Problem Set 2

BUAN 6356

Due: Tuesday, 2017-09-26-11:59pm

Deliverable: an R source-code file named ps2.r

Question 1

Data The data attend.csv contains 680 observations on students in a course on microeconomic principles.

Analysis

- Read the data attend.csv into a new variable 'context1' and familiarize yourself with the data.
- Create a new variable 'attendrt' for the attendance rate (classes attended out of the total number of classes). You can find information about the total number of classes in the labels .txt file.
- Create a new variable 'hwrt' for the homework completion rate (homeworks turned in of out the total number of homework assignments). You can find information about the total number of homework assignments in the labels .txt file.
- Run the following linear model using the 'lm' function. Store the result in: model1

$$\text{termGPA}_i = \beta_0 + \beta_1 \text{priGPA}_i + \beta_2 \text{ACT}_i + \beta_3 \text{attendrt}_i + \beta_4 \text{hwrt}_i + e_i \quad (1)$$

Interpretations

a. Interpret the estimated coefficient on attendrt from model1 (eq 1).

b. Interpret the estimated coefficient on hwrt from model1 (eq 1).

c. Predict the termGPA for a student with a 32 ACT and a 2.2 priGPA who attended 28 lectures and turned-in 8 homework assignments.

d. Predict the termGPA for a student with a similar attendence and homework pattern who had a 20 ACT and a 3.9 priGPA.

e. Intuitively, which variable is more important to the termGPA, ACT or priGPA?

f. Predict the termGPA for a student with a 25 ACT and a 3.0 priGPA who attends all the classes, but only finishes half the homework assignments.

g. Predict the termGPA for a similarly qualified student who turns in all the homwork assignments, but only attends half the classes.

h. Intuitively, which variable is more important to the termGPA, attendance or homework completion?

i. Why is it easier to compare attendrt and hwrt than it is to compare priGPA and ACT score?

Question 2

Data The data set in CEOSAL2.csv contains information on chief executive officers for U.S. corporations. The variable salary is annual compensation, in thousands of dollars.

Analysis

- Read the data CEOSAL2.csv into a new variable 'context2' and familiarize yourself with the data.
- Run the following linear model using the 'lm' function. Store the result in: model2 (Remember, you can use the natural log of a variable in the model just by adding 'log( )' around the variable)

$$\ln[\text{salary}_i] = \beta_0 + \beta_1 \ln[\text{mktval}_i] + \beta_2 \text{profits}_i + \beta_3 \text{ceoten}_i + e_i \tag{2}$$

- Run the following linear model using the 'lm' function. Store the result in: model3

$$\ln[\text{salary}_i] = \beta_0 + \beta_1 \ln[\text{mktval}_i] + \beta_2 \text{profits}_i + \beta_3 \text{ceoten}_i + \beta_4 \ln[\text{sales}_i] + e_i \tag{3}$$

Interpretations

j. We used natural logs on all the dollar-valued quantities except profits in models 2 & 3 (eq. 2 & 3). Why did we not take the log of profits?

k. Interpret the estimated coefficient on log mktval in model 2 (eq 2).

l. Interpret the estimated coefficient on log mktval in model 3 (eq 3).

m. Compare the test statistics on log mktval between model 2 and model 3. Please explain the differences you find in terms of the biases we discussed in class.

n. Is the coefficient on profits significant in model 3 (eq 3)?

o. Interpret the estimated coefficient on log sales in model 3 (eq 3).

Question 3

Data hprice1.csv contains data on 88 U.S. houses, their characteristics, and their prices at the time of sale.

Analysis

- Read the data hprice1.csv into a new variable 'context3' and familiarize yourself with the data.
- Run the following linear model using the 'lm' function. Store the result in: model4

$$\text{price}_i = \beta_0 + \beta_1 \text{bdrms}_i + \beta_2 \ln[\text{lotsize}_i] + \beta_3 \ln[\text{sqrft}_i] + \beta_4 \text{colonial}_i + e_i \tag{4}$$

- Run the following linear model using the 'lm' function. Store the result in: model5

$$\ln[\text{price}_i] = \beta_0 + \beta_1 \text{bdrms}_i + \beta_2 \ln[\text{lotsize}_i] + \beta_3 \ln[\text{sqrft}_i] + \beta_4 \text{colonial}_i + e_i \tag{5}$$

Interpretations

p. Interpret the estimated coefficient on log lotsize from model4 (eq 4).

q. Interpret the estimated coefficient on log lotsize from model5 (eq 5).

r. Interpret the estimated coefficient on colonial from model4 (eq 4, please ignore significance here).

s. Which model (4 or 5) better fits the data for this data set? On what criterion/criteria are you basing your judgement?

t. Suppose your house is worth \$300k. You are considering an expansion of your home to add a master suite (+1 bedroom to your home). This expansion would increase your square-footage by 10% and would cost \$50k. You have valued your enjoyment of the additional space at \$20k, so you would only be willing to consider the build if it were to also increase your property value accordingly. Does the appropriate model indicate that you should pursue the expansion?

Question 4

Data The data in JTRAIN2.csv come from a job training experiment conducted for low income men in the U.S. during 1976-1977; see Lalonde (1986).

Analysis

- Read the data JTRAIN2.csv into a new variable 'context4' and familiarize yourself with the data.
- Run the following linear model using the 'lm' function. Store the result in: model6

$$\text{re78}_i = \beta_0 + \beta_1 \text{re75}_i + \beta_2 \text{train}_i + \beta_3 \text{educ}_i + \beta_4 \text{black}_i + e_i \tag{6}$$

Interpretations

u. Interpret the estimated coefficient on re75 from model6 (eq 6).

v. Interpret the estimated coefficient on train from model6 (eq 6). Is the coefficient significant?

w. Interpret the estimated coefficient on black from model6 (eq 6).