Lecture 3 –Demystifying Regression

In IEM, you had some exposure to regression but you likely felt rushed and maybe didn't completely understand it.

There are multiple courses dedicated to regression at every university and hundreds researchers who primarily study properties of regression. I could design this class around properties of regression and never run out of material (and it would be very boring).

Instead, we are going to focus on the intuition for what regression is doing and how to interpret it.

Regression as Line Fitting

One way to think about regression is as a line fitting exercise (see picture on the board)

What is the best way to fit this line? We all have some sort of intuitive idea. Regression formalizes this by minimizing the sum of squared residuals.

The idea we have in our mind can be represented by

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

On the figure: What is the residual? Why use the square?

Mathematically it is doing the following:

$$\min_{\beta} \sum_i (Y_i - \beta_1 X - \beta_0)^2$$

What does β tell us? The best guess at the relationship between X and Y

Maybe in IEM you learned the formula. In the two variable case it is

$$\hat{\beta} = \frac{covariance(Y, X)}{variance(X)}$$

The covariance is how much do and y move together. The variance is how much variation is there in X

What is the interpretation of $\hat{\beta}$?

How much do we expect Y to go up if X goes up by 1 unit? This is the number you put in a policy memo

If this were a regression class we could prove things about how good a predictor this is, etc. I won't go into detail, but it is very good. In fact, in typical conditions it is the best.

Multiple Regression

Okay, so now we hopefully have some sense of what single variable regression is. However, regression is most useful when we have more than one X variable. E.g.

$$Wages_i = \beta_0 + \beta_1 ParentIncome_i + \beta_2 Education_i + \epsilon_i$$

Why might you want to account for both of these things? People whose parents have high income may have different amounts of education and that may be what is driving differences in wages, not the education.

What you would ideally have is a dial where you can turn the years of education up by 1 year but keep parent income the same. We might also want the opposite (keep education the same but change parent income).

Can we do this? Yes! The math is slightly different but the idea is more or less the same.

Here is one way to think about it. First you regress wages on parent income (only). Then you get the $\tilde{Y}$ which the leftover part of Y after accounting for your best guess at the relationship of wages and parent income. You also do this for the Xs and generate an $\tilde{X}$ (e.g. you regress parent income on education)

Then you do the same thing except you use $\tilde{Y}$ (instead of Y) and $\tilde{X}$ and do a line fitting exercise with that. The resulting slope is the best fit for how much education matters for wages *after accounting for parent income*.

This is known as the Frisch–Waugh–Lovell theorem, but you do not need to remember the name—just the intuition.

There is nothing special about the order here, you could do it in reverse.

Why is this so useful? You can say how much does this particular X predict Y after accounting for other factors? (This was mind blowing to me when I first learned it).

You can sort out all sorts of things (potentially). Do rich people live longer because they're rich or because they have more education? What are some uses you can think of for this?

Causal versus prediction questions

It is very important to distinguish between causal and prediction questions.

Causal question: What is the effect of X on Y? Prediction question: How are X and Y related?

Is this a rain dance or umbrella question? Many policy mistakes come from treating predictive relationships as causal.

So far, I have talked about regression in entirely a prediction way (line fitting). This can be very useful.

However, think back to the formula for bivariate OLS, all that it tells us is how two things move together in the data. Why might two things move together? (Causal, reverse causality, third omitted factor, part of a system)

However we can infer causality if we are careful about what we feed into OLS and what assumptions we make.

Put another way, if we are very careful about what we feed into OLS, we can get something useful out. If we're not, we may not get something useful out.

Regression has several limitations. You must specify a functional form relationship between the Xs and Ys. If this is wrong, you can get misleading results.

Regression also makes it easy to extrapolate out of sample. You can add all of the βs but there many not be anyone in the world who has the Xs you picked. This concern is more important for prediction than it is for causal inference.

Another issue with OLS is that you are limited by the variation in your data. For instance, if you want to know what the relationship between gender and wages and your entire sample is female, you will not be able to estimate that relationship.

A linear probability model is where you have the Y variable be a binary variable (0/1). Why do we use linear probability models in modern empirical work? The are way easier to interpret than other options (logit/probit)

What is interpretation in the bivariate regression case where both are binary?

Notation

Some basic notation things

Here is a pretty standard regression equation:

$$Y_{ijt} = \beta_0 + \mathbf{X_{ijt}}\beta + \varepsilon_{ijt} \tag{1}$$

In this equation Y_{ijt} is the outcome for person i in group j in time period t. $\mathbf{X_{ijt}}$ is a vector of characteristics that can vary at some combination of person, group, or time level. For instance, a person's birth year doesn't change over time so that would be an example of something that varies at X_i, the national unemployment rate could

change by year (but not by person) which would be an example of an X_t, and the geographic size of a state is an example of X_j.

We can modify this in many ways, an important one is the addition of fixed effects as below.

$$Y_{ijt} = \beta_0 + \mathbf{X_{ijt}}\beta + \gamma_j + \varepsilon_{ijt} \tag{2}$$

In this equation γ_j are fixed effects for groups. These amount to dummy variables. Much of modern empirical economics relies heavily on fixed effects.

Omitted Variable Bias

When is observational OLS causal?

You may have discussed omitted variable bias in IEM. The math derivation is not so important but the last equation is extremely practical/useful.

Assume that the true relationship is the following:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \varepsilon_i \tag{3}$$

If we estimate this using OLS, we will get unbiased estimates of β_0, β_1, and β_2. What if we estimate $Y_i = \alpha_0 + \alpha_1 \cdot X_{1i} + \eta_i$ instead? Why might this happen?

In this case, we know that our estimate of $\hat{\alpha_1}$ will be

$$\frac{cov(X_1, Y)}{var(X_1)}$$

Well we can plug in what we know about Y into that equation:

$$\hat{\alpha_1} = \frac{cov(X_1, Y)}{var(X_1)}$$

$$= \frac{cov(X_1, \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \varepsilon_i)}{var(X_1)}$$

$$= \frac{cov(X_1, \beta_0) + var(X_1\beta_1) + cov(X_1, \beta_2 X_2) + \beta_2 cov(X_1, \varepsilon)}{var(X_1)}$$

$$= \frac{0 + \beta_1 var(X_1) + \beta_2 cov(X_1, X_2) + 0}{var(X_1)}$$

$$= \beta_1 + \beta_2 \cdot \frac{cov(X_1, X_2)}{varX_1}$$

So when we have omitted variable bias, it will depend on the 1) the true relationship between the omitted variable (X_2 above) and the outcome, and the covariance of our included variable (X_1) and our excluded variable (X_2), $cov(X_1, X_2)$

Put another way, we are okay to exclude a variable if either 1) it is not related to our outcome or 2) it is related to our outcome but it isn't related to our included variable of interest.

How does an RCT help with omitted variable bias?

In Mastering Metrics they go through examples of short and long regressions. The short regression does not include the omitted variable of interest. The long regression does. You can actually determine the amount of omitted variable bias if using the formula we derived above. See the text for details.

Let's say you're interested in the relationship between college graduation and financial aid. You regress graduation on financial aid. However, you're worried about the relationship between financial aid and income. What sign will the omitted variable bias be?

We need 2 pieces of information. 1) the true relationship between income and graduation (+) and the covariance between income and aid (-).

Hence the omitted variable bias is negative.

When you read a study trying to understand if it yields a causal question, how can you use omitted variable bias?

Robustness of Regression Results

In MM, they discuss Dale and Krueger's study on the effect of attending a private college on earnings. After accounting for the set of schools students apply to and are accepted at, the inclusion of additional controls doesn't change the estimate. This is very desirable–why?

It means that additional controls are not strongly correlated with our coefficient of interest. That is, these additional controls are not the source of omitted variable bias. People often then argue that unobserved, additional controls would not change the interpretation (are they right?) This argument is suggestive, not a proof.