Econometrics HW #5

Ryan Safner

Due: Wednesday, December 6, 2018

Theory & Concepts

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

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

---

2. 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."

3. 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?

4. 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:

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

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` 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. Let's reexamine the `speeding_tickets` dataset, now that we have some more models to try out.

a. Run a regression of `Amount` on `Age`. Write out the estimated regression equation, and interpret the coefficient on Age.

b. Is the effect of `Age` on `Amount` nonlinear? Run a quadratic regression. Write out the estimated regression equation. Is this model an improvement?

c. Write an equation for the marginal effect of `Age` on `Amount`.

d. Predict the marginal effect on `Amount` of being one year older when you are 18. How about when you are 40?

e. Our quadratic function is a U-shape. According to the model, at what age is the amount of the fine minimized?

f. Create a scatterplot between Amount and Age and overlay it with your predicted quadratic regression curve. The regression curve, just like any regression *line*, is a `geom_smooth()` layer on top of the `geom_point()` layer. We will need to customize `geom_smooth()` to `geom_smooth(method="lm", formula="y~poly(x,2)`. This is the same as a regression line (`method="lm"`), but we are modifying the formula to a polynomial of degree 2 (quadratic): $y = a + bx + cx^2$.

g. It's quite hard to see the quadratic curve with all those data points. Redo another plot and this time, only keep the `geom_smooth()` and leave out `geom_point()`. This will only plot the regression curve.

h. Should we use a higher-order polynomial equation? Run a cubic model, and determine whether it is necessary.

i. Run an F-test to check if a nonlinear model is appropriate. Your null hypothesis is $H__0: $\beta_2 + \beta_3 = 0$ from the regression in pert (h). The command is `linearHypothesis(reg_name, c("var1", "var2"))` where `reg_name` is the name of the `lm` object you saved your regression in, and `var1` and `var2` (or more) in quotes are the names of the variables you are testing. This function requires (installing and) loading the "`car`" package (additional regression tools).

j. Now let's take a look at speed (`MPHover` the speed limit). Running a simple regression between `Amount` and `MPHover` each time, run three regressions:

- a linear-log model
- a log-linear model
- a log-log model Write down each estimated regression equation and interpret the coefficient on the `MPHover` variable.

l. Which log model from the previous part has the best fit?

m. Return to the quadratic model. Run a quadratic regression of `Amount` on `Age`, `Age`^{2}, `MPHover`, and the race dummy variables. Test the null hypothesis: "the race of the driver does not matter at all."