Homework 4

Simple Regression Model (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: HW4-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 data used in this problem comes from a small sample of older, in-hospital patients with information on average daily step count (measured over 5 days with a pedometer) and their length of stay (los) in the hospital.

$$\widehat{log(los)} = 3.01997 - 0.17800\ log(avgSteps)$$
$$n = 148,\ R^2 = 0.09109$$

(i) Interpret the coefficient on log(avgSteps). Is the sign of this estimate what you expect it to be?

(ii) What other factors may effect length of stay? Might these be correlated with average steps?

(iii) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of los with respect to avgSteps? Hint: which of the SLR assumptions is/are violated?)

Question 2 Consider the savings function:

$$sav = \beta_0 + \beta_1 inc + u, \qquad u = inc.e$$

where e is a random variable with $E(e) = 0$ and $var(e) = \sigma_e^2$. Assume that e is independent of inc.

(i) Show that $E(u|inc) = 0$, so that the key zero condition mean assumption is satisfied.

(ii) Show that $Var(u|inc) = \sigma_e^2 inc$, so that the homoskedasticity Assumption SLR.5 is violated. In particular, the variance of sav increases with inc. [Hint: $Var(e|inc) = Var(e)$ if e and inc are independent].

(iii) Provide a discussion that supports the assumption that the variance of savings increases with family income.

Computer Exercises

Question 3 These data were used in the doctoral dissertation of Jeffrey Blend, Department of Agricultural Economics, Michigan State University, 1998. The thesis was supervised by Professor Eileen van Ravensway. Drs. Blend and van Ravensway kindly provided the data, which were obtained from a telephone survey conducted by the Institute for Public Policy and Social Research at MSU. Data loads lazily. The variable faminc is family income measured in thousands of dollars and the variable hhsize is the household size

(i) Find the average family income and the average house hold size in the sample.

(ii) Now, estimate the simple regression equation

$$\widehat{faminc} = \hat{\beta}_0 + \hat{\beta}_1 hhsize$$

and report the results along with the sample size and R-squared

(iii) Interpret the intercept in your equation. Interpret the coefficient on *hhsize*.

(iv) Find the predicted *faminc* when $hhsize = 11$. Is this a reasonable prediction? Explain what is happening here.

(v) How much of the variation in $faminc$ is explained by *hhsize*? Is this a lot in your opinion?

Question 4 Use the data in GPA1 for this question. We want to explore the relationship between the college GPA(*colGPA*) and classes skipped. (*skipped*)

(i) Do you think each additional class skipped has the same effect on the pass rate, or does an exponential effect seem more appropriate? Explain.

(ii) In the population model

$$colGPA = \beta_0 + \beta_1 log(skipped) + u$$

Argue that $\frac{\beta_1}{10}$ is the percentage point change in *colGPA* given a 10% increase in *skipped*

(iii) Use the data in GPA1to estimate the model from part (ii). Report the estimated equation in the usual way, including the sample size and R-squared

(iv) How big is the estimated skipped effect? Namely, if skipping class increases by 10%, what is the estimated point decrease in *colGPA*.

(v) One might worry that regression analysis can produce fitted values for *colGPA* that are less than 0.0. Why is this not much of a worry in this data set?

Question 5 Use the data in VOTE1 From M. Barone and G. Ujifusa, The Almanac of American Politics, 1992. Washington, DC: National Journal to answer the following questions:

(i) What is the average campaign expenditure by democrats (expendA)? What percentage of candidates were democrats?

(ii) What is the average proportion of voters that voted for candidate A each district? What are the minimum and maximum values?

(iii) Estimate the model

$$voteA = \beta_0 + \beta_1 expendA + u$$

by OLS and report the results in the usual way, including the sample size and R-squared.