ECNS 561

Qualitative Information in Multiple Regression Analysis

Qualitative Information

- What are some examples of information that may come in binary form (i.e., “dummy” variables)?
  - At individual level
    - Gender
    - Race
    - Education levels
    - Married
  - At county or state level
    - Whether or not a certain law is in place
    - Geographic location

Consider the following wage equation

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \varepsilon$$

Q. What is the interpretation of the coefficient estimate for *female*?

Ans. This is the ***level*** difference between wages for men and women.

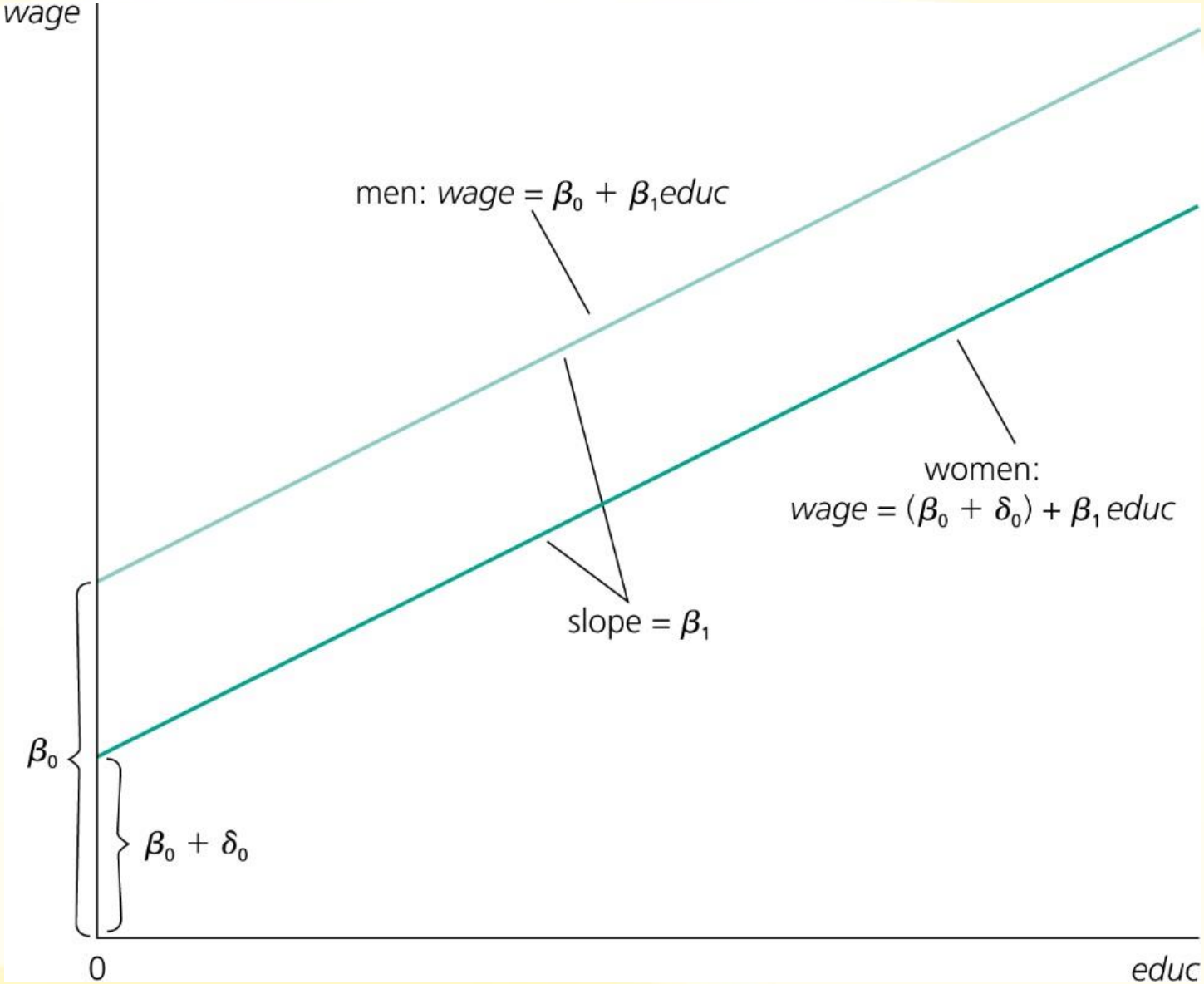
Q. If we can satisfactorily control for productivity, what does else does this coefficient estimate tell us?

Ans. If $\delta_0 < 0$, holding all else equal, there is discrimination against women.

-We can write δ_0 as

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

Q. δ_0 simply represents a shift in which parameter?


wage
men: wage = β_0 + $\beta_1$educ
women:
wage = (β_0 + δ_0) + β_1 educ
slope = β_1
β_0
β_0 + δ_0
0
educ

- Q. Why do we also not control for a dummy variable *male*?

Ans. Because this would introduce perfect collinearity

$$female + male = 1$$

- Q. What happens to the gap between our wage estimates for men and women when we add more relevant control variables, such as *exper* and *tenure*?

What if we specify our dependent variable as log(y)?

-Consider our KS crime equation:

$$\log(violent\ crime_c) = \beta_0 + \beta_1 Wet\_Law_c + \boldsymbol{X_c}'\boldsymbol{\beta_2} + \varepsilon$$

where *violent crime* is the violent crime rate in county *c*, *Wet_Law* is equal to 1 if county *c* allows by-the-drink alcohol sales for on-premises consumption, and ***X*** is a vector of county-level attributes.

-Suppose when we estimate the above equation with data from all 105 KS counties, we obtain $\widehat{\beta_1}$ = .102.

-So, wet counties have about 10.2% more violent crime than dry counties (using our percentage change approximation)

-But, we can be more precise than this and compute the exact percentage difference in predicted crime.

-What we want is the proportionate difference in crime between wet and dry counties

$$(\widehat{violent\ crime}_{Wet} - \widehat{violent\ crime}_{Dry})/\widehat{violent\ crime}_{Dry}$$

-From our coefficient estimate on *Wet_Law*, what we really have is

$$\log(\widehat{violent\ crime}_{Wet}) - \log(\widehat{violent\ crime}_{Dry}) = .102$$

-Exponentiating and subtracting 1 from both sides, we have

$$(\widehat{violent\ crime}_{Wet} - \widehat{violent\ crime}_{Dry})/\widehat{violent\ crime}_{Dry} = \exp(.102) - 1 = .107$$

-So, this more accurate estimate implies that a wet county is, on average, going to have a 10.7% higher violent crime rate than a dry county.

-In general,

$$100*[\exp(\widehat{\beta_1}) - 1]$$

Dummy variables for multiple categories

- If the regression model is to have different intercepts for g groups, then we need to include $g - 1$ dummies.
- Let's say we expect there to be a regional effect in our wet law and crime examination (SW KS, NW KS, SE KS, SW KS)
- Q. Why might this matter?
- Ans. Maybe region proxies for religiosity, urbanicity, voting preferences, etc. that may influence whether or not a county votes to go wet and is simultaneously correlated with crime rates.

- Suppose, we have

$$\log(violent\ crime_c) = \beta_0 + \beta_1 Wet\_Law_c + \boldsymbol{X_c'\beta_2} + \beta_3 \text{NWdum}_c + \beta_4 \text{NEdum}_c + \beta_5 \text{SWdum}_c + \varepsilon$$

- Q. Why have we only included 3 of the 4 dummies in our model?
- Q. How do we interpret the regression coefficients on our regional dummies?

Ordinal Information in Regression

- Sometimes we will encounter an ordinal variable in regression analysis.

- Suppose we are interested in estimating a wage equation for academic economists as a function of the rank of the department where they received their PhD, where the variable *rank* is defined as:

  = 1 if went to a 4^{th} tier ranked department

  = 2 if went to a 3^{rd} tier ranked department

  = 3 if went to a 2^{nd} tier ranked department

  = 4 if went to a 1^{st} tier ranked department.

- We could enter this variable on its own in the regression as follows:

$$\log(wage) = \beta_0 + \beta_1 rank + \varepsilon$$

- Q. What is the interpretation here?
- Q. Is there a better way to specify this regression?
- A preferred approach might be to create four dummy variables for each tier:

$$\log(wage) = \beta_0 + \beta_1 tier1 + \beta_2 tier2 + \beta_3 tier3 + \varepsilon$$

Interacting Dummy Variables

- Similar to interacting variables with quantitative meaning.
- Consider the following estimated county crime regression

$$\widehat{\log(crime)} = .231 + .121Wet + .098College + .034Wet * College$$

-*crime*: crime rate per 1,000 population in county *c*

-*Wet:* = 1 if county *c* allows by-the-drink alcohol sales, = 0 otherwise

-*College:* = 1 if county *c* is home to a major college campus, = 0 otherwise

- This equation allows the effect of the wet law to depend on whether or not there is a major college campus in the county.
- Four possible scenarios
  - 1.) *Wet* = 1 & *College* = 1
  - 2.) *Wet* = 1 & *College* = 0
  - 3.) *Wet* = 0 & *College* = 1
  - 4.) *Wet* = 0 & *College* = 0

- This equation allows us to obtain the crime rate differentials across all four groups
- The intercept for wet counties without a major college campus would, for example, be

$$.231 + .121 = .325$$

  - Q. How do we interpret this? What is our reference group?
  - Ans. Wet counties without a major college campus have crime rates that are over 30% higher than dry counties without a major college campus.

Allowing for Differing Slopes

- Consider the interaction of *Wet* with an explanatory variable that is not a dummy

$$\log(crime) = \beta_0 + \beta_1 Wet + \beta_2 Income + \beta_3 Wet*Income + \varepsilon$$

-where *Income* is equal to real income per capita in county *c*

- Q. What is the intercept for dry counties? What is the slope on income for dry counties?
  - If we set *Wet* = 0, then we see that the intercept for dry counties is β_0 and the slope on income for dry counties is β_2.
- Q. What is the intercept for wet counties? What is the slope on income for wet counties?
  - If we set *Wet* = 1, then we see that the intercept for wet counties is $\beta_0 + \beta_1$ and the slope on income for wet counties is $\beta_2 + \beta_3$.

Chapter 7 #1

Using individual-level data the following equation was estimated

$$\widehat{sleep} = 3{,}840.83 - .163totwrk - 11.71educ - 8.70age$$
$$(235.11) \quad (.018) \quad (5.86) \quad (11.21)$$
$$+ .128age^2 + 87.75male$$
$$(.134) \quad (34.33)$$

$N = 706$, $R^2 = .123$

where *sleep* is total minutes per week spent sleeping at night, *totwork* is totally weekly minutes spent working, *educ* and *age* are measured in years, and *male* is a gender dummy.

(i) All other factors equal, is there evidence that men sleep more than women? How strong is the evidence?

Ans. Coefficient on *male* is 87.75, so a man is estimated to sleep one and one-half hours more per week than a comparable woman. This is significant at the 1% level.

$$\widehat{sleep} = 3{,}840.83 - .163totwrk - 11.71educ - 8.70age$$
$$(235.11) \quad (.018) \quad (5.86) \quad (11.21)$$
$$+ .128age^2 + 87.75male$$
$$(.134) \quad (34.33)$$

N = 706, R^2 = .123

(ii) Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?

Ans. The coefficient estimate on *totwrk* is negative and significant at the 1% level. One more hour of work is associated with .163(60) ≈ 9.8 minutes less sleep.

(iii) What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?

Ans. The null we are interested in testing here is that the coefficients on *age* and age^2 are jointly zero. So, we would need to run a restricted version of the regression above where these two variables are omitted and calculate…

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1-R_{ur}^2)/(n-k-1)}$$

Chapter 7, #3

Consider the following student-level estimated equation

$$\widehat{sat} = 1{,}028.10 + 19.30hsize - 2.19hsize^2 - 45.09female$$

(6.29) (3.83) (.53) (4.29)

$$- 169.81black + 62.31female*black$$

(12.71) (18.15)

N = 4.137, R^2 = .0858

where *sat* is SAT score, *hsize* is the size of the student's high school graduating class (in hundreds), *female* is a gender dummy variable, and *black* is a race dummy variable.

(i) Is there strong evidence that $hsize^2$ should be included in the model? From these estimates what is the optimal school size?

Ans. Yes, it is significant at the 1% level.

Optimal school size would be 19.3/(2*2.19) ≈ 4.41 (or 441 students)

$$\widehat{sat} = 1{,}028.10 + 19.30hsize - 2.19hsize^2 - 45.09female$$
$$(6.29) \quad (3.83) \quad (.53) \quad (4.29)$$
$$- 169.81black + 62.31female*black$$
$$(12.71) \quad (18.15)$$

$N = 4{,}137$, $R^2 = .0858$

(ii) Holding *hsize* fixed, what is the estimated difference in SAT score between nonblack females and nonblack males? How statistically significant is this estimated difference?

Ans. This is given by the coefficient on *female*: nonblack females have SAT scores about 45 points lower than nonblack males. The *t* stat is about -10.51, so the difference is very statistically significant.

(iii) What is the estimated difference in SAT score between nonblack males and black males? Test the null that there is no difference between their scores, against the alternative that there is a difference.

Ans. Because *female* = 0, the coefficient on *black* implies that a black male has an estimated SAT score nearly 170 points less than a comparable nonblack male. This estimate is clearly precisely estimated.

$$\widehat{sat} = 1{,}028.10 + 19.30hsize - 2.19hsize^2 - 45.09female$$

(6.29) (3.83) (.53) (4.29)

$$- 169.81black + 62.31female*black$$

(12.71) (18.15)

$N = 4{,}137$, $R^2 = .0858$

(iv) What is the estimated difference in SAT score between black females and nonblack females? What would you need to do to test whether the difference is statistically significant?

Ans. Calculating the difference is easy…

-For black females, we plug in *female* = 1 and *black* = 1

-45.09 – 169.81 + 62.31

-For nonblack females, we plug in *female* = 1 and *black* = 0

-45.09

-The difference between these two estimates is simply -169.81+62.31 = -107.50. So, a black female has an estimated SAT score about 108 points lower than a nonblack female.

-Calculating whether this difference is statistically significant is more difficult, because the estimate depends on two coefficients. How might we rewrite our regression equation, so that we could test this?

Ans. The easiest approach would be to define dummy variables for three of the four race/gender combos and choose nonblack females as the reference group (i.e. omitted category)

Chapter 7, #4

$$\widehat{\log(salary)} = 4.59 + .257\log(sales) + .011roe + .158finance$$
$$(.30) \quad (.032) \qquad\qquad (.004) \qquad (.089)$$
$$+ .181consprod - .283utility$$
$$(.085) \qquad\qquad (.099)$$

N=209, R^2= .357

where *sales* is firm sales, *roe* is return on equity, *finance*, *consprod*, and *utility* are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation.

(i) Compute the approximate percentage difference in estimated salary between the utility and transportation industries, holding sales and return on equity fixed. Is this difference significant?

Ans. The rough difference is -28.3% and this is statistically significant.

$$\widehat{\log(salary)} = 4.59 + .257\log(sales) + .011roe + .158finance$$
$$(.30) \quad (.032) \qquad\qquad (.004) \qquad (.089)$$
$$+ .181consprod - .283utility$$
$$(.085) \qquad\qquad (.099)$$

N=209, R^2= .357

(ii) What is the exact percentage difference in estimated salary between the utility and transportation industries and compare this to the answer in part (i)?

Ans. 100*[exp(-.283) – 1] = -24.7%...so the exact difference is slightly smaller in magnitude

(iii) What is the approximate percentage difference in estimated salary between the consumer products and finance industries?

Ans. The proportionate difference is .181 - .158 = .023 or 2.3%.

Q. How can we test whether this is statistically significant?

Ans. Rewrite the equation so that the finance industry is the omitted category.