Practicing regression

Econ 140 Spring 2025, Section 4

Jonathan Old

Syllabus/OH bCourses Website Feedback form (**Always open**) RStudio

Roadmap

1. Time for your questions
2. Omitted Variable Bias
3. Bad Controls
4. Quadratic Terms
5. Interaction terms
6. Logs
7. Topics we've glossed over so far
8. Exam practice

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

Omitted Variable Bias

The most important slide of this course (until the midterm)

Let Y_i be the outcome variable, X_i our regressor of interest, W a series of control variables, and Z_i the "omitted" variable.

$$\text{[Long regression]} \quad Y_i = \alpha_L + \beta_L X_i + \lambda Z_i + W\beta_{WL} + e_i^S$$

$$\text{[Short regression]} \quad Y_i = \alpha_S + \beta_S X_i + W\beta_{WS} + e_i^L$$

$$\text{[Auxiliary regression]} \quad Z_i = \pi_0 + \pi_1 X_i + W\pi_{WP} + v_i$$

Then, the **Omitted variable bias formula** states that:

$$\underbrace{\beta_S}_{\text{Short}} = \underbrace{\beta_L}_{\text{Long}} + \underbrace{\lambda}_{\text{Omitted}} \cdot \underbrace{\pi_1}_{\text{Included}}$$

The OVB formula describes what happens to our coefficient of interest, β, as we include one additional variable Z in the regression. We call $\lambda\pi_1$ the **omitted variable bias**. Direction of bias: multiply our guesses for the signs of λ and π_1.

Bad Controls

Control variables

Control variables are additional variables (or covariates) **included in a regression**. We do this for **various reasons** (in decreasing order of importance):

- To remove selection bias / omitted variable bias
- To increase precision of our estimates
- To know about the (conditional/partial) correlation of other variables
- To better predict the outcome

Bad controls

- Some controls are called ”bad controls”. These are:
  1. Variables that are themselves outcomes of a treatment: What happens if you control for the change in English test scores in the regression below?
  2. Variables that moderate the treatment effect, e.g. controlling for occupation choice in gender wage gap regression . . .
- Rule of Thumb: Good controls are either pre-determined or immutable characteristics.
- Another way to think about it: Controls help us make ”apples to apples” comparisons. We should think before what exactly we want to compare to each other.

Mathematically, good and bad controls are the same thing.

We need to use our 🧠 to distinguish them!

Quadratic Terms

Making OLS more interesting

- We can extend the simple OLS framework

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

  to something richer:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

- All questions of the type *"how is Y_i expected to change if we change X_i", keeping all other variables in the regression fixed* can be solved with **partial derivatives** – in this case:

$$\frac{\partial Y_i}{\partial X_i} =$$

Making OLS more interesting

- We can extend the simple OLS framework

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

  to something richer:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

- All questions of the type *"how is Y_i expected to change if we change X_i", keeping all other variables in the regression fixed* can be solved with **partial derivatives** – in this case:

$$\frac{\partial Y_i}{\partial X_i} = \beta_1 + 2 \cdot \beta_2 \cdot X_i$$

Example for Quadratic Terms

Let us see how to use quadratic terms on `Datahub`

Interaction terms

How to think about interaction terms

We will cover this as an exercise. My most important advice to you is that you should use partial derivatives!
The slides after this will be skipped in class, but feel free to use them for your own learning.

Interaction terms: Making OLS more interesting

Let us consider the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

where Y_i is a country's GDP per capita, X_{1i} the value of its natural resources, and X_{2i} a measure of how democratic it is.

1. How do we interpret β_1?

Interaction terms: Making OLS more interesting

Let us consider the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

where Y_i is a country's GDP per capita, X_{1i} the value of its natural resources, and X_{2i} a measure of how democratic it is.

1. How do we interpret β_1?
   Keeping democracy fixed, increasing the value of a country's natural resources by one unit is associated with β_1 higher GDP per capita.
2. How do we interpret β_2?

Interaction terms: Making OLS more interesting

Let us consider the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

where Y_i is a country's GDP per capita, X_{1i} the value of its natural resources, and X_{2i} a measure of how democratic it is.

1. How do we interpret β_1?
   Keeping democracy fixed, increasing the value of a country's natural resources by one unit is associated with β_1 higher GDP per capita.
2. How do we interpret β_2?
   Keeping natural resources fixed, increasing a country's democracy score by one unit is associated with β_2 higher GDP per capita.

Interaction Terms (ii)

Now, let us extend the model to:

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

1. What is the ”effect” of X_{1i} on Y_i?

Interaction Terms (ii)

Now, let us extend the model to:

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

1. What is the ”effect” of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$

Interaction Terms (ii)

Now, let us extend the model to:

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

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1?

Interaction Terms (ii)

Now, let us extend the model to:

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

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**

Interaction Terms (ii)

Now, let us extend the model to:

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

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2?

Interaction Terms (ii)

Now, let us extend the model to:

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

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2? **The effect of an additional unit of X_{2i} , if X_{1i} is equal to 0.**

Interaction Terms (ii)

Now, let us extend the model to:

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

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2? **The effect of an additional unit of X_{2i} , if X_{1i} is equal to 0.**
4. How do we interpret $\beta_1 + \beta_3$?

Interaction Terms (ii)

Now, let us extend the model to:

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

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2? **The effect of an additional unit of X_{2i} , if X_{1i} is equal to 0.**
4. How do we interpret $\beta_1 + \beta_3$? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 1.**

Rule of thumb: Always use partial derivatives to make sure that you are right!

Interactions with only dummy variables

- Take any two binary variables, for example: Studies at UC Berkeley or not (UCB_i), and is from California or not (Cali_i).
- The regression with interactions looks like this:

$$Y_i = \underbrace{\beta_0}_{17.77} + \underbrace{\beta_1}_{2.28} \text{UCB}_i + \underbrace{\beta_2}_{0.95} \text{Cali}_i + \underbrace{\beta_3}_{-0.97} \text{UCB}_i \times \text{Cali}_i$$

- We can write this in a table, and get all group averages

	Cali = 1	Cali=0
UCB = 1	$\beta_0 + \beta_1 + \beta_2 + \beta_3$ =17.77+2.28+0.95-0.97	$\beta_0 + \beta_1$ = 17.77 + 2.28
UCB = 0	$\beta_0 + \beta_2$ = 17.77 + 0.95	β_0 =17.77

- Take differences between cells to get different effects!

Logs

Notes on logarithms

- We can take logs of whole equations to get linear models (problem set)
- We can also take logs of specific variables, especially when they have long tails (wealth in the US, GDP per capita, etc.)
- We can get to the right interpretation of log-specifications with just math

Notes on logarithms

- We can take logs of whole equations to get linear models (problem set)
- We can also take logs of specific variables, especially when they have long tails (wealth in the US, GDP per capita, etc.)
- We can get to the right interpretation of log-specifications with just math
- But I will make your life easier with a cheat sheet.

Logs: Cheatsheet

Model	LHS	RHS	A change in x by ...	is associated with a change in y by ...
Level-Level	y	x	1 unit	β_1 units
Level-Log	y	$\log(x)$	1%	$\beta_1/100$ units
Log-Level	$\log(y)$	x	1 unit	$100\beta_1\%$
Log-Log	$\log(y)$	$\log(x)$	1%	$\beta_1\%$

If you want to get a bonus star from me, write ”approximately” in log-interpretations.

Topics we've glossed over so far

What if the outcome variable is binary (a dummy variable)?

Let's run the regression

$$\text{Defaulted}_i = \alpha + \beta \widetilde{\text{Credit Score}}_i + e_i$$

where Defaulted_i is equal to 1 if individual i has ever defaulted on a loan (mortgage, credit card, auto loan, etc.), and $\widetilde{\text{Credit Score}}_i$ is i's credit score, **minus the average credit score in the sample** (Note: US credit scores range from 300 to 850 points).

1. You run a regression and get $\hat{\alpha}$=0.1. How do you interpret this? Does this number make sense here?
2. Your estimate for β is $\hat{\beta} = 0.001$. Interpret.

What if the outcome variable is binary (a dummy variable)?

Let's run the regression

$$\text{Defaulted}_i = \alpha + \beta \tilde{\text{Credit Score}}_i + e_i$$

where Defaulted_i is equal to 1 if individual i has ever defaulted on a loan (mortgage, credit card, auto loan, etc.), and $\tilde{\text{Credit Score}}_i$ is i's credit score, **minus the average credit score in the sample** (Note: US credit scores range from 300 to 850 points).

1. You run a regression and get $\hat{\alpha}$=0.1. How do you interpret this? Does this number make sense here?
2. Your estimate for β is $\hat{\beta} = 0.001$. Interpret.

With a dummy dependent variable, changing X_i by one unit increases the probability of $Y_i = 1$ by $\hat{\beta} \cdot 100$ percentage points.

Exam practice

Practice exam question: 1a)

The Ministry of Truth is interested in a rumour that **air pollution could impact mental health**. One of the most harmful pollutants is fine particulate matter PM2.5, which comes from operations that involve the burning of fuels such as wood, oil, coal, etc. A research team is sent to investigate the rumour. The team **randomly** selects and surveys 19,920 people across 71 districts of the country. The key variable, Exposure E_i, is a **dummy variable** equal to 1 if the individual i is exposed to a large amount of PM2.5 in the last 24 hours, and 0 otherwise. The team also conducts a standardised questionnaire to record **depressive symptoms** in the last month, called the Kessler Psychological Distress scale (K6). The questionnaire results in a score, Depression_i, that ranges from 0 to 24; and the higher the score, the more severe the depressive symptoms for individual i. The variable has a sample average of 2.96. **Running regressions with Depression D_i as the dependent variable, you obtain the following results**:

Practice exam question: 1a)

Dependent variable: Depression $_i$

Regressor	(1)	(2)	(3)
Exposure $_i$	0.834	0.614	0.554
	(0.032)	(0.045)	(0.042)
Exposure $_i \times$ Female $_i$		0.065	
		(0.024)	
Female $_i$		−0.739	−0.825
		(0.036)	(0.066)
Age_i			0.452
			(0.132)
Age_i^2			0.524
			(0.121)

Notes: All estimations contain a constant term. Robust standard errors are in the parentheses. Age_i is the age (years old) of individual i, and Age_i^2 is the square of Age_i.

Pratice Exam question: 1a)

a) Interpreting the coefficient in Column (1), a journalist, Katherine, claims: ”Since participants are randomly selected, we can infer that exposure to a large amount of PM2.5 does cause depression.”

i. Explain carefully why Katherine is wrong, specifying the direction of bias(es) if there is any. Which assumption(s) would she need to impose for the causality claim to hold?

ii. What is the correct interpretation from Column (1) that Katherine should have made?

(Detailed) Suggested Answer: 1a)

i. **Random selection is not the same thing as random assignment to treatment!** Survey respondents may be systematically different from each other in ways that are correlated with depression and pollution exposure. Therefore, the results from a regression can not be interpreted causally (and are biased). A priori, it is unclear in which direction the bias would go, but we could imagine that (**Only one explanation needed for exam**):
On rainy days, pollution is lower (-) and people may be reporting more depression symptoms (+), leading to downward bias. More wealthy people choose to live in less polluted areas (-) and they may have less depression (e.g., better access to mental health resources) (-), leading to upward bias.
For the regression causality claim to hold, we need to assume that: People exposed to pollution and those not exposed to pollution would have, on average, the same depression level, had they been exposed to the same level of pollution. In other words: **Both groups would have to have the same potential depression outcomes**.

ii. On average, people that were exposed to pollution had a 0.8 points higher score on the depression scale. The difference between the two groups is significant at the 5% level.

Pratice Exam question: 1b)

b) Interpret column (2) of the regression table

i. A colleague notes the the coefficient on Female_i is significant, and states: ”The effect of being female on depression is significantly different from zero”. Do you agree with the statement? Why or why not?

ii. How is pollution exposure related to depression, for men? And how for women?

(Detailed) Suggested Answer: 1b)

i. It is difficult to make such interpretations when interaction terms are involved. Taking partial derivatives, the "effect" of being female is:

$$\frac{\partial \text{Depression}_i}{\partial \text{Female}_i} = -0.739 + 0.065 \cdot \text{Exposure}_i$$

We can do inference (and test significance) at $\text{Exposure}_i = 0$ (just looking at the coefficient for female, -0.739 is significant). We can also do it for any other level of exposure, but for that we also need to take the other coefficient into account and cannot just use the table.

Comment 1: Less relevant for the exam, but important to be aware of!
Comment 2: We can always do inference on the interaction term, which is significant here!

ii. Taking partial derivatives, the "effect" of pollution exposure is:

$$\frac{\partial \text{Depression}_i}{\partial \text{Exposure}_i} = 0.614 + 0.065 \cdot \text{Female}_i$$

Hence, the "effect" for Males is 0.614 and the "effect" for Females is larger ($0.614 + 0.065 = 0.779$). The effect of pollution is also **signficantly** larger than for men, because the coefficient on the interaction term is significantly different from zero.

If time permits: Practice exam question, midterm 2022.