Mostly Harmless Econometrics

Reading Group Discussion Guide

Chapter 3: Making Regression Make Sense (3.3-3.4)

Chapter 3.3: Heterogeneity and Nonlinearity

Set Up

Let Y be an outcome, X be a covariate, and D be treatment status. Let X be a variable that can take on three values: $\{a, b, c\}$. Let D be a binary variable: $\{0, 1\}$. Let i, j index different people from the sample $\mathcal{N}$. Thus, an observation in the data would be (Y_i, X_i, D_i).

Heterogeneity

For any individual, there is a treatment effect of:

$$\delta_i = Y_{i,1} - Y_{i,0}, \tag{1}$$

but we only observe $Y_i = (1-D_i)\cdot Y_{i,0} + D_i \cdot Y_{i,1}$. Using the above, we can write this as: $Y_i = Y_{i,0} + \delta_i \cdot D_i$.

If $\delta_i = \delta_j$ for all $i, j \in \mathcal{N}$, then there is no heterogeneity. However, we tend to assume that δ_i is different for different people.

Average Treatment Effect

Assuming there is treatment effect heterogeneity, we often want to summarize this information. The simplification often chosen is the Average Treatment Effect (ATE) or the Average Treatment on the Treated (ATT):

$$\delta_{ATE} = \mathsf{E}[\delta_i] \tag{2}$$

$$\delta_{ATT} = \mathsf{E}[\delta_i \mid D_i == 1]. \tag{3}$$

Question 1: Why would these be different? Can you think of an example?

Conditional Average Treatment Effect

While the ATE and the ATT are good aggregate measures, we may be interested in conditional effects. Suppose we are interested in how the treatment varies according the observed characteristics, X, such as parents' income or education status. We can calculate:

$$\delta_X = \mathsf{E}[\delta_i \mid X_i]. \tag{4}$$

The δ_X term is general for the random variable X, while the conditional treatment effect at a specific value would be $\delta_a = \mathsf{E}[\delta_i \mid X_i = a]$.

We can go back to the ATE from the CATE using the Law of Iterated Expectations:

$$\delta_{ATE} = \mathsf{E}[\delta_i] = \mathsf{E}[\mathsf{E}[\delta_i \mid X_i]] = \mathsf{E}[\delta_X]. \tag{5}$$

Estimation

How does β_{OLS} relate to δ_i?

Let's focus on $Y_i = Y_{i,0} + \delta_i \cdot D_i$.

If we were take this to linear regression, then we would probably do:

$$\text{M1} \qquad Y_i = \alpha + \beta D_i + u_i. \tag{6}$$

C1 If (A1) $\mathsf{Cov}(D_i, u_i) = 0$, (A2) $\mathsf{Cov}(\delta_i, u_i) = 0$, and (A3) $\mathsf{Cov}(\delta_i, D_i) = 0$, then $\beta = \delta_{ATE}$.

C2 If (A1) $\mathsf{Cov}(D_i, u_i) = 0$, (A2) $\mathsf{Cov}(\delta_i, u_i) = 0$, and (A3) $\mathsf{Cov}(\delta_i, D_i) \neq 0$, then $\beta = \delta_{ATT}$.

Question 2: What are the different assumptions? What do they mean?

Question 3: How does C2 relate to Question 1?

Adding X Variables

Lets suppose that $\delta_i = \delta(X)$, and consider the case where $Y_{i,0} = v_i + \gamma X_i$.

Consider three regressions:

M2 $Y_i = \alpha + \beta_1 D_i + u_i$

M3 $Y_i = \alpha + \beta_2 D_i + \theta X_i + u_i$

M4 $Y_i = \alpha + \beta_3 D_i + \theta X_i + \mu_X D_i \cdot X_i + u_i$

Lets continue to assume (A1) $\mathsf{Cov}(D_i, u_i) = 0$ and (A2) $\mathsf{Cov}(\delta_i, u_i) = 0$.

1. If (A1, A2) and $\mathsf{Cov}(\delta_i, D_i) = 0$, then $\beta_1 = \mathsf{E}[\delta_X]$; however, if $\mathsf{Cov}(\delta_i, D_i) \neq 0$, then $\beta_1 \neq \mathsf{E}[\delta_X]$.
2. If (A1, A2) and $\mathsf{Cov}(\delta_i, D_i) = 0$, then $\beta_2 = \mathsf{E}[\delta_X]$; however, if $\mathsf{Cov}(\delta_i, D_i) \neq 0$, then $\beta_2 \neq \mathsf{E}[\delta_X]$... but it is likely closer.
3. If (A1, A2), then
   - $\mathsf{E}[\delta_X \mid X_i = a] = \beta_3$
   - $\mathsf{E}[\delta_X \mid X_i = b] = \beta_3 + \mu_b$
   - $\mathsf{E}[\delta_X \mid X_i = b] = \beta_3 + \mu_c$.

   Thus, when we fully interact D and X, then we can ignore the $\mathsf{Cov}(D, X)$ assumption.

Matching

Matching estimators assume there is heterogeneity in treatment effects that is based on some set of observed variables. Essentially, (1) one estimates the a regression of Y on D for each value of X, and then saves the β_x, and (2) calculates the X-probability weighted average of β_x to get δ_{ATE}. Note, if one calculates the X-probability-conditional-on-$D = 1$ weighted average of β_x, then one gets δ_{ATT}.

The key is that one needs to turn all the covariates into discrete groups. For example, if covariates are (1) parents' income, (2) test scores, and (3) education level, then all three variables need to be turned into categorical variables and then one needs to create an overall categorical variable. For example,

- $X = 1$ if $PI \in [0, 50k]$ and $TS \in [80, 90]$ and $E =$ Less HS
- $X = 2$ if $PI \in [0, 50k]$ and $TS \in [80, 90]$ and $E =$ HS
- $X = 2$ if $PI \in [0, 50k]$ and $TS \in [80, 90]$ and $E =$ BA+
- $X = 4$ if $PI \in [0, 50k]$ and $TS \in [90, 100]$ and $E =$ Less HS
- $X = 5$ if $PI \in [0, 50k]$ and $TS \in [90, 100]$ and $E =$ HS
- ...

Then

$$\delta_{ATE} = \sum_x \delta_x \cdot \Pr(X_i = x) \tag{7}$$

$$\delta_{ATT} = \sum_x \delta_x \cdot \Pr(X_i = x \mid D_i = 1) \tag{8}$$

Skip from "Even More..." to Section 3.4.1

You can read section 3.3.3, but don't worry about other things.

Section 3.4.2: Marginal Effects

One somewhat controversial aspect of the MHE is that the book is a strong supporter of OLS even with 'limited dependent variables;' e.g., binary dependent variables, count data dependent variables, or bounded dependent variables. Some examples of these are:

- Binary - yes or no variables: did Luke buy coffee today
- Count - 'how many' variables: how many times did Luke buy coffee today
- Bounded - continuous but with upper and/or lower bounds: what percent of Luke's beverages today were coffee, how much coffee did Luke drink today

MHE argues that as long as we have (quasi-) experimental variation for our treatment, then OLS of the outcome on the treatment will tell us something useful and 'well defined' about the relationship. This is because our definition of potential outcomes and the ATT (as the difference in conditional means) does not depend on any assumptions about the distribution of the dependent variable.

Probit Example

Suppose we observe a treatment $D \in \{0, 1\}$ and outcome $Y \in \{0, 1\}$. Let's assume that Y as having some sort of threshold crossing characteristics to it.

For example, let Y be a question of whether I bought a coffee in the afternoon. Suppose Y is ultimately a question of how much energy I think I need in the afternoon: low $\rightarrow$ buy; high $\rightarrow$ not buy. Now let's say D is whether I was given a surprise task in the morning (for example, suppose

I need to reschedule a seminar because the speaker's flight was delayed). If I get an expected task, then I might feel more mentally tired in the afternoon. However, how tired I am is also related to the work I planned to do and how much sleep I got the night before. Thus, we can hypothesize that there is some relationship between Y and D but that D is plausibly random relative to my baseline amount of energy.

We will formalize this example. Let's say that we can represent my decision about whether to buy a coffee, Y, is a threshold crossing function of a latent variable of my tiredness, $Y^\star$. Thus:

$$Y = 1[Y^\star > 0], \tag{9}$$

where we are normalizing the equation such that if my latent variable $Y^\star$ is greater than zero, then I buy the coffee.

Now, we can suppose that this latent variable has the following form:

$$Y^\star(D) = Y^\star(0) + \gamma_1 D = \gamma_0 + \gamma_1 D + u, \tag{10}$$

where $Y^\star(0)$ is my potential latent outcome without treatment, which we assume is equal to $\gamma_0 + u$.

Now, we start making distributional assumptions. If $Y^\star(0) \sim \mathcal{N}(\gamma_0, \sigma_u^2)$, then $u \sim \mathcal{N}(0, \sigma_u^2)$. When this is case, then we can write the conditional expected value of Y as:

$$\mathsf{E}[Y \mid D] = \Phi\left(\frac{\gamma_0 + \gamma_1 D}{\sigma_u}\right), \tag{11}$$

where $\Phi(\cdot)$ is the Normal Distribution CDF.

It turns out we can rewrite the conditional expectation in equation 11 as:

$$\mathsf{E}[Y \mid D] = \Phi\left(\frac{\gamma_0}{\sigma_u}\right) + \left(\Phi\left(\frac{\gamma_0 + \gamma_1 D}{\sigma_u}\right) - \Phi\left(\frac{\gamma_0}{\sigma_u}\right)\right) \cdot D. \tag{12}$$

Note, given equation 11, this is just: $\mathsf{E}[Y \mid D] = \mathsf{E}[Y \mid D = 0] + (\mathsf{E}[Y \mid D = 1] - \mathsf{E}[Y \mid D = 0]) \cdot D$. The key insight about equation 12 is that this looks like the the OLS conditional expectation relationship if we write it as:

$$\mathsf{E}[Y \mid D] = \alpha + \beta D, \tag{13}$$

where we have relabeled the terms as regression coefficients, which makes sense given what the objects are.

Remember, this is just for me, so we should write α_{Luke} and β_{Luke}. If we had data on many people, then the regression would be averaging over all the people, so we should get: $\beta_{\text{OLS}} = \mathsf{E}[\beta_i]$. Because of this, some people will refer to this as an Average Marginal Effect; because the difference embedded in the OLS coefficient is approximately like an average derivative. MHE talks a little about this around equation 3.4.8 on page 104.

Comment: I hope this example highlights two things: (1) how easily we can develop a model for a question (*Will Luke by afternoon coffee?*), and (2) that regression coefficients are not just fixed numbers estimate but take on a meaning from that model and represent relationships between variables. For example: if we had a different assumption about u (e

The rest of the chapter

The rest of the chapter talks about more complicated settings (e.g., what if the variable is either 0 or continuous and positive; like hours worked), adding covariates, and explaining the origin of the term 'regression.'