Homework 6

Multiple Regression Model Estimation - part 2 (25 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: HW6-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 Consider the multiple regression model containing three independent variables, under Assumptions MLR.1 through MLR.4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

You are interested in estimating the sum of the parameters on $x_1$and x_2; call this $\theta_1 = \beta_1 + \beta_2$.

(i) Show that $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ is an unbiased estimator of θ_1.

(ii) Find $var(\hat{\theta}_1)$ in terms of $var(\hat{\beta}_1)$, $var(\hat{\beta}_2)$, and $corr(\hat{\beta}_1, \hat{\beta}_2)$

Question 2 Suppose that average worker productivity at manufacturing firms (avgprod) depends on two factors, average hours of training (avgtrain) and average worker ability (avgabil):

$$avgprod = \beta_0 + \beta_1 avgtrain + \beta_2 avgabil + u$$

Assume that this equation satisfies the Gauss-Markov assumptions. If grants have been given to firms whose workers have less than average ability, so that avgtrain and avgabil are negatively correlated, what is the likely bias in $\tilde{\beta}_1$ obtained from the simple regression of avgprod on avgtrain?

Question 3 The following estimated equations use the data in MLB1, which contains information on major league baseball salaries. The dependent variable, lsalary, is the log of salary. The two explanatory variables are years in the major leagues (years) and runs batted in per year (rbisyr):

$$\widehat{lsalary} = \underset{(.098)}{12.373} + \underset{(.0132)}{.1770} years$$

$$n = 353 \quad SSR = 326.196 \quad SER = .964 \quad R^2 = .337$$

$$\widehat{lsalary} = \underset{(.084)}{11.861} + \underset{(.0118)}{.0904} years + \underset{(.0020)}{.0302} rbisyr$$

$$n = 353 \quad SSR = 198.475 \quad SER = .753 \quad R^2 = .597$$

(i) How many degrees of freedom are in each regression? How come the SER is smaller in the second regression than the first?

(ii) The sample correlation coefficient between years and rbisyr is about 0.487. Does this make sense? What is the variance inflation factor (there is only one) for the slope coefficients in the multiple regression? Would you say there is little, moderate, or strong collinearity between *years* and *rbisyr*?

(iii) How come the standard error for the coefficient on $years$ in the multiple regression is lower than its counterpart in the simple regression?

Question 4 In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

(i) In the model

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u,$$

does it make sense to hold sleep, work, and leisure fixed, while changing study?

(ii) Explain why this model violates Assumption MLR.3.

(iii) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption MLR.3?

Computer Exercises

Question 5 Use the data in MEAPSINGLE to study the effects of single-parent households on student math performance. These data are for a subset of schools in southeast Michigan for the year 2000. The socioeconomic variables are obtained at the ZIP code level (where ZIP code is assigned to schools based on their mailing addresses).

(i) Run the simple regression of math4 on pctsgle and report the results in the usual format. Interpret the slope coefficient. Does the effect of single parenthood seem large or small?

(ii) Add the variables lmedinc and free to the equation. What happens to the coefficient on pctsgle? Explain what is happening.

(iii) Find the sample correlation between lmedinc and free. Does it have the sign you expect?

(iv) Does the substantial correlation between lmedinc and free mean that you should drop one from the regression to better estimate the causal effect of single parenthood on student performance? Explain.

(v) Find the variance inflation factors (VIFs) for each of the explanatory variables appearing in the regression in part (ii). Which variable has the largest VIF? Does this knowledge affect the model you would use to study the causal effect of single parenthood on math performance?

Question 6 Use the data in HTV to answer this question. The data set includes information on wages, education, parents' education, and several other variables for 1,230 working men in 1991.

(i) What is the range of the educ variable in the sample? What percentage of men completed twelfth grade but no higher grade? Do the men or their parents have, on average, higher levels of education?

(ii) Estimate the regression model $educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + u$ by OLS and report the results in the usual form. How much sample variation in educ is explained by parents' education? Interpret the coefficient on motheduc.

(iii) Add the variable abil (a measure of cognitive ability) to the regression from part (ii), and report the results in equation form. Does "ability" help to explain variations in education, even after controlling for parents' education? Explain.

(iv) Now estimate an equation where abil appears in quadratic form:

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u.$$

Using the estimates $\hat{\beta}_3$ and $\hat{\beta}_4$, use calculus to find the value of abil, call it $abil^*$, where educ is minimized. (The other coefficients and values of parents' education

variables have no effect; we are holding parents' education fixed.) Notice that abil is measured so that negative values are permissible. You might also verify that the second derivative is positive so that you do indeed have a minimum.

(v) Argue that only a small fraction of men in the sample have "ability" less than the value calculated in part (iv). Why is this important?

(vi) This part is optional! (1 point extra credit)
If you have access to a statistical program that includes graphing capabilities, use the estimates in part (iv) to graph the relationship between the predicted education and abil. Set motheduc and fatheduc at their average values in the sample, 12.18 and 12.45, respectively.