14.74 Recitation 4

Categorical Variables, Joint determination problem, Omitted Variables Bias

Maheshwor Shrestha

October 04, 2013

Agenda

- Regression with categorical variable
- Joint determination
- Omitted Variables Bias

Regression with dummy variable

- Last recitation, we had $x_i \in \{0, 1\}$ and the regression

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- What was the interpretation of β?
  - β is the difference in mean between the treated $(x = 1)$ and the control $(x = 0)$ groups
- What happens when x_i takes discrete values?
  - For example, in a survey people are asked to assess their own health and people pick: very good, good, fair, bad, very bad
  - People are asked to rate the quality of the doctors they visited on a scale of 1 to 5
  - People choose the type of health care providers: government doctors, private doctors, bhopas
- In these context, the numerical values generally does not have any significance.

Regression with categorical variable

- For simplicity, assume that x_i takes 3 values $\{0, 1, 2\}$. How could we run a regression of some outcome on x?
  - How about
    $$y_i = \alpha + \beta x_i + \varepsilon_i$$
    what is wrong with this regression? Limitations?
- Suppose you create 3 dummy variables
  - $C0_i = 1\,(x_i == 0)$
  - $C1_i = 1\,(x_i == 1)$
  - $C2_i = 1\,(x_i == 2)$
- and you run the regression
  $$y_i = \alpha + \beta_0 C0_i + \beta_1 C1_i + \beta_2 C2_i + \varepsilon_i$$
  What is wrong with this regression?

Interpretation

- The previous regression suffers from perfect collinearity

$$C0_i + C1_i + C2_i = 1$$

  and we cannot run this regression.

- Instead, drop one of the categories (say $C0$) and run

$$y_i = \alpha + \beta_1 C1_i + \beta_2 C2_i + \varepsilon_i$$

  what do the coefficients give? Do we need an interaction term?
    - α: mean outcome for category 0
    - β_1: difference in mean outcome between category 1 and category 0
    - β_2: difference in mean outcome between category 2 and category 0

- Which category to omit?
    - depends upon how you want to interpret. Usually omit the "default" category or the largest category

Interpretation of regression coefficients

- Suppose you estimate the regression

$$Y_i = \alpha + \beta \cdot M_i + \gamma \cdot E_i + \varepsilon_i$$

  where ε_i is a random error term not correlated with any variables (all OLS assumptions are satisfied, no selection bias, omitted variables or reverse causality)

- What is the interpretation of β? Of γ?
  - β tells us how much will Y change if M changes by 1 unit, holding E constant

$$\beta = \frac{\partial Y}{\partial M}$$

  - γ tells us how much will Y change if E changes by 1 unit, holding M constant

$$\gamma = \frac{\partial Y}{\partial E}$$

Joint determination

- Suppose malaria (M) causes education (E) to go down and also per-capita income (Y) to go down
  - Malaria jointly determines education and per-capita income
- Suppose education also leads to higher per-capita income
- Suppose there are no other causal channels, reverse causality and or omitted variables. You estimate a cross-country regression

$$Y_i = \alpha + \beta \cdot M_i + \gamma \cdot E_i + \varepsilon_i$$

- What is the interpretation of β? Of γ?

Interpretation

- β tells us how much will Y change if M changes by 1 unit, holding E constant
  - But if M changes, then it will cause E to change...
- So, the total causal effect of malaria on income is

$$\frac{\partial Y}{\partial M} = \underbrace{\beta}_{<0} + \underbrace{\gamma}_{>0} \underbrace{\frac{\partial E}{\partial M}}_{<0}$$

- Is β an overestimate or underestimate the magnitude?
  - β will be smaller in magnitude than the total effect because it ignores the second part (underestimate)

Bad controls? Bad regression?

- What if we just dropped E from our equation. That is, estimate

$$Y_i = \pi_0 + \pi_1 M_i + \varepsilon_i$$

- Assuming our previous assumptions still hold, will omitting education, E create a problem? What is the interpretation of π_1? Will there be an omitted variable bias?
- When is it a bad idea to omit E?
  - When changes in education causes malaria to change. That is, when education jointly determines malaria (the regressor) and income (outcome)
- What does the real world look like?
  - Suppose Education also causes malaria reduction. And you run the original specification. How do you interpret the coefficients?

Omitted Variables Bias

- Suppose that world is simple as before:
  - Malaria causes a fall in education and also a fall in per-capita income
  - Education causes rise in per-capita income
- And you run the regression

$$Y_i = \theta_0 + \theta_1 \cdot E_i + \varepsilon_i$$

  to find the causal effect of education on income
  - Note that here you have omitted Malaria which has a causal effect both on education and income
- Will θ_1 be a causal impact of the world?
  - No. Because changes in malaria causes income and education to co-move, and hence causing a spurious correlation between them. This is the **omitted variable bias**.

OVB Formula

- The true model of the world (our simple world from last slide) is

$$Y_i = \alpha + \gamma \cdot E_i + \beta \cdot M_i + \varepsilon_i$$

  and we wanted to estimate γ
- But the regression we ran was:

$$Y_i = \theta_0 + \theta_1 \cdot E_i + \varepsilon_i$$

- What will θ_1 tell us?

$$\begin{aligned}
\theta_1 &= \frac{Cov\left(Y_i, E_i\right)}{Var\left(E_i\right)} \\
&= \frac{Cov\left(\alpha + \gamma \cdot E_i + \beta \cdot M_i + \varepsilon_i, E_i\right)}{Var\left(E_i\right)} \\
&= \frac{Cov\left(\alpha, E_i\right) + Cov\left(\gamma \cdot E_i, E_i\right) + Cov\left(\beta \cdot M_i, E_i\right) + Cov\left(\varepsilon_i, E_i\right)}{Var\left(E_i\right)}
\end{aligned}$$

OVB Formula (cont...)

$$\theta_1 = \frac{Cov(\alpha, E_i) + Cov(\gamma \cdot E_i, E_i) + Cov(\beta \cdot M_i, E_i) + Cov(\varepsilon_i, E_i)}{Var(E_i)}$$
$$= \frac{\gamma \cdot Cov(E_i, E_i) + \beta \cdot Cov(M_i, E_i)}{Var(E_i)}$$
$$= \gamma + \beta \frac{Cov(M_i, E_i)}{Var(E_i)}$$

- The first term γ is the true causal effect of education on income
- The second term $\beta \frac{Cov(M_i, E_i)}{Var(E_i)}$ is the omitted variable bias
  - Depends upon the effect of malaria on income β AND
  - Depends upon the effect of malaria on education $\frac{Cov(M_i, E_i)}{Var(E_i)}$
    - Note that this term is the regression coefficient of malaria on education

When OK to omit?

- It is ok to omit variable z_i from regression of y_i on x_i when:
  - z_i does not have a causal impact on y_i
  - z_i does not have a causal impact on x_i
- It is also okay to omit variable z_i from regression of y_i on x_i when x_i has a causal effect on z_i and z_i has a causal effect on y_i (and no other causalities)
  - This is the case where leaving z_i from the regression gives the total causal impact of x_i on y_i