Lecture 2

Limited Variables 1 of 5: Binary (Dummy) **Explanatory** Variables (Dummy Variables I)

Michael Curran

Trinity College Dublin

JS Econometrics

Lecture 2 Outline

Introduction

Overview of Limited Variables

Lecture 2 Outline

Qualitative Information

- Quantitative variables: magnitudes convey useful information. Examples?
- Qualitative variables: e.g. gender, race, industry, region, etc.
- Binary information: **binary variable** / zero-one variable / **dummy variable** (DV). Examples?
- How to define a DV? Decide which event to assign the value 1 and which event to assign the value 0. NB: clear variable names can be helpful!
- When investigating real GDP growth performance differences between Eurozone members and countries outside the Eurozone, is the name *country* a wise choice for a DV? What might be a more helpful name?

Lecture 2 Outline

Single Binary Independent Variables

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u \tag{1}$$

Intercept shift:
Note $E(u|female, educ) = 0$ implies

$$\delta_0 = E(wage|female = 1, educ) - E(wage|female = 0, educ)$$

or rewriting more simply since $female = 1$ corresponds to females and v.v. for 0,

$$\delta_0 = E(wage|female, educ) - E(wage|male, educ)$$

Key: level of educ same in both expectations; difference δ_0 due to gender only.
Dummy variable trap

Single Binary Independent Variables

Base group / **benchmark group**: group against which comparisons are made. Write the model so to choose females as the base group:

$$wage = \alpha_0 + \gamma_0 male + \beta_1 educ + u$$

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

Testing? Estimate model by OLS and use usual t statistic.
Difference? Interpretation of coefficient on DV.

Single Binary Independent Variables

Examples

Logarithmic Dependent Variables

- DV coefficients have a *percentage* interpretation when the dependent variable is $\log(y)$.
- Approximation (adequate when coefficient only suggests a *small* proportionate changes in y): $100 \times \hat{\delta}_i$ where $\hat{\delta}_i$ is the estimate of the coefficient of the DV.
- Generally, if $\hat{\beta}$ is the coefficient on DV (x_1) when $\log(y)$ is the dependent variable, the exact percentage difference in predicted y when $x_1 = 1$ versus when $x_1 = 0$ is

$$100 \cdot [\exp(\hat{\beta}_1) - 1] \tag{2}$$

- The estimate $\hat{\beta}_1$ can be positive or negative and it is important to preserve its sign in computing (2).

Lecture 2 Outline

Multiple Categories

$$\log(wage) = \beta_{0,1} + \delta_{1,1} female + \beta_{1,1} educ + \beta_{2,1} exper + \beta_{3,1} tenure + u_1 \quad (3)$$

$$\log(wage) = \beta_{0,2} + \delta_{1,2} female + \delta_{2,2} married + \beta_{1,2} educ + \beta_{2,2} exper + \beta_{3,2} tenure + u_2 \quad (4)$$

$$\log(wage) = \beta_{0,3} + \delta_{1,3} marrmale + \delta_{2,3} marrfem + \delta_{3,3} singfem + \beta_{1,3} educ + \beta_{2,3} exper + \beta_{3,3} tenure + u_3 \quad (5)$$

- Limitation of (4): marriage premium ($\delta_{2,2}$) assumed to be identical across gender, but relaxed in (5).
- Which group is base group?
- Have we avoided the DV trap?
- Possible to estimate differences between any two groups, but have to go further to test if the difference is statistically significant!
- DV trap: if there are g groups or categories, include $g - 1$ DV plus an intercept.

Multiple Categories

Ordinal Information

Ordinal variables – examples? Credit ratings.

Wrong way to incorporate ordinal variables into models:

$$MBR = \beta_0 + \beta_1 CR + otherfactors$$

If we don't want to assume constant partial effects, since CR only takes on a few values, define a DV for each value of CR:

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + otherfactors \quad (6)$$

Constant partial effect is contained in special case of (6):

$$MBR = \beta_0 + \delta_1 (CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + otherfactors$$

F statistic for testing constant partial effects restriction:

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n - k - 1)} = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/df_{ur}}$$

What about when an ordinal variable takes on too many values? Partition and test.

Lecture 2 Outline

Interactions

Introduction

$$\log(wage) = \beta_0 + \delta_1 female + \delta_2 married + \delta_3 female \cdot married + \ldots$$

- Marriage premium depends on gender like in (5).
- Need to add coefficients, e.g. intercept for married men found by setting $female = 0$ and $married = 1$, then adding $\beta_0 + \delta_2$.
- Be careful plugging in the right combination of zeros and ones!
- Easier for testing certain hypotheses. . . but trickier for others!

Interactions

Slope Differentials

Interact DV with non-DV explanatory variables to allow for **differences in slopes**.

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

To estimate by OLS. write with interaction between men and women:

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

Return to education same across gender: H_0: $\delta_1 = 0$
Average wages conditional on education are identical across gender: F test of $H_0 : \delta_0 = 0, \delta_1 = 0$.
Complications.

Interactions

Testing

H_0: two populations or groups follow same regression function.
H_1: one or more of the slopes differ across the groups.
Test if same regression model describes college GPA for male and female college athletes:

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u$$

Intercept difference: include DV for males or females. Slope to depend on gender: interact appropriate variable with e.g. *female*. *Any* difference between men and women: need to allow for intercept and all slopes to be different across the two groups.

$$cumgpa = \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u$$

H_0: *cumgpa* follows same model for males and females:

$$H_0 : \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$$

Interactions

Testing

Complications:

1. F statistics preferred to individual t statistics for testing joint hypotheses.
2. Interpreting: interaction terms matter when obtaining differences between groups.
3. What about when we have many independent variables? Tedious to have many interaction terms.

Solution to third point: use SSR form of F statistic. Let k be the number of explanatory variables, include an intercept and two groups $g = 1$ and $g = 2$. Test whether intercept and all slopes are same across two groups:

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + \ldots + \beta_{g,k}x_k + u \quad g = 1, 2$$

Interactions

Testing

Chow statistic:

$$F = \frac{[SSR_p - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k+1)]}{k+1} \tag{7}$$

Only valid under homoscedasticity; normality not needed for asymptotic analysis. F statistic is same as R squared form in models with and without interaction terms. Limitation: null hypothesis allows for no differences (intercept and slope) between the groups. How can we allow intercepts to differ under the null hypothesis?

1. Include the group dummy and all interaction terms but then test the joint significance of interaction terms only.
2. Form F stat as in (7) but where restricted sum of squares SSR_p is obtained by the regression that allows an intercept shift only, i.e., run a pooled regression and just include the DV that distinguishes the two groups.

Lecture 2 Outline

Summary

- Qualitative information can be described by variables such as binary variables.
- Important to avoid dummy variable trap!
- Standard estimation (OLS) and tests.
- Base group changes interpretation – keep in mind.
- Interesting applications via interactions.

References

- Dummy Independent Variables: Wooldridge 7.1-4.