ECO 5350 T. Fomby
Intro. Econometrics Fall 2016

EXERCISE 7
KEY

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.

For the Key R program for this exercise see the file **Exercise 7_Key.R**.

The code is

```
library(foreign)
# Downloading the datasumma
mlb1 <- read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/mlb1.dta")
# Running the initial regression
summary(lm(log(salary) ~ years+gamesyr, data=mlb1))
reg<-lm(log(salary) ~ years+gamesyr, data=mlb1)
anova(reg)
# Getting the histograms of salary and lsalary
salary <- mlb1$salary
hist(salary)
lsalary <- mlb1$lsalary
hist(lsalary)

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

summary(step(lm(log(salary)~ years+gamesyr+bavg+fldperc+hruns+sbases, data=mlb1),
direction="backward"))
summary(step(lm(log(salary)~ years+gamesyr+bavg+fldperc+hruns+sbases, data=mlb1),
direction="forward"))
```

summary(step(lm(log(salary)~ years+gamesyr+bavg+fldperc+hruns+sbases, data=mlb1), direction="both"))

(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.

Answer:

The first R program statement output is: summary(lm(log(salary) ~ years+gamesyr, data=mlb1))

This regression analyzes the log of the salaries of major league baseball players as a function of the "years in the major leagues" (years) and the "games per year in the major leagues (gamesyr)."

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

Answer:

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.66858 -0.46412 -0.01177  0.49219  2.68829

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.223804   0.108312 103.625  < 2e-16 ***
years        0.071318   0.012505   5.703  2.5e-08 ***
gamesyr      0.020174   0.001343  15.023  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7527 on 350 degrees of freedom
Multiple R-squared:  0.5971,    Adjusted R-squared:  0.5948
F-statistic: 259.3 on 2 and 350 DF,  p-value: < 2.2e-16
```

The overall F-statistic is 259.3 with a p-value of < 0.00001. Therefore, the regression is statistically significant, overall.

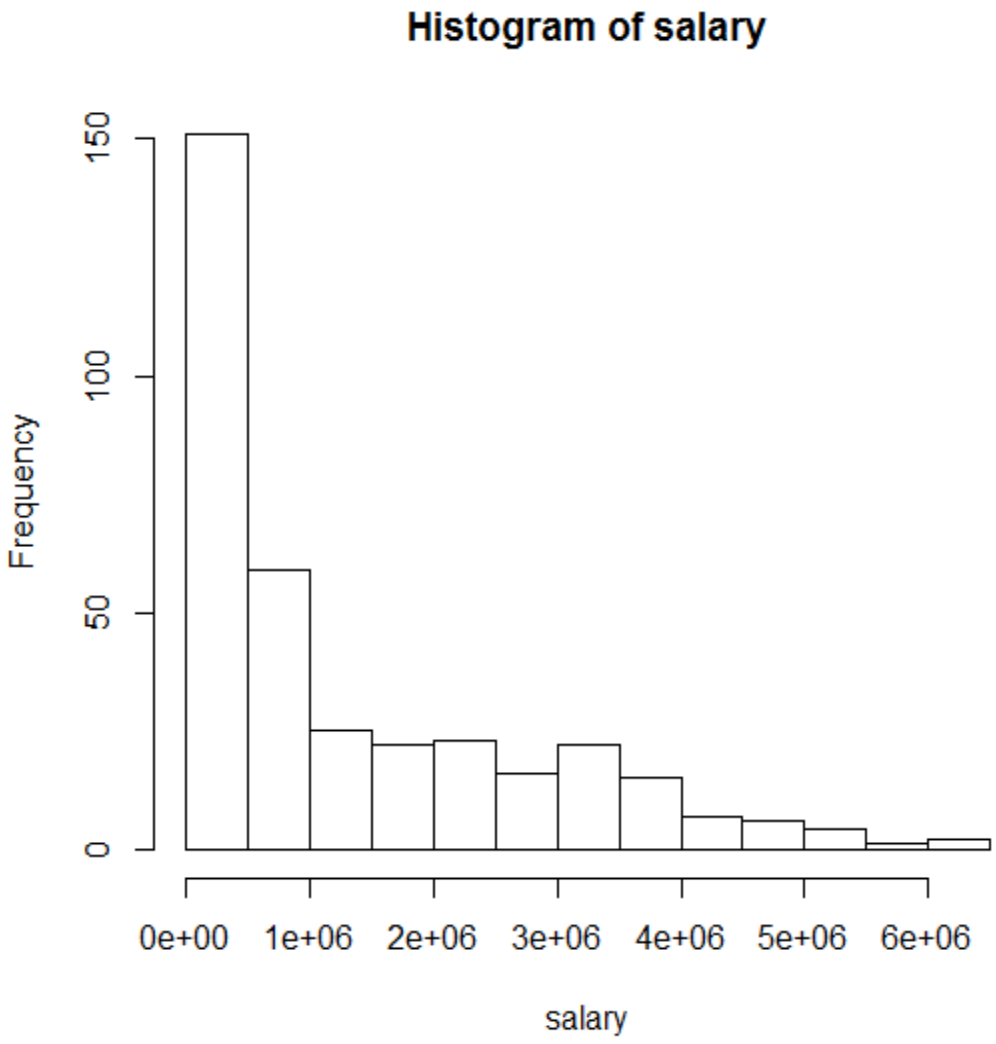
(c) Are the individual explanatory variables of the model statistically significant? Explain your answer. Do the coefficients have the signs that you would expect? Explain.

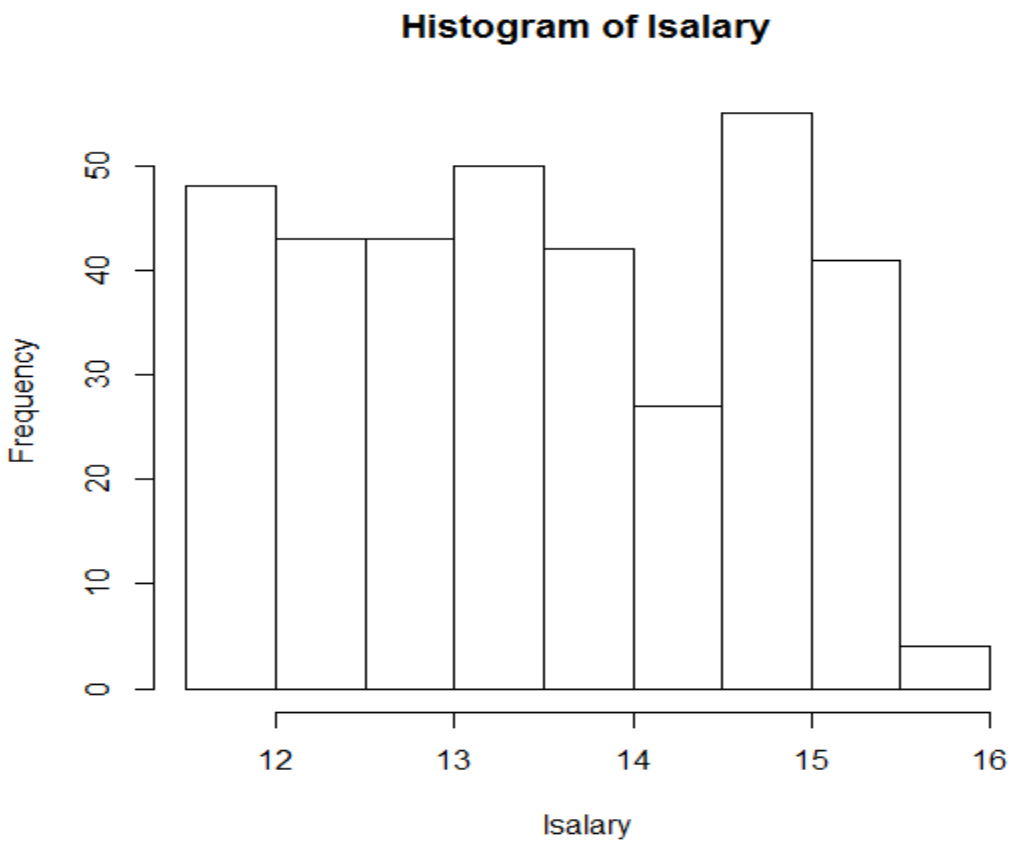
Answer:

Both of the individual coefficients are statistically significant (the p-values of the t-ratios of the coefficient estimates are less than 0.05. They also have the expected positive signs. Both "years in the major leagues" (years) and the "games per year in the major leagues (gamesyr) are expected to positively affect the salary of baseball players.

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

The *lsalary* plot looks more normal (at least symmetric). The *salary* plot is strongly skewed to the right.





(e) Suppose that we have the competing models

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

$$salary = \gamma_0 + \gamma_1 years + \gamma_2 gamesyr + v0.05.)$$

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

Answer:

The R code to execute this is something that I am still trying to figure out but here is my STATA solution to the problem:

*** Use MLB1.dta**
regress salary years gamesyr
*** R^2_y_hat = 0.4209, SST = 6.9718e+14**
regress lsalary years gamesyr
predict lsalary_hat
generate salary_tilda = exp(lsalary_hat + 0.5666/2)
generate SSR_tilda= 353*[(salary - salary_tilda)^2]

summarize SSR_tilda
*** SSR_tilda = 4.50e+14**
*** R^2_y_tilda = 1 - SSR_tilda/SST = 1 - 4.40e+14/6.9718e+14**
*** = 1 - 4.4/6.9718 = 0.3688**
*** Since R^2_y_hat > R^2_y_tilda, we favor the *salary* equation.**

(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?

Answer:

All three regressions have the same dependent variable. Therefore, we can compare the desirability of the equations by looking for equation that has the highest adjusted R^2. The adjusted R^2s of the above three regression equations are, respectively, 0.4176, 0.6004, and 0.5943. Therefore, the second equation is preferred.

(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**.

Answer:

The backward selection technique chose the following model:

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

Residuals:

Min	1Q	Median	3Q	Max
-3.10857	-0.46387	-0.02655	0.49460	2.63220

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.134e+01	1.167e-01	97.188	< 2e-16	***
years	5.149e-02	1.451e-02	3.549	0.000439	***
gamesyr	1.899e-02	1.405e-03	13.517	< 2e-16	***

```
hruns       2.026e-03  7.697e-04   2.632 0.008862 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7464 on 349 degrees of freedom
Multiple R-squared:  0.6049,    Adjusted R-squared:  0.6015
F-statistic: 178.1 on 3 and 349 DF,  p-value: < 2.2e-16
```

The conventional reporting of the backward selected model is

$$\widehat{log}(salary) = \underset{(0.1167)}{11.34} + \underset{(0.01451)}{0.05149}years + \underset{(0.001405)}{0.01899}gamesyr + \underset{(0.0.0007697)}{0.002026}hruns$$

(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**.

The forward selection technique chose the following model:

```
lm(formula = log(salary) ~ years + gamesyr + bavg + fldperc +
   hruns + sbases, data = mlb1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.09982 -0.48348 -0.02861  0.49723  2.68923

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.045e+01  2.087e+00   5.005 8.91e-07 ***
years       5.061e-02  1.614e-02   3.135  0.00187 **
gamesyr     1.847e-02  1.506e-03  12.271  < 2e-16 ***
bavg        1.402e-03  1.118e-03   1.253  0.21096
fldperc     5.990e-04  2.091e-03   0.286  0.77470
hruns       2.039e-03  7.823e-04   2.607  0.00954 **
sbases      2.179e-05  4.395e-04   0.050  0.96048
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7479 on 346 degrees of freedom
Multiple R-squared:  0.6067,    Adjusted R-squared:  0.5999
F-statistic: 88.97 on 6 and 346 DF,  p-value: < 2.2e-16
```

The conventional reporting of the forward selected model is

$$\widehat{log}(salary) = \underset{(2.087)}{10.45} + \underset{(0.01614)}{0.05061}years + \underset{(0.001506)}{0.01847}gamesyr + \underset{(0.001118)}{0.001402}bavg$$

$$+ 0.000599fldperc + 0.002039hruns + 0.00002sbases$$

(0.002091) (0.0007823) (0.0004395)

The both (stepwise) selection technique chose the following model:

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

Residuals:

Min	1Q	Median	3Q	Max
-3.10857	-0.46387	-0.02655	0.49460	2.63220

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.134e+01	1.167e-01	97.188	< 2e-16	***
years	5.149e-02	1.451e-02	3.549	0.000439	***
gamesyr	1.899e-02	1.405e-03	13.517	< 2e-16	***
hruns	2.026e-03	7.697e-04	2.632	0.008862	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7464 on 349 degrees of freedom
Multiple R-squared: 0.6049, Adjusted R-squared: 0.6015
F-statistic: 178.1 on 3 and 349 DF, p-value: < 2.2e-16

The conventional reporting of the both (stepwise) selected model is

$$\widehat{log}(salary) = \underset{(0.1167)}{11.34} + \underset{(0.01451)}{0.05149}years + \underset{(0.001405)}{0.01899}gamesyr + \underset{(0.0007697)}{0.002026}hruns$$

The stepwise and backward selection methods chose the same model.