San Francisco State University **Michael Bar**
ECON 312 **Fall 2018**

Problem set 3

Due Thursday, October 11, in class

Names of collaborators (type below):

Name 1 MINJIN PARK

Name 2 NISHANLANG KHONGLAH

Assignment Rules

1. Homework assignments must be typed. For instruction on how to type equations and math objects please see notes "Typing Math in MS Word".
2. Homework assignments must be prepared within this template. Save this file on your computer and type your answers following each question. Do not delete the questions.
3. Your assignments must be stapled.
4. No attachments are allowed. This means that all your work must be done within this word document and attaching graphs, questions or other material is prohibited.
5. Homework assignments must be submitted at the end of the lecture, in class, on the listed dates.
6. Late homework assignments will not be accepted under any circumstances, but the lowest homework score will be dropped.
7. The first homework assignment cannot be dropped.
8. You are encouraged to work on this homework assignment in groups of up to 3 people, and submit one assignment with up to 3 names typed on this page. Sharing the electronic version of your assignment with other teams is absolutely prohibited.
9. All the graphs should be fully labeled, i.e. with a title, labeled axis and labeled curves.
10. In all the questions that involve calculations, you are required to show all your work. That is, you need to write the steps that you made in order to get to the solution.
11. This page must be part of the submitted homework.

Unbiasedness of OLS Estimators

1. (10 points). We showed in the notes that the OLS estimators can be written as the sum of the true coefficient and a linear combination of the error terms:

$$b_2 = \beta_2 + \sum_{i=1}^{n} a_i u_i$$

$$b_1 = \beta_1 + \sum_{i=1}^{n} c_i u_i$$

Prove that OLS estimators are unbiased.

The observations on X are non-random, therefore the weights a_i and c_i are also non-random.

Taking the expectation of b_2, we have

$$\begin{aligned} E(b_2) &= E\left(\beta_2 + \sum_{i=1}^{n} a_i u_i\right) \\ &= \beta_2 + E\left(\sum_{i=1}^{n} a_i u_i\right) \\ &= \beta_2 + \sum_{i=1}^{n} a_i E(u_i) \ \text{ By assumption A0, X is non} - \text{random and } X_{is} \text{ are fixed} \\ &= \ \beta_2 + 0 \ \text{ By assumption A3, the error term has zero expectation} \\ &= \beta_2 \end{aligned}$$

Therefore, we can observe that b_2 is an unbiased estimator of β_2

Similarly, taking the expectation of b_1, we have

$$\begin{aligned} E(b_1) &= E\left(\beta_1 + \sum_{i=1}^{n} c_i u_i\right) \\ &= \ \beta_1 + E\left(\sum_{i=1}^{n} c_i u_i\right) \\ &= \beta_1 + \sum_{i=1}^{n} c_i E(u_i) \ \text{ By assumption A0, X is non} - \text{random and } X_{is} \text{ are fixed} \\ &= \beta_1 + 0 \text{ By assumption A3, the error term has zero expectation} \\ &= \beta_1 \end{aligned}$$

Therefore, we can observe that b_1 is an unbiased estimator of β_1

2. (10 points). Suppose that the true model is

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Instead of OLS, a researcher is using a different estimator of β_2:

$$\tilde{b}_2 = \frac{\bar{Y}_n}{\bar{X}_n}$$

In other words, the estimator is the ratio of the sample mean of *Y* and *X*.

a. Prove that $\tilde{b}_2$ is a biased estimator of β_2 and find the bias.

We know that

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ and } \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Therefore,

$$\widetilde{b_2} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{Y_i}{X_i}$$

$$\text{Let} \frac{1}{\mathrm{X_i}} = \mathrm{M}$$

$$\begin{aligned} \widetilde{b_2} &= Y_i \,\mathrm{M} \\ &= (\beta_1 + \beta_2 X_i + u_i)\,\mathrm{M} \\ &= \mathrm{M}\,\beta_1 + \beta_2 + \mathrm{M}\,u_i \qquad (X_i \text{ and M cancel out}) \end{aligned}$$

$$\begin{aligned} E(\widetilde{b_2}) &= \mathrm{E}\,(\mathrm{M}\,\beta_1 + \mathrm{M}\,\beta_2 + \mathrm{M}\,u_i\,) \\ &= \mathrm{E}\,(\mathrm{M}\,\beta_1) + \mathrm{E}\,(\,\beta_2) + \mathrm{E}\,(\mathrm{M}\,u_i\,) \\ &= \mathrm{M}\,\beta_1 + \beta_2 \quad \text{By assumption A3, the error term has zero expectation} \end{aligned}$$

We have shown that $\widetilde{b_2}$ is biased.

$$\begin{aligned} \text{Bias}\left(\widetilde{b_2}\right) &= \mathrm{M}\,\beta_1 + \beta_2 - \beta_2 \\ &= \mathrm{M}\,\beta_1 \end{aligned}$$

b. Prove that if $\beta_1 = 0$, then $\tilde{b}_2$ is unbiased.

$$bias(\tilde{b}_2) = \frac{\beta_1}{\bar{X}_n} = 0 \quad because\ \beta_1 = 0$$

From (a) we have,

$$E(\widetilde{b_2}) = \mathrm{M}\,\beta_1 + \beta_2$$

$$\text{If } \beta_1 = 0$$

$$E(\widetilde{b_2}) = \beta_2$$

We have shown that if $\beta_1 = 0$, then $\tilde{b}_2$ is unbiased and bias $(\tilde{b}_2) = 0$

Create an R script, which performs all the analysis for questions 3 and 4. You can name the script `HW3.R`. You can either print out the script and attach it as a separate page at the end of this assignment, or copy and paste its content at the end of your assignment. Make sure to add comments explaining which question you are solving, and every command in your script.

t-tests of regression coefficients

3. (80 points). For this exercise use data from http://www.stata-press.com/data/r12/auto.dta, which is in Stata format. We investigate the relationship between `mpg` (gas mileage) and `weight` (car's weight in pounds). The variable `foreign` indicates whether the car is domestic or foreign made. This is a categorical variable, with categories "`Domestic`", and "`Foreign`".

   a. Create a new variable `weight100`, which is the weight in hundreds of pounds. The R command is:

```
auto$weight100 <- auto$weight/100.
```

   Present summary statistics of the variables `mpg`, `weight100,` and `foreign`.

```
mpg                    weight100               foreign
Min.:      12.00       Min.:      17.60        Domestic:  52
1st Qu.:  18.00        1st Qu.:   22.50        Foreign:     22
Median:  20.00         Median:  31.90
Mean:     21.30        Mean:     30.19
3rd Qu.:  24.75        3rd Qu.:  36.00
Max.:      41.00       Max.:      48.40
```

   b. Based on the summary statistics in the last section, how many foreign made cars and how many domestic cars are in the sample?

   Based on the summary statistics, there are 52 domestic-made cars and 22 foreign-made cars.

   c. Present a summary statistics of `mpg` by categories of `foreign`. The R command is:

```
tapply(auto$mpg, auto$foreign,  summary, na.rm=TRUE)
```

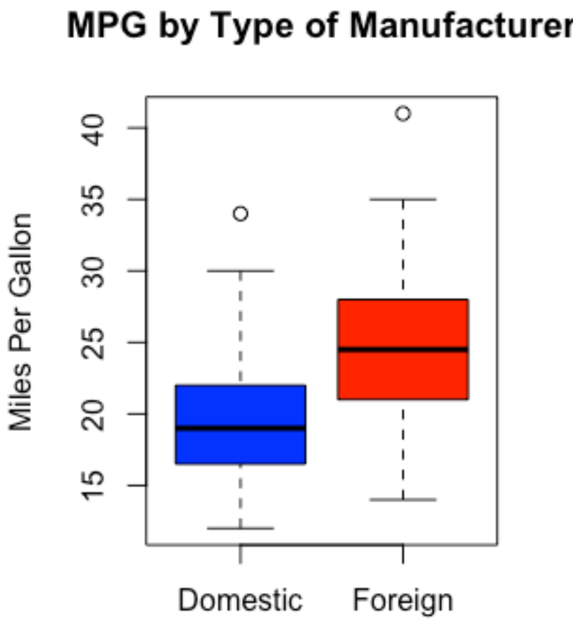
```
$Domestic
Min.    1st Qu.  Median    Mean     3rd Qu.   Max.
 12.00   16.75    19.00     19.83    22.00     34.00

$Foreign
Min.    1st Qu.  Median    Mean    3rd Qu.   Max.
14.00   21.00    24.50     24.77   27.50     41.00
```

   d. Based on the summary statistics in the last section, what is the mean `mpg` of domestic cars, and what is the mean `mpg` of foreign made cars?

   Based on the summary statistics, the mean mpg of domestic cars is 19.83 miles per gallon, while the mean mpg of foreign made cars is 24.77 miles per gallon.

   e. Present a boxplot diagram that visually summarizes the `mpg` by categories of `foreign`. The R command is:

```
boxplot(mpg~foreign,data=auto,
        main="MPG by Type of Manufacturer",
        ylab="Miles Per Gallon",col=c("blue","red"))
```



f. Run a regression of `mpg` on `weight100`. Present the R commands and R output here.

```
#Regression of mpg vs weight100
model1 <- lm(mpg ~ weight100, data=auto)
summary(model1) #Output of regression

lm(formula = mpg ~ weight100, data = auto)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9593 -1.9325 -0.3713  0.8885 13.8174

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.44028    1.61400   24.44   <2e-16 ***
weight100   -0.60087    0.05179  -11.60   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.439 on 72 degrees of freedom
Multiple R-squared:  0.6515,  Adjusted R-squared:  0.6467
F-statistic: 134.6 on 1 and 72 DF,  p-value: < 2.2e-16
```

g. Interpret the estimated regression coefficients.

   b2 = -0.60
   This means that for every additional 1 pound in the car's weight, the gas mileage is predicted to decrease by 0.60 miles per gallon.

   b1 = 39.44
   This is the predicted gas mileage of a car with 0 weight, which does not make real-world sense.

h. Suppose that we want to test whether there is relationship between `mpg` and `weight100`, and we choose significance level of $\alpha = 0.05$.

   i. State the null and alternative hypotheses.

      The null hypothesis is given by
      $$H_0 : \beta_2 = 0$$
      The alternative hypothesis is given by
      $$H_1 : \beta_2 \neq 0$$

   ii. Find the critical t-values for this test. In R, the command for the lower-tail critical value ($-t_c$) is `qt(alpha/2, df)`, where alpha is significance level and `df` is the RSS degrees of freedom: $df = n - k = 74 - 2 = 72$. You can automatically calculate the degrees of freedom after estimation, by `df.residual(model1).`

      The critical t-value for this test is -1.99

   iii. Calculate the test statistic.

      The test statistic for this test is -11.6

   iv. Write the conclusion of the test.

      Because the calculated t-value = -11.6, which is < critical t-value = -1.99, we **reject the null hypothesis H_0 at significance level of $\alpha = 0.05$** and conclude that **a car's weight has some impact on gas mileage**.

i. Suppose that we want to test whether there is relationship between `mpg` and `weight100`, and we choose significance level of $\alpha = 0.01$.

   i. State the null and alternative hypotheses.

The null hypothesis is given by

$$H_0 : \beta_2 = 0$$

The alternative hypothesis is given by

$$H_1 : \beta_2 \neq 0$$

ii. Find the critical t-values for this test.

The critical t-value for this test is -2.65.

iii. Calculate the test statistic.

The test statistic for this test is -11.6

iv. Write the conclusion of the test.

Because the calculated t-value = -11.6, which is < the critical t-value = -2.65, we **reject the null hypothesis H_0 at significance level of α = 0.01** and conclude that **a car's weight has some impact on gas mileage**.

j. Suppose that theory tells us that if weight affects the gas mileage, the effect must be **negative**. Assume significance level of $\alpha = 0.05$.

i. State the null and alternative hypotheses.

The null hypothesis is given by

$$H_0 : \beta_2 = 0$$

The alternative hypothesis is given by

$$H_1 : \beta_2 < 0$$

ii. Find the critical t-values for this test. In R, the command for the left (lower-tail) critical value is `qt(alpha, df)`, where alpha is significance level and `df` is the RSS degrees of freedom.

The critical t-value for this test is -1.66

iii. Calculate the test statistic.

The test statistic for this test is -11.6

iv. Write the conclusion of the test.

Because the calculated t-value = -11.6, which is < the critical t-value = -1.66, we **reject the null hypothesis H_0 at significance level of α = 0.05** and conclude that **a car's weight has a negative impact on gas mileage**.

4. (40 pt). This question uses the wage21 data set, posted on the course webpage in several formats, and was used in HW2.

The key variables that you need to know for this assignment are:
EARNINGS – hourly earnings, in $ per hour.
S – schooling, in years.
EXP – years of experience.

a. Run the regression of earnings on schooling and experience and present the R command and R regression output.

   The R command is given by the following:

```
model2 <- lm(EARNINGS ~ S + EXP, data = wage) #OLS estimation
summary(model2)
```

   The output is the following:

```
lm(formula = EARNINGS ~ S + EXP, data = wage)

Residuals:
   Min       1Q      Median     3Q        Max
 -28.077  -6.717   -2.146     3.976     89.112



Coefficients:
               Estimate      Std. Error      t value Pr(>|t|)
(Intercept)    -28.3098      4.0274          -7.029 6.33e-12 ***
S               2.8057       0.2189          12.820  < 2e-16 ***
EXP             0.5671       0.1218           4.657 4.04e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.87 on 537 degrees of freedom
Multiple R-squared:  0.238,    Adjusted R-squared:  0.2352
F-statistic: 83.88 on 2 and 537 DF,  p-value: < 2.2e-16
```

b. Write the fitted equation and interpret the estimated regression coefficients.

   The fitted equation can be written as the following

$$\widehat{EARNINGS} = b_1 + \mathrm{S}\, b_2 + \mathrm{EXP}\, b_3 + u_i$$

   Where b_1, b_2, and b_3 are the estimated regression coefficients, S and EXP denote schooling and experience, respectively, and u_i denotes all influences on earnings other than schooling and experience.

   The estimated regression coefficient b2 = 2.80. This means that every additional year of schooling increases hourly earnings by $2.80/hour.

The estimated regression coefficient b3 = 0.57. This means that every additional year of work experience increases hourly earnings by $0.57/hour.

The estimated regression coefficient b1 = -28.30. This means that the predicted hourly earnings for someone with 0 years of schooling and no work experience is $ -28.30/hour, which makes no real-world sense.

c. Suppose that we want to test whether there is any relationship between earnings and experience, and we choose significance level of $\alpha = 0.05$.

   i. State the null and alternative hypotheses.

      The null hypothesis is given by
      $$H_0 : \beta_3 = 0$$
      The alternative hypothesis is given by
      $$H_1 : \beta_3 \neq 0$$

   ii. Find the critical t-values for this test. To find the lower tail critical value in R, use `qt(alpha/2, df)`, where alpha is the significance level of the test, and `df` is the degrees of freedom of RSS: $df = n - k = 540 - 3 = 537$.

      The critical t-value for this test is 1.96.

   iii. Calculate the test statistic.

      The test statistic for this test is 4.66.

   iv. Write the conclusion of the test.

      Because the calculated t-value = 4.66, which is > the critical t-value = 1.96, we **reject the null hypothesis H$_0$ at significance level of α = 0.05** and conclude that **experience has some impact on a person's earnings**.

d. Suppose that economic theory tells us that if experience has any effect on earnings, the effect must be positive. Choose significance level of $\alpha = 0.05$.

   i. State the null and alternative hypotheses.

      The null hypothesis is given by
      $$H_0 : \beta_3 = 0$$
      The alternative hypothesis is given by
      $$H_1 : \beta_3 > 0$$

   ii. Find the critical t-values for this test. In R use the command: `qt(alpha, df, lower.tail = FALSE)`, which gives the upper-tail critical value.

      The critical t-value for this test is 1.65.

iii. Calculate the test statistic.

The test statistic for this test is 4.66.

iv. Write the conclusion of the test.

Because the calculated t-value = 4.66, which is > the critical t-value = 1.65, we **reject the null hypothesis H_0 at significance level of $\alpha = 0.05$** and conclude that **experience has a positive impact on a person's earnings**.

e. Suppose that economic theory tells us that each year of schooling increases the hourly earnings by $3 per hour. Choose significance level of $\alpha = 0.05$.

i. State null and alternative hypotheses.

The null hypothesis is given by

$$H_0 : \beta_2 = 0$$

The alternative hypothesis is given by

$$H_1 : \beta_2 > 0$$

ii. Find the critical t-values for this test.

The critical t-value for this test is -1.64.

iii. Calculate the test statistic.

The test statistic for this test is -0.89

iv. Write the conclusion of the test.

Because the calculated t-value = -0.89, which is < the critical t-value = -1.64, we **do not reject the null hypothesis H_0 at significance level of $\alpha = 0.05$** and conclude that **the theory is true, that is, schooling increases the hourly earnings of a person by $3/hour**.