Econometrics HW #4

Ryan Safner

Due: Wednesday, November 28, 2018

Theory & Concepts

For the following questions, please answer the questions completely but succinctly (2-3 sentences).

1. In your own words, describe what the "dummy variable trap" means. What precisely is the problem, and what is the standard way to prevent it?

---

2. In your own words, describe what an interaction term is used for, and give an example. You can use any type of interaction to explain your answer.

Theory Problems

For the following questions, please *show all work* and explain answers as necessary. You may lose points if you only write the correct answer. You may use `R` to verify your answers, but you are expected to reach the answers in this section "manually."

3. Suppose data on many countries' legal systems (Common Law or Civil Law) and their GDP per capita gives us the following summary statistics:

Legal System	Average GDP Growth Rate	Std. dev	n
Common Law	1.84	3.55	19
Civil Law	4.97	4.27	141
Difference	-3.13	1.02	-

a. Using the group means, write a regression equation for a regression of GDP Growth rate on Common Law, including the standard errors in parentheses below the coefficients. Define

$$\text{Common Law}_i = \begin{cases} 1 & \text{if country } i \text{ has common law} \\ 0 & \text{if country } i \text{ has civil law} \end{cases}$$

b. How do we use the regression to find the average GDP Growth rate for common law countries? For civil law countries? For the difference?

c. Looking at the coefficients, does there appear to be a statistically significant difference in average GDP Growth Rates between Civil and Common law countries?

d. Is the estimate on the difference likely to be unbiased? Why or why not?

e. Now using the same table above, reconstruct the regression equation if instead of Common Law, we had used:

$$\text{Common Law}_i = \begin{cases} 1 & \text{if country } i \text{ has common law} \\ 0 & \text{if country } i \text{ has civil law} \end{cases}$$

4. Suppose a real estate agent collects data on houses that have sold in a particular neighborhood over the past year, with the following variables:

- $Price_h$: price of house h (in thousands of \$)
- Bed_h : number of bedrooms in house h
- $Bath_h$: number of bathrooms in house h
- $Pool_h : \begin{cases} = 1 & \text{if house } h \text{ has a pool} \\ = 0 & \text{if house } h \text{ does not have a pool} \end{cases}$
- $View_h : \begin{cases} = 1 & \text{if house } h \text{ has a nice view} \\ = 0 & \text{if house } h \text{ does not have a nice view} \end{cases}$ \end{itemize}

a. Suppose he runs the following regression:

$$\widehat{Price}_h = \underset{(129.42)}{119.20} + \underset{(9.82)}{29.76}\, Bed_h + \underset{(10.23)}{24.09}\, View_h + \underset{(9.49)}{14.06}\, BDR_h * View_h$$

What does each coefficient mean?

b. Write out two separate regression equations, one for houses *with* a nice view, and one for homes *without* a nice view. Explain each coefficient in each regression.

c. Suppose he runs the following regression:

$$\widehat{Price_h} = \underset{(129.42)}{189.20} + \underset{(9.82)}{42.40}\, Pool_h + \underset{(10.23)}{12.10}\, View_h + \underset{(9.49)}{12.09}\, Pool_h * View_h$$

What does each coefficient mean?

d. Find the expected price for:

- a house with no pool and no view
- a house with no pool and a view
- a house with a pool and without a view
- a house with a pool and with a view

e. Suppose he runs the following regression:

$$\widehat{Price} = \underset{(1.18)}{87.90} + \underset{(0.22)}{53.94}\, Bed + \underset{(0.22)}{15.29}\, Bath + \underset{(0.04)}{16.19}\, Bed * Bath$$

What is the marginal effect of adding an additional **bedroom** if the house has 1 bathroom? 2 bathrooms? 3 bathrooms?

f. What is the marginal effect of adding an additional bathroom if the house has 1 bedroom? 2 bedrooms? 3 bedrooms?

R Problems

Answer the following problems using R. Round to 2 decimal places. If using R Markdown, simply create code chunk(s) for each question and be sure all input code is displayed (i.e. echo=TRUE) and feel free to just turn in a single html or pdf output file for your entire homework.

If you are NOT using R Markdown, please follow our standard procedure: Attach/write the answers to each question on the same document as the previous problems, but also include a printed/attached (and commented!) *.R* script file of your commands to answer the questions.

7. Download the HeightWages.dta dataset from Blackboard (under Data). This data is a part of a larger dataset from the National Longitudinal Survey of Youth (NLSY) 1979 cohort: a nationally representative sample of 12,686 men and women aged 14-22 years old when they were first surveyed in 1979. They were subsequently interviewed every year through 1994 and then every other year afterwards. There are many included variables, but for now we will just focus on:

- wage96: Adult hourly wages ($/hr) reported in 1996
- height85: Adult height (inches) reported in 1985
- height81: Adolescent height (inches) reported in 1981
- male: Dummy variable = 1 if person is male, = 0 if person is female
- hgc96: Highest grade of education completed in 1996 (0-20)

a. Using R to examine the data, find the mean earnings (wage96) for men and women, calculate the difference, and run a *t*-test to determine if this difference is statistically significant (at the 5% level). Note there are some missing values (NAs). To avoid getting NA for your means, add , na.rm=TRUE inside mean(). This tells R to *remove* (rm) the missing (NA) values.

b. Run a regression of wage96 on male, and write down the estimated regression equation. Use the regression coefficients to find:

- the average wage for males
- the average wage for females
- the difference between the average for males and females

c. Now recode (with `ifelse()`) the sex of a person by generating a variable $female_i$ instead of $male_i$. Rerun the regression in part (b) using `female` instead of `male`. Write down the estimated regression equation, and use the regression coefficients to find:

- the average wage for males
- the average wage for females
- the difference between the average for males and females

---

d. Education `hgc96` probably has a lot to do with wages. Run a regression of wages on female and education. Write down the estimated regression equation, and interpret each coefficient (note there are no interaction effects here). What happened to the estimate on Female?

---

e. Does the effect of education on wages differ between men and women? Run a regression of `wage96` on `female`, `hgc96`, and an interaction term. Write down the estimated regression equation.

---

f. What we actually have are two different regression lines. Visualize this with a scatterplot between `wage96` (Y) and `hgc96` (X). Note there are several outliers about $300, you may wish to drop them to get a better view.

g. Do the two regression lines have the same intercept? The same slope? Use the original regression in part (e) to test these possibilities.

h. Take your regression equation from part (e) and rewrite it as two separate regressions. Interpret the coefficients for each.

i. Double check your calculations in (h) are correct by running the regression in (e) twice, once for only males and once for only females. Hint: subset your data conditionally before each regression. Note `R` may give you errors with the regression, but will still print the relevant coefficients.

i. Now let's examine the effect of regions on earnings. The four possible regions coded into the data are `norest96`, `norcen96`, `south96`, and `west96`. Run a regression of `wage96` on `norest96`, `norcen96`, `south96`, and `west96`. What happens, and why?

j. Rerun the regression of wages on regional dummies, only this time omit `norest96`. Write down the estimated regression equation and interpret each coefficient. Then use the regression coefficients to find the average wage in the

- Northeast
- North Central
- South
- West

k. Rerun the regression of wages on regional dummies, only this time omit `west96`. Write down the estimated regression equation and interpret each coefficient. Then use the regression coefficients to find the average wage in the

- Northeast
- North Central
- South
- West

l. Use `stargazer` to make a nice output table of all of your regressions from parts b,d,e,j,k.

8. Lead is toxic, particularly for young children, and for this reason government regulations severely restrict the amount of lead in our environment. In the early part of the 20^{th} century, the underground water pipes in many U.S. cities contained lead, and lead from these pipes leached into drinking water. This exercise will have you investigate the effect of these lead pipes on infant mortality. Download the `LeadMortality.dta` dataset from Blackboard. This dataset contains data on:

- `infrate`: infant mortality rate (deaths per 100 in population)
- `lead`: = 1 if city has lead water pipes, = 0 if did not have lead pipes
- `pH`: water pH

and several demographic variables for 172 U.S. cities in 1900.

We want to figure out what is the effect of height on wages (e.g. do taller people earn more on average than shorter people?)

a. Using `R` to examine the data, find the average infant mortality rate for cities with lead pipes and for cities without lead pipes. Calculate the difference, and run a t-test to determine if this difference is statistically significant (at the 5% level).

b. Run a regression of `infrate` on `lead`, and write down the estimated regression equation. Use the regression coefficients to find:

- the average infant mortality rate for cities with lead pipes
- the average infant mortality rate for cities without lead pipes
- the difference between the averages for cities with or without lead pipes

---

---

c. We see again that lead by itself appears to not be significant. Perhaps the estimate on lead is biased. Does the pH of the water matter? Find some statistical evidence for including pH.

---

d. Include `ph` in your regression from part (b). Write down the estimated regression equation, and interpret each coefficient (note there is no interaction effect here). What happens to the estimate on `lead`?

---

e. The amount of lead leached from lead pipes normally depends on the chemistry of the water running through the pipes: the more acidic the water (lower pH), the more lead is leached. Create an interaction term between lead and pH, and run a regression of `infrate` on `lead`, `pH`, and your interaction term. Write down the estimated regression equation. Is this interaction significant?

f. What we actually have are two different regression lines. Visualize this with a scatterplot between `infrate` (Y) and `ph` (X) by `lead`, in a way similar to the last problem (though there is no need to jitter and there are no pesky outliers).

g. Do the two regression lines have the same intercept? The same slope? Use the original regression in part (e) to test these possibilities.

h. Take your regression equation from part (e) and rewrite it as two separate regressions. Interpret the coefficients for each.

h. Double check your calculations in (g) are correct by running the regression in (e) twice, once for cities without lead pipes and once for cities with lead pipes.

i. Use `stargazer` to make a nice output table of all of your regressions from parts b,d,e.