Week 2: OVB

JJ Chen

January 23, 2015

Anoucement: Today's Tasks

- OVB:
  - Regression Anatomy and OVB formula
  - Typical Usage of OVB formula
  - Bad controls
- Examples:
  - *Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables* (Dale and Krueger (2002))
  - *Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools* (Altonji, Elder, and Taber (2005))

Regression Review

- What kind of regressions do we want to run?
- By regression justifications #1 and #3 in MHE, a regression "gives" us a CEF
  - but most of the time, we only interested in one kind of CEF
  - we want a CEF that can be supported by a believable CIA, so that we will comparing like with like
  - this is about the same as runing a regression without OVB
- Questions about CIA?
  - recall from last section, CIA is hard to justified in observational studies

Regression Anatomy (the FWL Theorem)

- Consider a bivariate population regression: $\mathrm{Y}_i = \alpha + \beta \mathrm{D}_i + e_i$
  - take covariance in either side, $\mathrm{Cov}(\mathrm{Y}_i, \mathrm{D}_i) = \mathrm{Cov}(\alpha + \beta \mathrm{D}_i + e_i, \mathrm{D}_i) \implies$
  - $\beta = \frac{\mathrm{Cov}(\mathrm{Y}_i, \mathrm{D}_i)}{\mathrm{Var}(\mathrm{D}_i)}$ (how about α?)
- Consider a multivariate population regression: $\mathrm{Y}_i = \alpha + \beta \mathrm{D}_i + \mathrm{X}_i' \gamma + e_i$
  - regression anatomy; how to get β?
  - Step 1 (optional): run a regression of Y_i on X_i and get the residual $\tilde{\mathrm{Y}}_i$
  - Step 2: run a regression of D_i on X_i and get the residual $\tilde{\mathrm{D}}_i$
  - Step 3: run a bivariate regression of $\tilde{\mathrm{Y}}_i$ on $\tilde{\mathrm{D}}_i$
- Thus $\beta = \frac{\mathrm{Cov}(\tilde{\mathrm{Y}}_i, \tilde{\mathrm{D}}_i)}{\mathrm{Var}(\tilde{\mathrm{D}}_i)}$, or $\beta = \frac{\mathrm{Cov}(\mathrm{Y}_i, \tilde{\mathrm{D}}_i)}{\mathrm{Var}(\tilde{\mathrm{D}}_i)}$ (Why?)

OVB Formula

- Consider a short and a long regression:
  - $\mathrm{Y}_i = \alpha^s + \beta^s \mathrm{D}_i + e_i^s$
  - $\mathrm{Y}_i = \alpha^l + \beta^l \mathrm{D}_i + \gamma^l \mathrm{X}_i + e_i^l$
- Taking covaraince for the short regression and substituding the long regression gives a simple OVB formula
  - $\beta^s = \frac{\mathrm{Cov}(\mathrm{Y}_i, \mathrm{D}_i)}{\mathrm{Var}(\mathrm{D}_i)} \implies$
  - $\beta^s = \frac{\mathrm{Cov}(\alpha^l + \beta^l \mathrm{D}_i + \gamma^l \mathrm{X}_i + e_i^l, \mathrm{D}_i)}{\mathrm{Var}(\mathrm{D}_i)} \implies$
  - $\beta^s = \beta^l + \gamma^l \frac{\mathrm{Cov}(\mathrm{X}_i, \mathrm{D}_i)}{\mathrm{Var}(\mathrm{D}_i)} = \beta^l + \gamma^l \pi_{\mathrm{XD}}$
    - where π_{XD} is from the auxiliary regression $\mathrm{X}_i = \phi + \pi_{\mathrm{XD}} \mathrm{D}_i + u_i$
- short = long + OVB = long + the effect of omitted $\times$ the regression of omitted on include (what if more than one OV?)

Typical Usage of the OVB Formula

- Often times, omitted variables are variables we wish we had
- OVB formula is thus used to predict the likely direction of bias
  - what's the sign of ability bias in a wage-schooling regression?
  - ability is positively correlated with earning
  - theory predicts that correlation between ability and schooling can be ambiguous (why?)
- More exercises:
  - $\text{sex}_{ir} = \alpha + \rho \text{HIV}_r + \text{X}'_{ir}\boldsymbol{\gamma} + u_i$
  - $\text{obisity}_{ir} = \alpha + \rho \text{residential}_\text{s}\text{prawl}_r + \text{X}'_{ir}\boldsymbol{\gamma} + u_i$
- Some time we want to estimate the extent of OVB
  - how large is the OVB if we want to explain away the entire estimate of your β^s
  - catholic school example

Bad Controls: What Should We Control for

- Recall from CIA, a good control X_i allows us to say:
  - conditional on X_i, the assignment of D_i is as good as ramdomly assigned
  - $E(Y_0 \mid D_i, X_i) = E(Y_0 \mid X_i)$
  - $E(\eta_i \mid D_i, X_i) = E(\eta_i \mid X_i)$
- By the logic of OVB, it seems that we should control for all variables X_i that are correlated with both the outcome variable Y_i and the treatment D_i
  - but we also said that X_i should be pre-treatment variables
  - what happened if we controls for post-treatment variables (or intermediate outcomes)?

Bad Controls: Examples

- Example: suppose we want to know the impact of parents' occupations on kids' test scores: $\mathrm{Y}_i = \alpha + \beta \mathrm{occu}_i + e_i$
  - should we control for family incomes (potentially affected by occupations)?
  - is it a good control in terms of giving $\mathrm{E}\left(\eta_i \mid \mathrm{occu}, \mathrm{income}\right) = \mathrm{E}\left(\eta_i \mid \mathrm{income}\right)$?
  - does it make sense to compare two people with the same income (say 200K) but different occupations? (give an example)
- Example from MHE: should we control for occupations in a wage-schooling regression?
  - occupation is correlated with both the treatment (years of schooling) and outcome (earning)

Bad Controls: Arguments

- Intuitively, having more schooling allow one to work in a high-paying job, which is part of the return of schooling
  - controlling for occupation eliminates this benefit
  - the channel by which wages are increased is partly through occupations
- More importantly, controlling for occupations introduces selection bias since it changes the composition of treatment group and control group

Bad Controls: Composition Effects

Table 1:Bad controls and slection bias

Person	Y_{0i}	Y_{1i}	W_{0i}	W_{1i}
A	*1*	1.5	Blue	Blue
B	*2*	*2.5*	Blue	White
C	3	*3.5*	White	White

- ATE: 0.5
- Controlling for occupations: 0

The Question and Selection Problem

- Let's discuss Dale and Kruger's paper and review basic concepts in this week
- Question: "Does the 'quality' of the college that students attend influence their subsequent earnings?"
- Selection Problem:
  - Can naive comparisons between students from selective colleges and non-selective colleges answer the question?
  - What is the possible selection bias?
  - If we control for family income, gender, race, SAT scores, and high school ranking, will the selection bias dispear?
  - If not, what will drive the bias term and how will you predict the sign of the bias?

Solutions to the Selection Problem

- Choices made by school admission comittees and students in the application, screen, and match process might reveal useful information
  - when applying for schools, students reveal their potential ability by the choice of schools they apply to
  - admission comittees have the chance to read students' essays, letters of recommendation, etc
  - given application outcomes, the decision of attending a particular school again reveals unobserved characteristics
- Extend "selection on the observables" to "selection on the observables and unobservables"
  - "match-applicant model"
  - "self-revelation model"

Discussion

- Table 1: Information from Application Results
  - What does the application results tell about unobservables
- Table 3: The effect of college selectivity on earnings
  - Selection-on-observables and selection-on-unobservables
  - How does the matched/rejected results and students' application choices reveal unobserved characteristics

References I

Altonji, Joseph G, Todd E Elder, and Christopher R Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113 (1). ERIC: 151.

Dale, Stacy Berg, and Alan B Krueger. 2002. "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables*." *The Quarterly Journal of Economics* 117 (4). MIT Press: 1491–1527.