Problem Set 5

ECON 480 - Fall 2022

Due by Monday, November 28, 2022

Theory and Concepts

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.

---

3. In your own words, describe when and why using logged variables can be useful.

4. In your own words, describe when we would use an F-test, and give some example (null) hypotheses. Describe intuitively and specifically (no need for the formula) what exactly F is trying to test for.

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."

5. 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	Avg. 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. 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:

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

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

Variable	Description
$Price_h$	price of house h (in thousands of \$)
$Bedrooms_h$	number of bedrooms in house h
$Baths_h$	number of bathrooms in house h
$Pool_h$	$\begin{cases} = 1 & \text{if house } h \text{ has a pool} \\ = 0 & \text{if house } h \text{ does not have a pool} \end{cases}$
$View_h$	$\begin{cases} = 1 & \text{if house } h \text{ has a nice view} \\ = 0 & \text{if house } h \text{ does not have a nice view} \end{cases}$

a. Suppose she runs the following regression:

$$\widehat{\text{Price}}_h = 119.20 + 29.76\,\text{Bedrooms}_h + 24.09\,\text{View}_h + 14.06\,(\text{Bedrooms}_h \times \text{View}_h)$$

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 she runs the following regression:

$$\widehat{\text{Price}}_h = 189.20 + 42.40\,\text{Pool}_h + 12.10\,\text{View}_h + 12.09\,(\text{Pool}_h \times \text{View}_h)$$

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 she runs the following regression:

$$\widehat{\text{Price}}_h = 87.90 + 53.94\,\text{Bedrooms}_h + 15.29\,\text{Baths}_h + 16.19\,(\text{Bedrooms}_h \times \text{Baths}_h)$$

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?

---

7. Suppose we want to examine the change in average global temperature over time. We have data on the deviation in temperature from pre-industrial times (in Celcius), and the year.

a. Suppose we estimate the following simple model relating deviation in temperature to year:

$$\widehat{\text{Temperature}}_i = -10.46 + 0.006\text{Year}_i$$

Interpret the coefficient on Year (i.e. $\hat{\beta}_1$)

b. Predict the (deviation in) temperature for the year 1900 and for the year 2000.

c. Suppose we believe temperature deviations are increasing at an increasing rate, and introduce a quadratic term and estimate the following regression model:

$$\widehat{\text{Temperature}}_i = 155.68 - 0.116\text{Year}_i + 0.000044\text{Year}_i^2$$

What is the marginal effect on (deviation in) global temperature of one additional year elapsing?

d. Predict the marginal effect on temperature of one more year elapsing starting in 1900, and in 2000.

e. Our quadratic function is a U-shape. According to the model, at what year was temperature (deviation) at its minimum?

---

8. Suppose we want to examine the effect of cell phone use while driving on traffic fatalities. While we cannot measure the amount of cell phone activity while driving, we do have a good proxy variable, the number of cell phone subscriptions (in 1000s) in a state, along with traffic fatalities in that state.

a. Suppose we estimate the following simple regression:

$$\widehat{\text{fatalities}}_i = 123.98 + 0.091\text{cell plans}_i$$

Interpret the coefficient on cell plans (i.e. $\hat{\beta}_1$)

b. Now suppose we estimate the regression using a linear-log model:

$$\widehat{\text{fatalities}}_i = -3557.08 + 515.81\ln(\text{cell plans}_i)$$

Interpret the coefficient on ln(cell plans) (i.e. $\hat{\beta}_1$)

c. Now suppose we estimate the regression using a log-linear model:

$$\ln(\widehat{\text{fatalities}}_i) = 5.43 + 0.0001\text{cell plans}_i$$

Interpret the coefficient on cell plans (i.e. $\hat{\beta}_1$)

d. Now suppose we estimate the regression using a log-log model:

$$\ln(\widehat{\text{fatalities}}_i) = -0.89 + 0.85\ln(\text{cell plans}_i)$$

Interpret the coefficient on cell plans (i.e. $\hat{\beta}_1$)

e. Suppose we include several other variables into our regression and want to determine which variable(s) have the largest effects, a State's cell plans, population, or amount of miles driven. Suppose we decide to *standardize* the data to compare units, and we get:

$$\widehat{\text{fatalities}}_Z = 4.35 + 0.002\text{cell plans}_Z - 0.00007\text{population}_Z + 0.019\text{miles driven}_Z$$

Interpret the coefficients on cell plans, population, and miles driven. Which has the largest effect on fatalities?

f. Suppose we wanted to make the claim that it is *only* miles driven, and neither population nor cell phones determine traffic fatalities. Write (i) the null hypothesis for this claim and (ii) the estimated restricted regression equation.

g. Suppose the R^2 on the original regression from (e) was 0.9221, and the R^2 from the restricted regression is 0.9062. With 50 observations, calculate the F-statistic.

R Questions

Answer the following questions using `R`. When necessary, please write answers in the same document (knitted `Rmd` to `html` or `pdf`, typed `.doc(x)`, or handwritten) as your answers to the above questions. Be sure to include (email or print an `.R` file, or show in your knitted `markdown`) your code and the outputs of your code with the rest of your answers.

9. 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. This dataset contains data on:

Variable	Description
`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.

a. Using `R` to examine the data, find the average infant mortality rate for cities *with* lead pipes and for cities *without* lead pipes. Then, calculate the difference in mortality rates, and run a t-test to determine if this difference is statistically significant.
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. Does the pH of the water matter? 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`?
d. 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 effect significant?
e. What we actually have are two different regression lines. Visualize this with a scatterplot between `infrate` (Y) and `ph` (X) by `lead`.
f. Do the two regression lines have the same intercept? The same slope? Use the original regression in part D to test these possibilities.
g. Take your regression equation from part D and rewrite it as two separate regression equations (one for no lead and one for lead). Interpret the coefficients for each.
h. Double check your calculations in G are correct by running the regression in D twice, once for cities without lead pipes and once for cities with lead pipes.**[1]
i. Use `huxtable` to make a nice output table of all of your regressions from parts B, C, and D.

---

[1] `filter()` the data first, then use the filtered data for the `data=` in each regression.

10. Let's look at economic freedom and GDP per capita using some data I sourced from Gapminder[2], Freedom House[3] and Fraser Institute Data[4] and cleaned up for you, with the following variables:

Variable	Description
`Country`	Name of country
`ISO`	Code of country (good for plotting)
`econ_freedom`	Economic Freedom Index score (2016) from 1 (least) to 10 (most free)
`pol_freedom`	Political freedom index score (2018) from 1 (least) top 10 (most free)
`gdp_pc`	GDP per capita (2018 USD)
`continent`	Continent of country

a. Does economic freedom affect GDP per capita? Create a scatterplot of `gdp_pc` (`Y`) against `econ_freedom` (`x`). Does the effect appear to be linear or nonlinear?
b. Run a simple regression of `gdp_pc` on `econ_freedom`. Write out the estimated regression equation. What is the marginal effect of `econ_freedom` on `gdp_pc`?
c. Let's try a quadratic model. Run a quadratic regression of `gdp_pc` on `econ_freedom`. Write out the estimated regression equation.
d. Add the quadratic regression to your scatterplot.
e. What is the marginal effect of `econ_freedom` on `gdp_pc`?
f. As a quadratic model, this relationship should predict an`econ_freedom` score where `gdp_pc` is at a *minimum*. What is that minimum Economic Freedom score, and what is the associated GDP per capita?
g. Run a cubic model to see if we should keep going up in polynomials. Write out the estimated regression equation. Should we add a cubic term?
h. Another way we can *test* for non-linearity is to run an F-test on all non-linear variables - i.e. the quadratic term and the cubic term ($\hat{\beta}_2$ and $\hat{\beta}_3$) and test against the null hypothesis that:

$$H_0 : \hat{\beta}_2 = \hat{\beta}_3 = 0$$

Run this joint hypothesis test, and what can you conclude?
i. Instead of a polynomial model, try out a logarithmic model. It is hard to interpret percent changes on an index, but it is easy to understand percent changes in GDP per capita, so run a *log-linear* regression. Write out the estimated regression equation. What is the marginal effect of `econ_freedom`?
j. Make a scatterplot of your log-linear model with a regression line.
k. Put all of your results together in a regression output table with `huxtable` from your answers in questions B, C, G, and H.

---

[2]GDP per capita (2018)
[3]Political freedom score (2018)
[4]Economic Freedom score (2016)