Mostly Harmless Econometrics

Reading Group Discussion Guide

Chapter 3: Making Regression Make Sense (3.1-3.2)

Chapter 3.1: Regression Fundamentals

The authors emphasize three key concepts in regression analysis:

- Randomization of treatment solves selection bias... however, we usually only have observational data
- Population first approach – we must define the objects of interest before we can think of estimating them
- The Conditional Expectation Function (CEF) and Linear Regression are tightly linked
- We can use statistical tools to quantify uncertainty; i.e., statistical inference

Observational Data

Most of this book concerns observational data as opposed to experimental data. With observational data, the goal is to figure out how to approximate experimental conditions to estimate causal effects of interest. However, the mechanical tools of regression and statistical analysis are the same for experimental or observational data – the emphasis is just different tools in the 'toolbox.'

Population First

Suppose that we had data on the entire population of interest. What could we calculate? This is similar to asking what is the 'ideal experiment.'

CEF and Regression

If we want to know the effect of a change in X on Y, then the CEF essentially contains that information:

$$\text{Effect}(x) = \mathsf{E}[Y \mid X = x + 1] - \mathsf{E}[Y \mid X = x]. \tag{1}$$

Thus, our task is to estimate CEFs.

We can link the CEF to Linear Regression / Ordinary Least Squares (OLS) two ways:

1. *Assume* that $\mathsf{E}[Y \mid X] = X\beta$,
2. *Approximate* $\mathsf{E}[Y \mid X]$ using $X\beta$.

Question 1: Is it clear how these are two different approaches?

Inference

If we use some statistical tools (CLT, Continuous Mapping Theorem, Delta Method), then we can find the asymptotic distribution of regression coefficients that yield standard errors. What gets lost in Section 3.1.3 is that where do standard errors come from and what do they tell us?

This section jumps to discussing estimation using samples from the population. Using the population approach, we know that the 'true' regression coefficient, β, is a single number (like a 3). However, given that we do not observe populations but samples for empirical work, our *estimator* for the 'true' coefficient is a *function* of random variables, (X, e). These random variables have distributions, so our estimator inherits the distributional features of the random variables – one of these features is the dispersion of the estimator around its population targe (i.e., the 'true' coefficient). Once we have an *estimate* of the regression coefficient, we can then use the properties of the *estimator* and the data to also estimate the dispersion, which we call the *standard error*.

If $Y = X\beta + e$, then

- True coefficient : $\beta = \mathsf{E}[X'X]^{-1}\,\mathsf{E}[X'Y]$ — a number
- Estimator : $\hat{\beta} = \left(\widehat{\mathsf{E}}[X'X]\right)^{-1}\widehat{\mathsf{E}}[X'Y]$ — a function of (X, Y) (hypothetical data)
- Estimate : $\hat{\beta}$ with real data — a number

It is annoying that we do not have separate notation for an estimator versus an estimate.

Question 2: Why do we need to think about standard errors?

Chapter 3.2: Regression and Causality

The section covers three major ideas:

- Causality linked to Conditional Independence Assumption (CIA) of treatment and potential outcomes
- Omitted Variable Bias is the most likely violation of CIA through induced correlation with treatment and unobservable
- Bad Controls are a subtle issue that shows the value of thinking in terms of population variables

Conditional Independence Assumption

The CIA is that the selection into treatment does not depend on either potential outcome: $\{Y_{0,i}, Y_{1,i}\}$; i.e., that treatment is 'essentially' randomly assigned.

Critically, it is an **assumption**. How can we think about whether it holds or not? Sketch out an economic model. Is it plausible that the agent would choose the treatment unrelated to the outcomes you are looking at? The authors discuss college choice and incomes.

Omitted Variable Bias

Omitted Variable Bias is the first threat to CIA. The authors motivate this with the example of the college wage premium. The authors suppose that incomes are a function of going to college, C,

ability A, and other factors, e:

$$Y_i = \alpha + \rho C_i + \gamma A_i + e_i. \tag{2}$$

The CIA for this question ('what is the college premium') is: $\{Y_{0,i}, Y_{1,i}\} \perp C_i \mid A_i$. If we observed all three variables (Y, C, A), then we could estimate ρ.[1] However, suppose that we do not observe underlying ability of people, then we do not have CIA, and so we cannot causally estimate ρ.

Digression: You have likely heard and/or read the phrase identification strategy. In the context of this example (observe (Y, C) but not A), an identification strategy is a description of how one will estimate how C affects Y where the 'variation' in C is not affected by not observing A. Alternatively, it is a situation where there is 'variation' in C but A does *not* vary. Usually, such variation comes from instituional or government policies.

Bad Controls

Understanding 'Bad Controls' is critical. It is actually a pretty subtle issue that trips up smart people.[2] Essentially, if you want to know the causal effect of X on Y, then you cannot include as a control anything else that X affects. Rather, you should include things that influence X and Y.

The authors use the idea of 'white-collar' (office) jobs to better estimate the college premium or to look at the effect of college degree on wages condition on 'white-collar' status. The problem is that the type of job you get is likely affected by college degree choice.

Next Week: Sections 3.3 and 3.4

Read 3.3 (Heterogeneity and Nonlinearity) and 3.4 (Regression Details) for next week.

Forward Guidance: The next two sections are a little technical with little payoff. Try to get intuition but worry less about the details. Also, do not be scared... most of what they do is take definitions we have talked about and insert the assumptions of more complicated models into the definitions.

My thoughts: Personally, I do not care as much about matching or propensity score. I think they have fallen a little out of fashion since publication. The main thing is that it does not provide anything that regression does not also provide. The most important idea they introduce is that the 'constant effect' assumption typically used is restrictive because the causal effect of X on Y might depend on other factors (so that there is heterogeneity in the causal effect). However, if one believes there is heterogeneity in the causal effect, then just model it using interactions.[3]

[1] Note that, $\{Y_{0,i}, Y_{1,i}\} \perp C_i \mid A_i$ does not imply $\{Y_{0,i}, Y_{1,i}\} \perp A_i \mid C_i$! Thus, just because we could causally estimate ρ does not mean we could causally estimate γ.

[2] Paul Samuelson famously noted that *Comparative Advantage* is one the simplest idea in economics that many fail to understand or believe. 'Bad Controls' is not quite like this, but it is close.

[3] My guess is that matching was used more when computing power was 'more expensive.'

Ways of Justifying OLS

Constant Effect CEF

Let $\mathsf{E}[Y \mid X = x+1] - \mathsf{E}[Y \mid X = x] = \beta(x)$. Note, we can write this as:

$$\frac{\Delta\,\mathsf{E}[Y \mid X = x]}{\Delta X} = \beta(x), \tag{3}$$

where $\Delta X = (x+1) - (x) = 1$.

Assume constant effects: suppose that $\beta(x) = \beta(z) = \beta$ for any value of x, z. Choose $x = x$ and $z = 0$ as the two values. Thus,

$$\frac{\Delta\,\mathsf{E}[Y \mid X = x]}{\Delta X} = \frac{\mathsf{E}[Y \mid X = x] - \mathsf{E}[Y \mid X = 0]}{x} = \beta \tag{4}$$

$$\implies \mathsf{E}[Y \mid X = x] = \mathsf{E}[Y \mid X = 0] + \beta \cdot x \tag{5}$$

$$:= \alpha + \beta x. \tag{6}$$

Minimization of Linear Approximation

See page 35. Suppose that you wish to approximate the $\mathsf{E}[Y \mid X]$ using a linear function. You choose to minimize based on the least squared error:

$$b^\star = \min_b \{\mathsf{E}[(Y - Xb)^2]\}. \tag{7}$$

The first order condition for minimization implies that at $b^\star$:

$$\mathsf{E}[X \cdot (Y - Xb^\star)] = 0. \tag{8}$$

Do the matrix algebra:

$$\mathsf{E}[X \cdot (Y - Xb^\star)] = 0 \tag{9}$$

$$\implies \mathsf{E}[XY] = \mathsf{E}[XXb^\star] \tag{10}$$

$$= \mathsf{E}[XX] \cdot b^\star \tag{11}$$

$$\mathsf{E}[XX]^{-1} \cdot \mathsf{E}[XY] = \mathsf{E}[XX]^{-1} \cdot \mathsf{E}[XX]b^\star \tag{12}$$

$$= b^\star. \tag{13}$$

Finally, note that $\mathsf{E}[XX]^{-1} \cdot \mathsf{E}[XY] = \beta$, so $b^\star = \beta$.

Covariance of Y and X

This is simplified version of page 35-36 (regression anatomy). Let X be a single variable (i.e., not a matrix). Let $Y = a + \beta X + u$. Take the covariance of Y with X and then substitute for Y: $\mathsf{Cov}(Y, X) = \mathsf{Cov}(a + \beta X + u, X)$. Now expand: $\mathsf{Cov}(a, X) + \mathsf{Cov}(\beta X, X) + \mathsf{Cov}(u, X)$. Covariance with a constant is zero: $\mathsf{Cov}(a, X) = 0$. Covariance with the same variable is variance and constants come out: $\mathsf{Cov}(\beta X, X) = \beta\,\mathsf{Var}(X)$. This yields the following relationship:

$$\mathsf{Cov}(Y, X) = \beta\,\mathsf{Var}(X) + \mathsf{Cov}(u, X) \implies \beta = \frac{\mathsf{Cov}(Y, X)}{\mathsf{Var}(X)} + \frac{\mathsf{Cov}(u, X)}{\mathsf{Var}(X)}. \tag{14}$$

We need to get rid of the second RHS term, so we assume that the unobservable and X have zero variance: $\mathsf{Cov}(u, X) = 0$.

Now, we use the definitions of covariance and variance: $\mathsf{Cov}(Y, X) = \mathsf{E}[(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])] = \mathsf{E}[Y \cdot X] - \mathsf{E}[Y]\,\mathsf{E}[X]$, and $\mathsf{Var}(X) = \mathsf{E}[X^2] - \mathsf{E}[X]^2$. Assume $\mathsf{E}[X] = 0$. This allows us to write:

$$\beta = \frac{\mathsf{E}[YX]}{\mathsf{E}[X^2]} = \mathsf{E}[X \cdot X]^{-1} \cdot \mathsf{E}[X \cdot Y], \tag{15}$$

which is the OLS β.

What assumptions did we use:

- X is single variable
- Y is linear in X
- u and X have zero covariance
- $\mathsf{E}[X] = 0$

Assumption 1 can be easily relaxed; it just requires more math. Assumption 2 is still needed, but can be justified as an approximation. Assumption 3 has the most bite, see below. Assumption 4 is taken care of as long as we have a constant in the regression.

Note, like above: $\mathsf{Cov}(u, X) = \mathsf{E}[(X - \mathsf{E}[X])(u - \mathsf{E}[u])] = \mathsf{E}[u \cdot X] - \mathsf{E}[u]\,\mathsf{E}[X]$. Thus, the covariance is zero if (1) $\mathsf{E}[u \cdot X] = 0$ – we tend to call this uncorrelated – and (2) $\mathsf{E}[u] = 0$. Condition (2) like Assumption 4 is true as long as we have a "constant" in the regression (i.e., these are more-or-less absorbed by the α coefficient).

This derivation is different from the other two for two connected reasons: (1) we started with (Y, X) which is data we observe rather than $\mathsf{E}[Y \mid X]$ is a population function that we wish to estimate; (2) an assumption about the relationship between the unobservable and the observable variables.