ECO 5350 T. Fomby
Intro. Econometrics Fall 2016

EXERCISE 7

Purpose: To learn more about the model selection of log(y) versus y dependent variable equations, how to use the adjusted R^2 criterion to choose between non-nested models that have the same dependent variable, to learn something about the backward, forward, and stepwise selection techniques in regression analysis, and to become somewhat more practiced with the open source program **R**. **This exercise is due on Thursday, November 3. Remember to write your work up in Word so that it is "neat and nice." I want you to turn in the .R program file that you used to complete this exercise along with your answers for the various parts below.**

Download **RStudio** to your computer and run the program **F-Test-MLB-restr.R**. Notice that the data is automatically accessible through the internet. There is an archive of all of the Stata files for your textbook on the website used by the "read.dta" procedure.

(a) Explain to me the regression that is being run. What are we analyzing here? For the definition of the variables and the source of the data you will need to go to the student resources website for the Wooldridge textbook to get the description file for the "mlb1.dta" data set.

(b) Run the regression that is given in the program. Is the model significant overall? Explain your answer.

(c) Are the individual explanatory variables of the model statistically significant? Explain your answer. Do the coefficients have the signs that you would expect? Explain.

(d) Generate a plot of log(salary). Generate a plot of salary. Which looks more normally distributed. Report both graphs.

(e) Suppose that we have the competing models

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + u$$

$$salary = \gamma_0 + \gamma_1 years + \gamma_2 gamesyr + v$$

Write code in your R program that will allow you to get the $R^2_{\hat{y}}$ and $R^2_{\tilde{y}}$ statistics that I presented in class. Which of the models do you prefer? Explain your answer.

(f) Add code to your program to estimate the following equations:

lm(log(salary) ~ years+gamesyr, data=mlb1)

lm(log(salary) ~ years+gamesyr+hruns+sbases, data=mlb1)

lm(log(salary) ~ years+gamesyr+bavg+fldperc, data=mlb1)

Which of these equations do you prefer and why?

(g) Consider the comprehensive equation lm(log(salary) ~ years+gamesyr+bavg+fldperc+hruns+sbases, data=mlb1)

Figure out a way of using R to select the important variables of this equation by using the so-called "backward elimination" technique. Hint: Take a look at the YouTube presentation at https://www.youtube.com/watch?v=TzhgPXrFSm8. Which final model was selected? Write out your selected model **in conventional form**.

(h) Separately, I want you to use R to get the models selected by the "forward selection" technique and, separately, the "stepwise" procedure. Write out the final models selected by these techniques in **conventional form**.