OLS Efficiency & Using Qualitative Data

Tyler Ransom

Univ of Oklahoma

Feb 5, 2019

Today's plan

1. Review reading topics

    1.1 Efficiency of OLS Estimators

    - The Gauss-Markov Theorem

    1.2 How to Use Qualitative Data

    - Dummy Variables
    - The "Dummy Variable Trap"
    - Linear Probability Models

2. In-class activity: Practice with dummy variables

Efficiency of OLS Estimators

The Gauss-Markov Theorem

- Under the Gauss-Markov Assumptions:
- OLS estimator $\hat{\beta}_0, \dots \hat{\beta}_k$ is the **best linear unbiased estimator (BLUE)**
- What is BLUE? Working backwards:

E: estimator—a rule to compute an estimate from a sample of data

U: unbiased— $E(\hat{\beta}_j) = \beta_j, j = 0, 1, \dots, k$

L: linear—the estimator is a linear function of y

B: best—has the **lowest sampling variance**

Efficiency

- On the second day of class we talked about efficiency:

- An estimator is **efficient** if it has a lower sampling variance than all other estimators

- Thus, $Var(\hat{\beta}_j) < Var(\tilde{\beta}_j)$ for all j, where $\tilde{\beta}_j$ is an alternative estimator

- This is what we mean by “best”

- What’s so great about OLS being efficient?

- Usually efficient estimators are not as simple to compute as $\hat{\beta}_j$!

How to Use Qualitative Data

Describing Qualitative Information

- Until now, all examples have used continuous variables, numerical values
- How to we describe binary qualitative information? (e.g. Yes/No)
  - A worker belongs to a union or does not
  - A firm offers a 401(k) pension plan or it does not
- Can be captured by defining a **binary variable** (or **dummy variable**)
- Must decide which outcome is assigned zero, which is one
- Choose variable name to be descriptive

Example

- to indicate gender, *female*, which is one if the person is female, zero if the person is male

- This is a better name than *gender* or *sex* (what does *gender* $= 1$ mean?)

colgpa	sat	hsperc	athlete	female	sex
3	810	66.66667	0	1	female
3.41	1110	96.2963	1	0	male
2.84	870	54.05405	1	1	female
3.61	1020	78.78788	0	1	female
2	860	79.62963	0	1	female
2.86	1150	81.81818	1	0	male

Why 0/1 and not some other pair of values?

- For distinguishing different types, any two different values would do
- But as we will see, 0/1 is convenient for use in regression analysis
- Also: can create more than two categories from multiple qualitative vars
- e.g. female athlete, female non-athlete, male athlete, male non-athlete

Single dummy variable

- Example: simple regression where x is binary

$$wage = \beta_0 + \delta_0 female + u$$

- Assuming $E(u|female) = 0$ holds,

$$E(wage|female) = \beta_0 + \delta_0 female$$

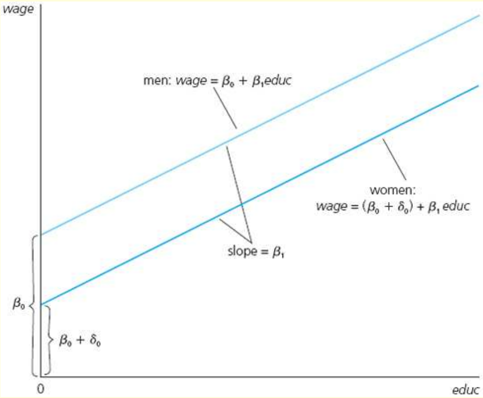
  is the population regression function

- with two values of *female* (0 and 1),

$$E(wage|female = 0) = \beta_0 + \delta_0 \cdot 0 = \beta_0$$
$$E(wage|female = 1) = \beta_0 + \delta_0 \cdot 1 = \beta_0 + \delta_0$$

- $\overline{wage}$ for men is β_0, for women is $\beta_0 + \delta_0$. δ_0 **is the difference on average**

Visualizing the dummy variable



- Visualization of the PRF of the equation

  $$wage = \beta_0 + \delta_0 female + \beta_1 exper + u$$

- δ_0 measures the gender difference in wages holding fixed *exper*

Properties of dummy variables

- Put in $M - 1$ dummy variables for a variable with M categories
- If put in M dummies, known as **dummy variable trap**
- Changing base group won't change estimates of non-dummy coeffs
- Will change sign (but not magnitude) of dummy variable coefficients
- Will change mangitude (and possibly sign) of intercept

Interpretation of dummy coefficients

- Interpret as difference in group means, holding fixed other x's

- Interpretation always relative to base group

- If y is in logs, interpretation is **approximately** % difference

$$\widehat{\log(wage)} = 2.413 - .343 female$$
$$n = 750, R^2 = .122$$

  so women earn about 34.3% less than men in this sample

- Sometimes you'll hear "34.3 log points" to distinguish from proper % change

Multiple categories

- Can have multiple groups if data has multiple qualitative variables
- e.g. marital status (Married/Single) and sex (Female/Male)
- Now there are four groups, so we should put in 3 dummies

Interpretation with multiple categories

- regress $\log(wage)$ on *female*, *married*, *female* $\times$ *married*, other x's

$$\widehat{\log(wage)} = .321 - .110 female + .213 married - .301 female \times married + \ldots$$

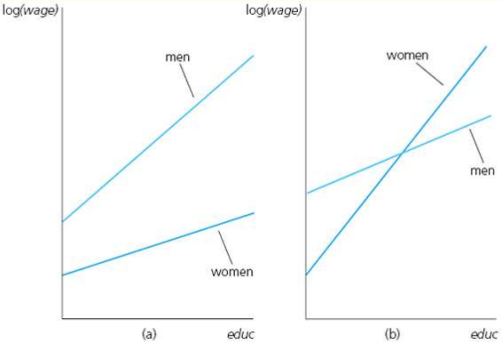
$$n = 526, R^2 = .461$$

- Need to be careful about plugging in 1's and 0's for different groups
- single M are reference group, their intercept $= .321$
- single W: $.321 - .110 = .211$ or 11% lower than single M
- married M: $.321 + .213 = .534$ or 21.3% more than single M
- married W: $.321 - .110 + .213 - .301 = .123$ or 41.1% less than married M

Allowing for different slopes

- Can also use dummies to allow for different slopes
- Multiply a dummy with a continuous variable
- This is known as an **interaction term**
- (Actually *female* $\times$ *married* is also an interaction term)
- Any product of two other variables is called an interaction term

Visualizing different slopes



- Visualization of the PRF of the equation

$$\log(wage) = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female) \times educ + u$$

- δ_0 is difference in intercept
- δ_1 is difference in slope

Qualitative data in R

- In R, qualitative data is handled by *factors*
- You define the levels of the factor (e.g. Yes/No, North/East/South/West)
- You tell R which level is the baseline
- R will handle the rest
- Practice with this in today's lab

Binary *y*

- Until now, we've talked only about dummy x's
- What about dummy y's?
- e.g. employed (Y/N), arrested (Y/N), etc.
- Here, the outcome is binary

Interpretation of β when y is binary

- How do we interpret the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

  when y is binary? y can only go from 0 to 1 (or 1 to 0)

- Key relationship when y is binary:

$$E(y|\mathbf{x}) = P(y = 1|\mathbf{x})$$

  where $P(y = 1|\mathbf{x})$ the **response probability**

- all partial effects are effects on the probability that $y = 1$

The Linear Probability Model

- We call

$$P(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

  the **linear probability model (LPM)**

- A change in x thus changes the probability that $y = 1$:

$$\Delta P(y = 1|\mathbf{x}) = \beta_j \Delta x_j, \text{ holding other } x\text{'s fixed}$$

- $\widehat{y}$ is now a predicted probability

Pros and Cons of LPM

- Pros:
  - Easy to estimate β's
  - Easy to interpret β's
- Cons:
  - Possible for $\widehat{y} < 0, \widehat{y} > 1$ (negative probability!?)
  - β's are constant through range of x's...
  - possibly leading to silly Δy's for a large Δx
- Overall, LPM is great if you care about partial effects and not prediction!