

- Check. How then would you expect `ethfrac` and `growth` to be related in your data in terms of correlation, covariance, and a regression coefficient?

Let’s try the regression model:

$$growth_i = \beta_0 + \beta_1 ethfrac_i + \epsilon_i$$

- *ethfrac* is a measure of ethnic fractionalization. It is the probability that two randomly selected people are from different groups, ranging from 0 to 1
  - `mean(ethfrac) = 0.45`
- *growth* is rate of growth in GDP per capita by year
  - `mean(growth) = 3.78`

Bivariate model

```
> summary(lm(rgdpgrowth ~ ethfrac, data = dat))
Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)   4.3121     0.2306   18.703    <2e-16 ***
ethfrac      -0.9517     0.4373   -2.176    0.0296 *
```

Let's review how to interpret this.

- The intercept $\hat{\beta}_0 = 4.31$. The expected *growth* rate when *ethfrac* $= 0$ (i.e. in perfectly homogeneous societies) is 4.31%.
- The coefficient on *ethfrac* is $\hat{\beta}_1 = -0.952$, meaning a 1-unit increase in *ethfrac* is associated with a predicted *decrease* of the GDP growth rate of 0.952 percentage points.
  - Check. What does a 1-unit increase in *ethfrac* represent? It represents going from a perfectly homogeneous society (*ethfrac* $= 0$) to a perfectly heterogeneous one (*ethfrac* $= 1$)!
- Do we think this slope is significantly different from zero?
  - Yes! Review Lecture 5 Hypothesis Testing and Chapter 4 of *Real Stats*.

Got causal?

Okay, but can you make causal claim? No!

Maybe we need to add some more variables into our regression to capture the relationship more fully.

The idea of multivariate regression is to try to “control” for or remove the effects of some potential confounders.

Let’s try adding *oilproduction*, the proportion of GDP from oil. Why?

- Perhaps by accident of history, places higher in ethnic fractionalization have higher oil production, and it is this difference in oil production that is slowing growth due to its effect on the economy (the “resource curse”, recall the midterm).
- We would call *oilproduction* a confounder.
- So we are interested in “controlling for” *oilproduction* or asking “at a fixed level of oil production, what is the relationship between ethnic fractionalization and growth”?

Let's try running our first multivariate OLS

```
summary(lm(rgdpgrowth ~ ethfrac + oilproduction_gdp, data = dat))

Call:
lm(formula = rgdpgrowth ~ ethfrac + oilproduction_gdp, data = dat)

Residuals:
Min       1Q  Median      3Q     Max
-67.558  -3.000   0.169   3.189  66.859

Coefficients:
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         3.3323      0.2919   11.414    <2e-16 ***
ethfrac            -0.4823      0.5290   -0.912     0.362
oilproduction_gdp  -1.3955      1.5634   -0.893     0.372
---
```

- When *ethfrac* $= 0$ and *oilproduction* $= 0$ (i.e. in perfectly homogeneous societies that have no oil production), the predicted *growth* rate of GDP is 3.33%.
- A 1-unit increase in *ethfrac* is associated with a *decrease* of the GDP growth rate of 0.48 percentage points, holding *oilproduction* constant.
- An increase of 1 percentage point of oil production as a share of GDP is associated with a *decrease* of the GDP growth rate of 1.40 percentage points, controlling for *ethfrac*.

How do I interpret significance?

- Just like in a bivariate regression, you read the significance for each X (independent variable).

- After controlling for other factors, do we still see there is a significant relationship between X and Y? Remember, these Xs are competing against each other to explain Y.

- We need to look at the p-value associated with that covariate (independent variable) – is it significant?
  - We have separate null and alternative hypotheses for *each* coefficient:
    1. $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$
    2. $H_0 : \beta_2 = 0$ vs. $H_A : \beta_2 \neq 0$

- What happened when we added *oilproduction* to our model? The coefficient on *ethfrac* went from being statistically significant ($p = 0.0296$) to losing statistical significance ($p = 0.362$).

Now, do we have a causal interpretation?

Suppose the coefficient on *ethfrac* had remained significantly negative in our multivariate regression model after adding *oilproduction*.

Then, would you feel comfortable saying *ethfrac* slows *growth*?

- "Controlling for" one potential confounder doesn't mean you have controlled for all of them! So NO, you are not clear to make causal claim.
- But you are probably better off having eliminated this one potential confounder or alternative explanation for the coefficient.

Recall that when critiquing a causal claim, you are obligated to point out potential confounders.

Meanwhile, the researcher's goal is to include those potential confounders as control variables, when possible to rule out those concerns.

Not as good as an experiment, but we do the best we can.

Political Science 15
Introduction to Research in Political Science
Lecture 6c: Visualization with Multivariate Regression

Alice Lépissier
University of California Santa Barbara

Multivariate regression - recap

Remember This

1. Multivariate OLS is used to estimate a model with multiple independent variables.
2. Multivariate OLS fights endogeneity by pulling variables from the error term into the estimated equation.
3. As with bivariate OLS, the multivariate OLS estimation process selects $\hat{\beta}$s in a way that minimizes the sum of squared residuals.

Chapter 5 of *Real Stats*, p. 137, digital version p. 207.

Multivariate regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

Example: “the resource curse”

Let’s keep working with the data-set from the midterm.

We want to understand whether an abundance of natural resources is associated with worse development outcomes.

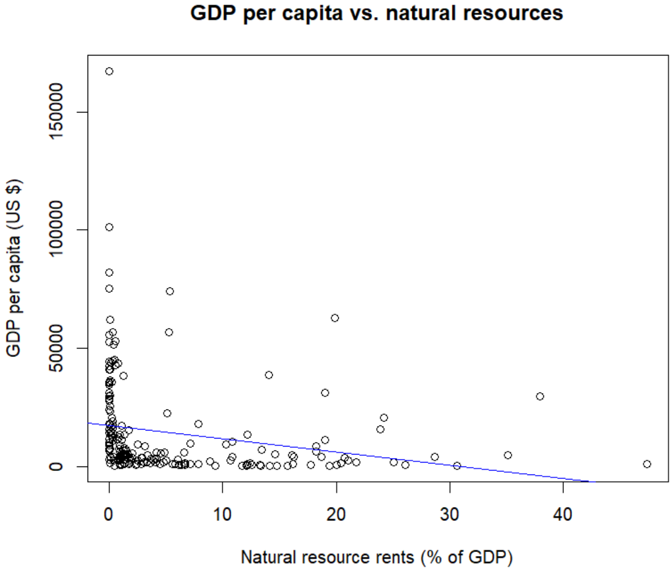
Let’s run a bivariate regression of GDP per capita on the percentage of natural resource rents in GDP.

```
> biv <- lm(GDPPerCap ~ NatResourceRents, data = development)
> summary(biv)

Coefficients:
                 Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)       17533.2       1793.1     9.778  <   2e-16 ***
NatResourceRents   -558.9        178.5    -3.131    0.00201 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The slope coefficient is negative and statistically significant.
- A 1-percentage point increase in the share of natural resources in GDP is associated with a **decrease** of $559 in GDP per person.

Bivariate regression of GDP on natural resources



Possible confounder

But one possible **confounder** is the quality of institutions. Perhaps economies that are better managed will have a lesser share of natural resources in their GDP, *and* will have better economic performance.

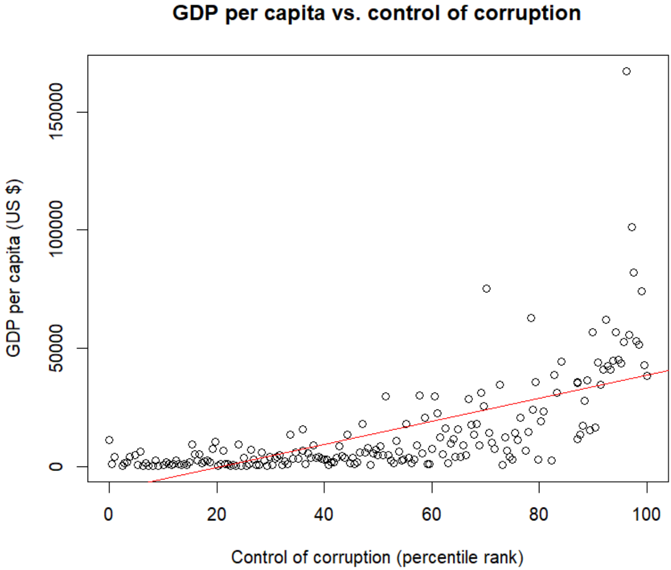
Let's run a bivariate regression of GDP per capita on the control of corruption (higher values ⇒ higher quality institutions).

```
> biv2 <- lm(GDPPerCap ~ ControlCorruption, data = development)
> summary(biv2)

Coefficients:
                  Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)       -9935.70      2303.42    -4.313   2.56e-05 ***
ControlCorruption   486.52        40.14    12.120    < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The slope coefficient is positive and statistically significant.
- A 1-unit increase in the control of corruption is associated with a **increase** of $487 in GDP per person.
- Whoa! What do you think is happening to the intercept?

Bivariate regression of GDP on control of corruption



Adding the confounder to our model

Recall the definition of a **confounder**: a variable that is associated both with the explanatory and the outcome variable.

Turns out, natural resource rents and institutional quality are correlated with each other.

```
> cor(development$NatResourceRents, development$ControlCorruption)
-0.4319271
```

Let's add *control of corruption* to our model.

So we estimate the following multivariate model:

$$GDPPerCap_i = \beta_0 + \beta_1 NatResourceRents_i + \beta_2 ControlCorruption_i + \epsilon_i$$

and hope that we do a better job (less bias, more precision).

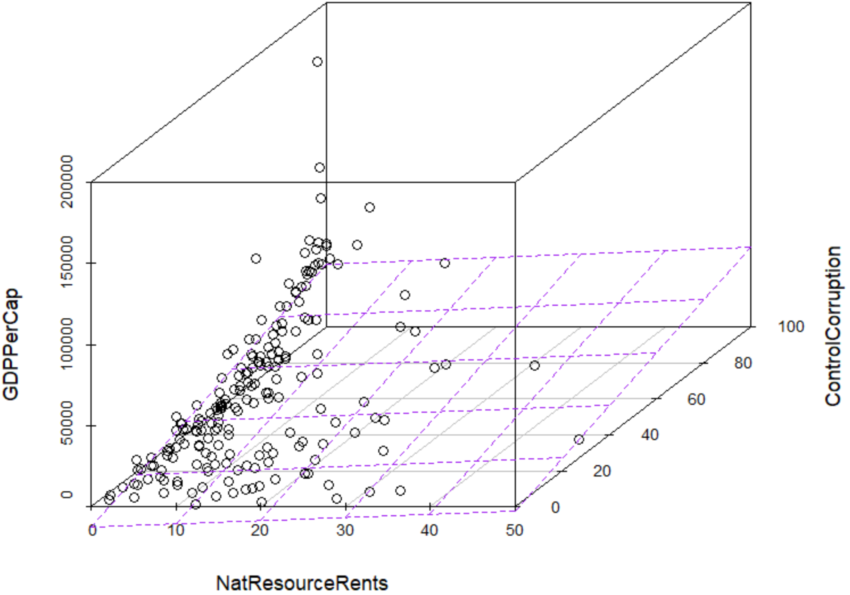
Adding the confounder to our model

```
> multiv <- lm(GDPPerCap ~ NatResourceRents + ControlCorruption,
data = development)
> summary(multiv)

Coefficients:
                   Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)       -12348.21      2939.93    -4.200   4.07e-05 ***
NatResourceRents     201.09       152.73     1.317       0.19
ControlCorruption    511.78        44.43    11.520    < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The coefficient $\hat{\beta}_1$ on *NatResourceRents* has switched sign and lost statistical significance.
- The coefficient $\hat{\beta}_2$ is still positive and statistically significant.
- Check. How do we interpret the coefficients?

Visualizing multivariate regression in 3D



Political Science 15
Introduction to Research in Political Science

Lecture 6d: Prediction with Multivariate Regression

Alice Lépissier
University of California Santa Barbara

Prediction with multivariate regression

We can use multivariate regression model as a "prediction machine" and obtain the **fitted values** $\hat{Y}$.

This is the predicted outcome (by your model), given a certain set of values for your independent variables Xs.

What *does* it mean to "hold other variables constant" when we are interpreting a multivariate regression?

This means you can, e.g., interpret $\hat{\beta}_1$ as the amount that Y will change when X_1 increases by 1 unit, when X_2 stays unchanged.

Prediction with multivariate regression

Let's extract our estimated $\hat{\beta}$ coefficients from our resource curse example.

```
> coef(multiv)
(Intercept)  NatResourceRents ControlCorruption
-12348.2116          201.0938          511.7794
> beta0hat <- coef(multiv)["(Intercept)"]
> beta1hat <- coef(multiv)["NatResourceRents"]
> beta2hat <- coef(multiv)["ControlCorruption"]
```

Prediction with multivariate regression

Let's see what happens to the predicted *GDPPerCap* when we change *NatResourceRents* by 1 unit, and keep *ControlCorruption* at a fixed level.

Turns out, the mean value of control of *ControlCorruption* is around 50.

```
> mean(development$ControlCorruption)
49.87735
```

Let's also pick a value of 10 for *NatResourceRents*.

```
> beta0hat + beta1hat*10 + beta2hat*50
15251.7
```

So, in a country where 10% of the GDP comes from natural resource rents and which ranks on the 50th percentile for control of corruption, the predicted GDP per capita is 15,252 dollars. Now, increase *NatResourceRents* by 1 unit.

```
> beta0hat + beta1hat*11 + beta2hat*50
15452.79
```

Calculate the difference. What number did we get? Our slope coefficient on natural resource rents, $\hat{\beta}_1$!

```
> 15452.79 - 15251.7
201.09
```

Prediction with multivariate regression

This works for any arbitrary level that you set your variables at!

```
> beta0hat + beta1hat*1 + beta2hat*50
13441.85
> beta0hat + beta1hat*2 + beta2hat*50
13642.94
> 13642.94 - 13441.85
201.09
```

This still returns our slope coefficient on natural resource rents, $\hat{\beta}_1$!

This works the other way too (varying *ControlCorruption* by 1 and keeping *NatResourceRents* constant).

```
> beta0hat + beta1hat*1 + beta2hat*50
13441.85
> beta0hat + beta1hat*1 + beta2hat*51
13953.63
> 13953.63 - 13441.85
511.78
```

Check. What does this number represent? Our slope coefficient on control of corruption, $\hat{\beta}_2$!

Prediction in R

So this is why we say “holding other factors constant” when we interpret a multivariate regression!

We can interpret each $\hat{\beta}$ as the “marginal effect” of the corresponding explanatory variable on the outcome, while controlling for the effect of the other independent variables.

Note: you can also use the `predict()` function in R to get your fitted values $\hat{Y}$.

```
> # You can do this
> predict(multiv,
          newdata = data.frame(NatResourceRents = 1,
                               ControlCorruption = 50))
13441.85

> # Or you can do that
> beta0hat + beta1hat*1 + beta2hat*50
13441.85
```

Political Science 15
Introduction to Research in Political Science
Lecture 6e: Using Multivariate Regression in Practice

Alice Lépissier
University of California Santa Barbara

Recap

- We talked about how multivariate regression can help us combat endogeneity in observational studies.
- If you are worried about omitted variable bias, you can include confounders in your model.
- This takes them out of the error term, and will lead to less bias and greater precision in your estimates.
- We discussed how to visualize multivariate regression, and how you can no longer use 2D scatter plots. (You can use a 3D scatter if you have 2 independent variables; after that, you're tapped out).
- So instead we rely on the interpretation of the $\hat{\beta}$ coefficients to understand the effect of shifting 1 variable, while holding the others constant.

Adding control variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

- Sometimes, our research question is mostly concerned with the effect of (for example) X_1 on Y, but we still include X_2 and X_3 in our model because we think they might be confounders.
- In this case, we refer to X_2 and X_3 as **control variables**, while X_1 is our primary **explanatory** or **predictor** variable.
- This is a distinction without a difference. What you consider to be a control variable and what you consider to be an explanatory variable depends *on your research question.*
- Back to the experimental setting for a moment. Intuition check. Say you have a randomized experiment, where X is a *treatment* and Y is your outcome. Do you need to include control variables in your regression? In principle, no! Randomization ensures that treatment and control groups are comparable across *all* possible confounders. In practice, adding control variables helps with "balance problems" in experiments (more later) and increases precision of estimates.

Using multivariate regression in practice

Let's talk about *how* to use multivariate regression in practice.

- How do you decide what goes in your model? This is called **model specification**.
- How do you assess your model's performance? This is called **model diagnostics**.
- What are potential issues that may come up (and how big of a deal are they)? We'll discuss **multicollinearity** and **irrelevant variables**.

Model specification

Say you want to understand the determinants of economic development. You run a regression of *GDPGrowth* on *NaturalResourcesRents* and *PolityScore*.

Your friend suggests that you control for *ColonialHistory*, because countries with, for example, richer natural resources are more likely to have been colonized, and this would bias your coefficients.

Next, your friend recommends that you add *EducationLevel* to your model, because forgetting about human capital might lead to omitted variable bias.

Then, your friend (who is starting to become annoying, but who is right), mentions that *Technology* is related both to human capital and growth.

Intuition check. Can you ever be sure you controlled for every possible confounder? (Can you ever shut your friend up?) No! Can't possibly control for everything (and some things can't be measured).

Model specification

In observational studies, you cannot achieve exogeneity by trying to control for everything!

You need a careful research design if you want to be able to make causal claims.

Another problem with the “throw everything in your regression” approach is the temptation to go **model fishing**.

Model fishing or **p-hacking** is when you run loads of regressions with different specifications (different controls, different samples, different measurements) and report only the results that support your hypothesis (such as those with statistically significant effects).

Model fishing

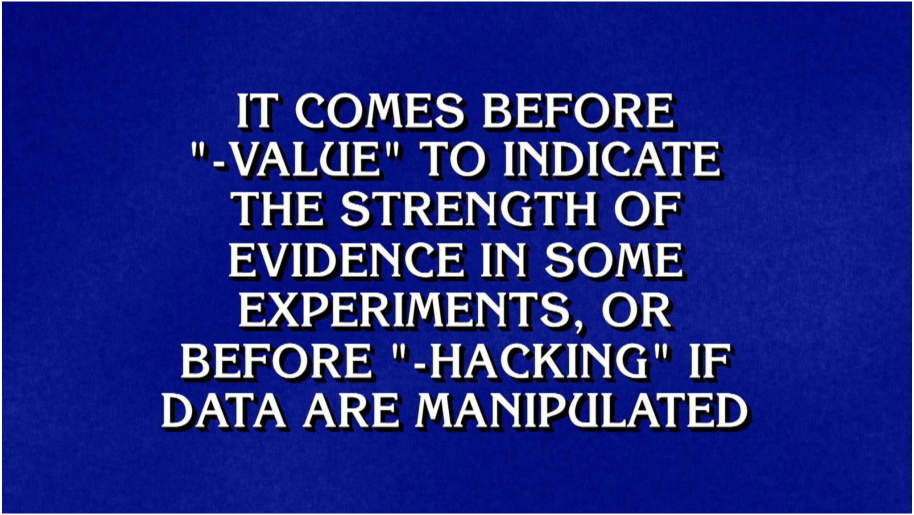
Model specification is hard!

Often times, if you don't have clear exogeneity (e.g. random assignment), you'll find that *how* you specify your model will dramatically affect the results you get (your $\hat{\beta}$ estimates).

This is disheartening! You don't know whether to trust your results. You can never be sure if there is stuff hiding out in your error term that is causing omitted variable bias.

So maybe you decide to go fishing: you run 100 models until you find one that you like, which shows that gun control causes more gun deaths, or that immigration is associated with lower wages for low-income Americans. What's the problem with this approach? (Apart from the fact that this is close to research malpractice...) Your results aren't **robust**.

What is "p", Alex?



IT COMES BEFORE
"-VALUE" TO INDICATE
THE STRENGTH OF
EVIDENCE IN SOME
EXPERIMENTS, OR
BEFORE "-HACKING" IF
DATA ARE MANIPULATED

Read a short article on the origin of "p-hacking" term: https://www.wired.com/story/were-all-p-hacking-now/.

Research reproducibility as a defence against model fishing

Increasingly, empirical social scientists are embracing the principles of open science. Part of this includes a greater expectation of **research reproducibility**.

This entails providing details of all the steps needed to reproduce your analysis/results (and sometimes the data, depending on constraints).

`RMarkdown` is so cool because it allows you to combine text, code, and outputs in one nifty document. Great to explain your methodology!

Science responsibly, folks!

Read more p. 157 of textbook, digital version p. 243, and *Chapter 2*.

Using RMarkdown for research reproducibility

Artwork by @allison_horst

Model diagnostics

Beyond the research design issues we just discussed, you might want to know **how well your model performs**, statistically speaking.

Recall that the R^2 is the proportion of variance in Y explained by the variance in your model. It is a measure of 'goodness of fit" of your model.

Problem: the R^2 will mechanically increase (or stay the same) as you add more independent variables to your model. That's not helpful.

Instead, in multivariate regression, we use a measure called **adjusted** R^2, which adds a *penalty term* if the variable you just added does add more in terms of your explanatory power.

- Use `Adjusted R-squared` in your `R` regression output, rather than `Multiple R-squared`.

Remember

A high R^2 is *not* the be-all and end-all. A biased model can have a high R^2! (Read more in Chapter 5 of *Real Stats* p. 151, digital version p. 231.)

Multicollinearity

This is a situation where your independent variables are (strongly) correlated with each other.

This will inflate your standard errors, that is, this will increase the variance of your $\hat{\beta}$ estimates.

Intuition check. What are the consequences of increased standard errors? It will make it harder for your coefficients to attain statistical significance.

But, and this is key: **multicollinearity does not cause bias**. Phew!

So, your $\hat{\beta}$ estimates will be centered around the true β (if you have exogeneity), but the distribution might be wider. That is, your $\hat{\beta}$ estimates will jump around more.

Read more p. 148 of textbook, digital version p. 226.

Including irrelevant variables

Suppose you didn't heed my earlier advice that you cannot "control your way to causal inference".

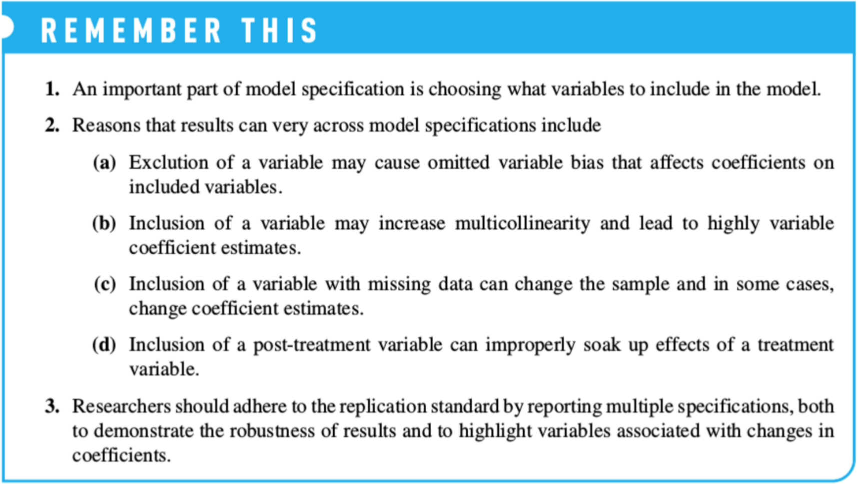
Say you decide to add an **irrelevant variable** to your multivariate regression model of *GDPGrowth*, such as the color of the national flag.

Like in the case of multicollinearity, the good news is that **including irrelevant variables does not cause bias**.

But there is a price to pay. Again, the variance of your coefficient estimates will increase.

Read more p. 151 of textbook, digital version p. 231.

Model specification recap



REMEMBER THIS

1. An important part of model specification is choosing what variables to include in the model.
2. Reasons that results can very across model specifications include
   - (a) Exclution of a variable may cause omitted variable bias that affects coefficients on included variables.
   - (b) Inclusion of a variable may increase multicollinearity and lead to highly variable coefficient estimates.
   - (c) Inclusion of a variable with missing data can change the sample and in some cases, change coefficient estimates.
   - (d) Inclusion of a post-treatment variable can improperly soak up effects of a treatment variable.
3. Researchers should adhere to the replication standard by reporting multiple specifications, both to demonstrate the robustness of results and to highlight variables associated with changes in coefficients.

Final thoughts

- We realized that we needed to include more variables in our model in order to fight endogeneity and omitted variable bias.
- Consequently, we used multivariate regression to bring variables out of our error term into our model.
- Although this can help in some circumstances, in practice we may not know how to correctly specify the model, and our model is not likely robust to alternate model specifications.
- For this reason, we often need to have some kind of random assignment to correctly and robustly identify causal effects.
- We will talk about this briefly next week as we cover experiments. The rest of the textbook covers alternative ways to obtain exogeneity in observational studies (e.g. instrumental variables, regression discontinuity designs). This is beyond the scope of this class, but you are very welcome to read the textbook if you are curious.