Assignment III

Due at 5:30pm of April 24 (Monday)

This assignment is related to Chapter 6 (problems 1, 2 and 3), 7 (problems 4 and 5) and 8 (problems 6, 7 and 8). You can use any formula taught in class without proof. If you want to use any result from other courses, please provide proof or justification.

Note: (i) You need to print out and turn in your STATA code for all empirical exercises. (ii) All STATA outputs are reported in the standard format as required in Assignment II, Problem 5(i).

1. (10 points) Read "Beta Coefficients" of Section 6-1a before answering this question. For a random variable X, define the *z-score* of X as $\frac{X-\overline{X}}{\widehat{\sigma}_X}$, where $\overline{X}$ is the sample mean and $\widehat{\sigma}_X$ is the sample deviation of X. In the simple linear regression,

$$y = \beta_0 + \beta_1 x + u, E\left[u|x\right] = 0,$$

we first calculate the z-scores of y and x, denoted as $\widetilde{y}$ and $\widetilde{x}$, and then run the regression of $\widetilde{y}$ on 1 and $\widetilde{x}$. Show that the resulting slope estimator is the sample correlation of x and y.

2. (10 points) Problem 2 of Chapter 6 in the textbook.

Note: Use the method suggested in the hint, which is different from what was taught in class.

3. (10 points) Computer Exercise C2 of Chapter 6 in the textbook.

4. (10 points) Problem 5 of Chapter 7 in the textbook.

5. (20 points) Use the data in SLEEP75.dta for this exercise. The equation of interest is

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u,$$

where *sleep* and *totwrk* (total work) are measured in minutes per week, *educ* and *age* are measured in years, and *yngkid* is a dummy variable for the presence of children less than 3 years old.

(i) Estimate this equation separately for men and women and report the results in the standard format. Are there notable differences in the two estimated equations?

(ii) Compute the Chow test for equality of the parameters in the sleep equation for men and women. Use the form of the test that adds *male* and the interaction terms $male \cdot totwrk$, $\ldots$, $male \cdot yngkid$ and uses the full set of observations. What are the relevant *df* for the test? Should you reject the null at the 5% level?

(iii) Now, allow for a different intercept for males and females and determine whether the interaction terms involving male are jointly significant.

(iv) Given the results from parts (ii) and (iii), what would be your final model?

6. (5 points) In the simple linear regression, suppose

$$y_i = \beta_0 + \beta_1 x_i + u_i, E\left[u_i|x_i\right] = 0, Var\left(u_i|x_i\right) = \sigma_i^2,$$

where σ_i^2 is known. Derive the formulae of the WLS estimator of (β_0, β_1).

7. (15 points) Suppose

$$\begin{aligned}
\widehat{lbwght} &= 7.96 - .0023 \cdot cigs + .0121 \cdot npvis - .00024 \cdot npvis^2 \\
&\quad (0.05)(.0012) \qquad (.0037) \qquad (.00012) \\
&\quad [0.05]\ [.0012] \qquad [.0051] \qquad [.00014] \\
&\quad -.00098 \cdot mage + .0022 \cdot fage - .0014 \cdot meduc + .0027 \cdot feduc \\
&\quad (.0015) \qquad (.0012) \qquad (.0030) \qquad (.0027) \\
&\quad [.0016] \qquad [.0012] \qquad [.0028] \qquad [.0027] \\
n &= 1624, R^2 = .0194
\end{aligned}$$

where $lbwght$ is the log of the birth weight, $cigs$ is the number of cigarettes, $npvis$ is the number of prenatal visits, $mage$ is mother's age, $fagem$ is father's age, $meduc$ is mother's education, and $feduc$ is father's education. The usual standard errors are in parentheses and the heteroskedasticity-robust standard errors are in brackets.

(i) Interpret the coefficient on $cigs$. Does the 95% confidence interval for β_{cigs} depend on which standard error you use?

(ii) Comment on the statistical significance of $npvis^2$, using both the usual and heteroskedasticity-robust standard errors.

(iii) If the four age and education terms are dropped from the regression (and the same set of observations is used), the R^2 becomes .0162. Develop the homoskedasticity-only test of H_0: $\beta_{mage} = 0$, $\beta_{fage} = 0$, $\beta_{meduc} = 0$, $\beta_{feduc} = 0$.

8. (20 points) Use the data in R&D_Sales_Profits.dta for this exercise. We want to explain the research and development (R&D) expenditures incurred by 18 industries. All data are in million of US dollars.

Consider the following regressions:

Model (1):

$$R\&D_i = \alpha_0 + \alpha_1 \cdot Profits_i + u_i$$

Model (2) : Variables in log

$$\ln R\&D_i = \beta_0 + \beta_1 \cdot \ln Profits_i + u_i$$

(i) Estimate both regressions and present your results in the standard format.

(ii) Using a graphical method, do you detect any evidence of heteroskedasticity in both regressions? What does this suggest about the log transformation?

(iii) Verify your qualitative conclusion in part (ii) by the White test.

(iv) Considering only Model (1), if there is evidence of heteroskedasticity,

(a) Name and apply a method to obtain robust standard errors, without changing efficiency. Compare with linear model results in (i).

(b) Name and apply a method to obtain efficient estimators. Explain how you proceed and your choices. Compare results with (a).

9. (Bonus) In this problem, we try to settle down some issues that were asked by some students or unsolved in class.

(i) (10 points) In slide 22 of Chapter 6, if we run a regression of y on $1, x_1, x_2$ and $(x_1 - \overline{x}_1)(x_1 - \overline{x}_2)$, then what is the relationship between $\left(\widehat{\alpha}_0, \widehat{\delta}_1, \widehat{\delta}_2\right)$ and the original regression coefficients $\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2\right)$? Justify your answer.

(ii) (10 points) In slide 36 of Chapter 6, we calculate $\widetilde{R}^2$ by first obtaining $\widehat{y}_i = \widehat{m}_i \widehat{\alpha}_0$, where $\widehat{m}_i = \exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik}\right\}$ and $\widehat{\alpha}_0 = n^{-1}\sum_{i=1}^{n} \exp(\widehat{u}_i)$, and then calculating $\widehat{Corr}(y, \widehat{y})^2$. An alternative goodness-of-fit measure is $\frac{\sum_{i=1}^{n}\left(\widehat{y}_i - \overline{\widehat{y}}\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$. We know that these two measures are equivalent in linear regression. Show that these two measures are not equivalent in this scenario. (**Hint**: show that the second measure is invariant to the value of $\widehat{\alpha}_0$ while the first measure is not.)