Homework 7

Multiple Regression Model - Inference (30 points)

Due Date: **Wednesday** March 22nd at 11:59 pm

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: HW7-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 Which of the following can cause the usual OLS t statistics to be invalid (that is, not to have t distributions under H_0)?

(i) Heteroskedasticity

(ii) A sample correlation coefficient of .95 between two independent variables that are in the model.

(iii) Omitting an important explanatory variable.

Hint: You need to check CLM assumptions.

Question 2 Are rent rates influenced by the student population in a college town? Let rent be the average monthly rent paid on rental units in a college town in the United States. Let pop denote the total city population, avginc the average city income, and pctstu the student population as a percentage of the total population. One model to test for a relationship is

$$log(rent) = \beta_0 + \beta_1 log(pop) + \beta_2 log(avginc) + \beta_3 pctstu + u$$

(i) State the null hypothesis that size of the student body relative to the population has no ceteris paribus effect on monthly rents. State the alternative that there is an effect.

(ii) What signs do you expect for β_1 and β_2?

(iii) The equation estimated using 1990 data from RENTAL data set for 64 college towns is

$$\widehat{log(rent)} = \underset{(.844)}{.043} + \underset{(.039)}{.066} log(pop) + \underset{(.081)}{.507} log(avginc) + \underset{(.0017)}{.0056} pctstu$$

$$n = 64 \quad R^2 = .458.$$

What is wrong with the statement: "A 10% increase in population is associated with about a 6.6% increase in rent"?

(iv) Test the hypothesis stated in part (i) at the 1% level.

Question 3 Consider the multiple regression model with three independent variables, under the classical linear model assumptions MLR.1 through MLR.6:

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

You would like to test the null hypothesis $H_0 : \beta_1 - 3\beta_2 = 1$

(i) Let $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the OLS estimators of β_1 and β_2. Find Var $(\hat{\beta}_1 - 3\hat{\beta}_2)$ in terms of the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ and the covariance between them. What is the standard error of $\hat{\beta}_1 - 3\hat{\beta}_2$?

(ii) Write the t statistic for testing $H_0 : \beta_1 - 3\beta_2 = 1$

(iii) Define $\theta_1 = \beta_1 - 3\beta_2$ and $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2$. Write a regression equation involving $\beta_0, \theta_1, \beta_2, \beta_3$ that allows you to directly obtain $\hat{\theta}_1$ and its standard error.

Question 4 In this question we are going to use an example testing the rationality of assessment of housing prices using a level-level formulation.

(i) In the simple regression model below, the assessment is rational if $\beta_1 = 1$ and $\beta_0 = 0$ i.e. if markets are efficient, everything should be reflected in "assess" and we don't need to control for any other explanatory variables out there.

$$price = \beta_0 + \beta_1 assess + u$$

Using the estimated equation given

$$\widehat{price} = \underset{(16.27)}{-14.47} + \underset{(.049)}{.976} assess$$

$$n = 88 \quad SSR = 165,644.51 \quad R^2 = .820$$

test the hypothesis that $H_0 : \beta_0 = 0$ against the two-sided alternative. Then, test $H_0 : \beta_1 = 1$ against the two-sided alternative. What do you conclude?

(ii) To test the joint hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$, we need the sum of squared residuals (SSR) in the restricted model. This amounts to computing $\sum_{i=1}^{n}(price_i - assess_i)^2$, where n =88, since the residuals in the restricted model are just $price_i - assess_i$ (No estimation is needed for the restricted model because both parameters are specified under H_0.). This turns out to yield SSR = 209,448.99. Carry out the F test for the joint hypothesis.

(iii) Now, test the joint hypothesis: $H_0 : \beta_2 = 0, \beta_3 = 0, and\ \beta_4 = 0$ in the model $price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u$. The R-squared from estimating this model using the same 88 houses is .829.

(iv) If the variance of price changes with assess, lotsize, sqrft or bdrms, what can you say about the F test from part (iii)? (Hint: Can you rely on F statistics in this case. Is there any violations of OLS assumption happening here? No calculation required!)

Computer Exercises

Question 5 The data set 401KSUBS contains information on net financial wealth ($nettfa$), age of the survey respondent (age), annual family income (inc), family size ($fsize$), and participation in certain pension plans for people in the United States. The wealth and income variables are both recorded in thousands of dollars. For this question, use only the data for single-person households (so $fsize = 1$).

(i) How many single person households are there in the dataset?

(ii) Use OLS to estimate the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + u$$

and report the results using the usual format. Be sure to use only the single-person households in the sample. Interpret the slope coefficients. Are there any surprises in the slope estimates?

(iii) Does the intercept from the regression in part (ii) have an interesting meaning? Explain.

(iv) Find the p-value for the test $H_0 : \beta_2 = 1$ against $H_1 : \beta_2 < 1$. Do you reject H_0 at the 1% significance level?

(v) If you do a simple regression of nettfa on inc, is the estimated coefficient on inc much different from the estimate in part (ii)? Why or why not?

Question 6 Use the data in DISCRIM to answer this question.

(i) Use OLS to estimate the model

$$log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 log(income) + \beta_3 prppov + u,$$

and report the results in the usual form. Is $\hat{\beta}_1$ statistically different from zero at the 5% level against a two-sided alternative? What about at the 1% level?

(ii) What is the correlation between log(income) and prppov? Is each variable statistically significant in any case? Report the two-sided p-values.

(iii) Are the *prpblck* and *prppov* jointly significant at 5% level?

(iv) To the regression in part (i), add the variable **log(hseval)**. Interpret its coefficient and report the two-sided p-value for $H_0 : \beta_{log(hseval)} = 0$.

(iv) In the regression equation from part (iv), test the hypothesis that the effect of *prpblck* on *psoda* is the same as the effect of *prppov* on *psoda*?