Homework 5

Multiple Regression Model Estimation - part 1 (30 points)

Instruction:

- This HW must be done in Rmarkdown!
- Please submit both the .rmd and the Microsoft word files. (Do not submit a PDF or any other image files as the TAs are going to give you feedback in your word document)
- Name your files as: HW5-groupnumber-name
- All the HW assignments are individual work. However, I highly encourage you to discuss it with your group members.
- Late homework assignments will not be accepted under any circumstances.

Problems

Question 1 The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u$$

where sleep and totwrk (total work) are measured in minutes per week and educ and age are measured in years.

(i) If adults trade off sleep for work, what is the sign of β_1?

(ii) What signs do you think β_2 and β_3 will have?

(iii) Using the data in SLEEP75, the estimated equation is

$$\widehat{sleep} = 3,638.25 - .148 totwrk - 11.13 educ + 2.20 age$$
$$n = 706 | R^2 = .113$$

If someone works five more hours per week, by how many minutes is sleep predicted to fall? Is this a large tradeoff?

(iv) Discuss the sign and magnitude of the estimated coefficient on educ.

(v) Would you say totwrk, educ, and age explain much of the variation in sleep? What other factors might affect the time spent sleeping? Are these likely to be correlated with totwrk?

Question 2 The median starting salary for new law school graduates is determined by

$$log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 log(libvol) + \beta_4 log(cost) + \beta_5 rank + u,$$

where LSAT is the median LSAT score for the graduating class, GPA is the median college GPA for the class, libvol is the number of volumes in the law school library, cost is the annual cost of attending law school, and rank is a law school ranking (with rank = 1 being the best).

(i) Explain why we expect $\beta_5 \leq 0$

(ii) What signs do you expect for the other slope parameters? Justify your answers.

(iii) Using the data in LAWSCH85, the estimated equation is

$$\widehat{log(salary)} = 8.34 + .0047 LAST + .248 GPA + .095 log(libvol) + .038 log(cost) - .0033 rank$$

$$n = 136, R^2 = .842.$$

What is the predicted ceteris paribus difference in salary for schools with a median GPA different by one point? (Report your answer as a percentage.)

(iv) Interpret the coefficient on the variable log(libvol).

(v) Would you say it is better to attend a higher ranked law school? How much is a difference in ranking of 20 worth in terms of predicted starting salary?

Question 3 The following equations were estimated using the data in LAWSCH85:

$$\widehat{lsalary} = \underset{(.24)}{9.90} - \underset{(.0003)}{.0041}rank + \underset{(.069)}{.294}GPA$$
$$n = 142 | R^2 = .8238$$

$$\widehat{lsalary} = \underset{(.29)}{9.86} - \underset{(.0004)}{.0038}rank + \underset{(.083)}{.295}GPA + \underset{(.00036)}{.00017}age$$
$$n = 99 | R^2 = .8036$$

How can it be that the R-squared is smaller when the variable age is added to the equation? (Hint: Look at the observation numbers in each model)

Computer Exercises

Question 4 Use the data in HPRICE1 to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

where price is the house price measured in thousands of dollars.

(i) Write out the results in equation form.

(ii) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

(iii) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).

(iv) What percentage of the variation in price is explained by square footage and number of bedrooms?

(v) The first house in the sample has $sqrft = 2,438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

(vi) The actual selling price of the first house in the sample was \$300,000 (so $price = 300$). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

Question 5 Use the data in DISCRIM to answer this question. These are zip code–level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks.

(i) Find the average values of prpblck and income in the sample, along with their standard deviations. What are the units of measurement of prpblck and income?

(ii) Consider a model to explain the price of soda, psoda, in terms of the proportion of the population that is black and median income:

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + u.$$

Estimate this model by OLS and report the results in equation form, including the sample size and R-squared. (Do not use scientific notation when reporting the estimates.) Interpret the coefficient on prpblck. Do you think it is economically large?

(iii) Compare the estimate from part (ii) with the simple regression estimate from psoda on prpblck. Is the discrimination effect larger or smaller when you control for income?

(iv) A model with a constant price elasticity with respect to income may be more appropriate. Report estimates of the model

$$log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 log(income) + u.$$

If prpblck increases by .20 (20 percentage points), what is the estimated percentage change in psoda? (Hint: The answer is 2.xx, where you fill in the "xx.")

(v) Now add the variable prppov to the regression in part (iv). What happens to $\widehat{\beta_{prpblck}}$?

(vi) Find the correlation between log(income) and prppov. Is it roughly what you expected?

(vii) Evaluate the following statement: "Because log(income) and prppov are so highly correlated, they have no business being in the same regression."