Practice Problems 5

Econometrics, Spring 2019
Belk College of Business, UNCC
Ercan Karadas

[1] Suppose that a researcher, using wage data on 250 randomly selected male workers and 280 female workers, estimates the OLS regression

$$\widehat{\text{Wages}} = \underset{(0.23)}{12.52} + \underset{(0.36)}{2.12}\ \text{Male}, \quad R^2 = 0.06, \quad SER = 4.2$$

where `Wage` is measured in dollars per hour and `Male` is a binary variable that is equal to 1 if the person is a male and 0 if the person is a female. Define the *wage gender gap* as the difference in mean earnings between men and women.

a) What is the estimated gender gap?

b) Is the estimated gender gap significantly different from 0? (Compute the p-value for testing the null hypothesis that there is no gender gap.)

c) Construct a 95% confidence interval for the gender gap.

d) In the sample, what tis the mean wage of women? men?

e) Another researcher uses theses same data but regresses `Wages` on `Female`, a binary variable that is equal to 1 if the person is a female and 0 if the person is a male. What are the regression estimates calculated from this regression? i.e.,

$$\widehat{\text{Wages}} = ??? + ??? \times \ \text{Female}, \quad R^2 = ???, \quad SER = ???$$

[2] **[SW 5.8]** Suppose that (Y_i, X_i) satisfy the least squares assumptions in Key Concept 4.3 and, in addition, u_i is $N(0, \sigma_u^2)$ and is independent of X_i. A sample of size $n = 30$ yields

$$\widehat{Y} = \underset{(10.2)}{43.2} + \underset{(7.4)}{61.5}X, \quad R^2 = 0.54, \quad SER = 1.52,$$

where the numbers in parenthesis are the homoskedastic-only standard errors for the regression coefficients.

a) Construct a 95% confidence interval for β_0.

b) Test $H_0 : \beta_1 = 55$ vs. $H_1 : \beta_1 \neq 55$ at the 5% level.

c) Test $H_0 : \beta_1 = 55$ (or $H_0 : \beta_1 \leq 55$) vs. $H_1 : \beta_1 > 55$ at the 5% level.

[3] In the 1980s, Tennessee conducted an experiment in which kindergarten students were randomly assigned to large and small classes and given standardized tests at the end of the year. (Large classes contained approx. 24 students; small contained approx.

15 students). Let SmallClass denote a binary variable equal to 1 if the student is assigned to a small class and equal to zero otherwise. A regression of TestScore on SmallClass yields:

$$\widehat{\text{TestScore}}_i = \underset{(1.6)}{918.0} + \underset{(2.5)}{13.9} \times \text{SmallClass}_i, \quad R^2 = 0.01, \quad SER = 74.6$$

Standard errors are presented in parentheses under the estimated intercept and slope coefficient.

a) Interpret the estimated slope coefficient.

b) Interpret the R^2.

c) Is the estimated effect of class size on test score statistically significant? Perform a test at the 5% level.

d) Construct a 95% confidence interval for the effect of `SmallClass` on `TestScore`.

e) Do you think the regression errors are plausibly homoskedastic?

Now let `LargeClass` denote a binary variable equal to 1 if the student is assigned to a large class and equal to zero otherwise. Suppose you regressed `TestScore` on `LargeClass` using the same data above.

f) Write the sample regression function you would obtain. (Hint: you should be able to work out what the estimated intercept and slope coefficient would be using the results presented in question 3.)

g) Would the R^2 you obtained be different from the R^2 reported in the original regression? Explain.

[4] Consider the following demand for cell phones regression:

$$\widehat{Y} = \underset{(14.4773)}{6.1523} + \underset{(0.00032)}{0.0022} \times X, \quad R^2 = 0.6023,$$

where Y = number of cell phone subscribers per hundred persons and X = purchasing-power adjusted per capita income (in dollars) and the sample consists of $n = 134$ countries. Standard errors are reported in parentheses.

a) Is the estimated intercept coefficient significant at the 5% level of significance? What is the null hypothesis you are testing?

b) Is the estimated slope coefficient significant at the 5% level? What is the null hypothesis you are testing?

c) Establish a 95% confidence interval for the true slope coefficient.

d) Interpret the R^2.

e) What is the predicted demand for cell phones in a country where per capita income is $9000?

f) Do you think the errors are plausibly homoskedastic? Explain.

[5] a) Define homoskedasticity and heteroskedasticity.

b) Provide an example of a regression (other than one weve looked at in class or in the homework) in which you think the errors would be heteroskedastic and explain your reasoning.

c) Briefly explain the consequences of using homoskedasticity-only standard in perform- ing hypothesis tests and constructing confidence intervals for the regression model you described in (b).

[6] Suppose a researcher wishes to examine the relationship between height and weight of people by using a simple regression model:

$$\text{Weight} = \beta_0 + \beta_1 \text{Height} + u_i$$

For that purpose, she collects weight (measured in pounds) and height (measured in inches) from 100 randomly chosen men and runs a regression which produces the following output:

$$\widehat{\text{Weight}} = \underset{(3.42)}{-79.24} + \underset{(0.42)}{4.16}\text{Height}, \quad R^2 = 0.72, \quad SER = 12.6$$

a) Interpret the intercept estimate?

b) Interpret the slope estimate?

c) Test the null hypothesis that the population slope coefficient is zero against the alternative that it is not at the 5% level of significance. What is your conclusion?

d) Test the null hypothesis that the weight increases at least 4 pounds for per inch at the 5% level of significance. What is your conclusion?

e) A man has a late growth spurt and grows 2 inches over the course of a year. Construct a 95% confidence interval for the person's weight gain.

f) What is the expected weight of a man who is 5' 9"?

g) Suppose OLS regression output above was based on a sample of only 25 randomly chosen men instead of 100. Which of the answers a) through f) would have been effected, discuss briefly why.

[7] A researcher investigates the relationship between age and circumference of orange trees. She employs the following population regression model:

$$\text{Circumference}_i = \beta_0 + \beta_1 X_i + u_i$$

where X_i is a binary variable that takes the value 1 if the age of tree i is greater than 900 (Old Trees) and 0 (Young Trees) otherwise. The OLS regression in R produces the following output:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    ...        8.092   7.498 1.28e-08 ***
age           96.583     10.704    ...      ...
---
Residual standard error: 31.34 on 73 degrees of freedom
Multiple R-squared:  0.7116
```

a) Using the regression output above write down the estimated sample regression line.

b) What is the estimated average circumference for young trees?

c) What is the estimated average circumference for old trees?

d) Interpret the slope estimate?

e) Is the difference between circumferences of old and young trees statistically significant? Explain.

   Another researcher uses `age` as the independent variable instead of working with a binary variable and she obtains the following regression output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.399650   8.622660   2.018   0.0518 .
age          0.106770   0.008277  12.900 1.93e-14 ***
```

f) Write the sample regression function that you would obtain in this case.

g) Would the R^2 you obtained here be different from the R^2 reported in the original regression? Speculate whether you would expect higher or lower R^2 in this model compare to the original.

Empirical Exercises

[8] E4.2 (Empirical Exercise 4.2 in SW)

[9] E5.1 (Empirical Exercise 5.1 in SW)