ECON471
Fall 2020
Problem Set 2
Due Wednesday September 30, by 11:59pm CST

Name: Albert Wiryawan

Section: B3

* For Question 1-3 Check below the problem set for the attached work

1. Show that the R^2 from the regression of Y on X is the same as the R^2 from the regression of X on Y.

2. Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + u_i$. Given the n observations $\{(x_i, y_i),\ \ i = 1,2, \dots \dots, n\}$, we estimate β_1 as

$$\tilde{\beta}_1 = \frac{1}{n-1} \sum_{i=2}^{n} \frac{(y_i - y_{i-1})}{(x_i - x_{i-1})}$$

a. Give a geometric interpretation of $\tilde{\beta}_1$.

b. Show that $\tilde{\beta}_1$ is an unbiased estimator of β_1. Be sure to state the assumptions needed to prove this.

3. Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + u_i$.

a. Show that the OLS estimator $\hat{\beta}_1$ can be written as $\hat{\beta}_1 = r_{x,y} \frac{s_y}{s_x}$, where $r_{x,y}$ is the sample correlation between x and y, and s_y and s_x are the sample standard deviations of y and x, respectively.

b. Show that the OLS estimator $\hat{\beta}_0$ is an unbiased estimator of β_0. (Hint: Use the fact that $\hat{\beta}_1$ is an unbiased estimator of β_1.)

4. Use the data in ATTEND.TXT for this exercise.

(i) Obtain the minimum, maximum, and average values for the variables $atndrte$, $priGPA$, and ACT, where $atndrte$ is percent classes attended, $priGPA$ is cumulative GPA prior to term, and ACT is the ACT score.

Code

```{r}
setwd("/Users/albertwiryawan/Code/Class_Repos/Econometrics/Problem Set #2")

#load in Attend txt data set for the exercise
attend <- read.table("attend.txt")

attach(attend)
```

```{r}
#atndrte
max_atndrte= max(V6)
min_atndrte= min(V6)
mean_atndrte= mean(V6)


#priGPA
max_priGPA = max(V3)
min_priGPA = min(V3)
mean_prGPA = mean(V3)

#ACT
max_ACT = max(V4)
min_ACT = min(V4)
mean_ACT = mean(V4)
print(max_ACT)
```

The maximum percent classes attended was 100% while the minimum percent classes attended was 6.25%. The average percent of class attendance was 81.71%.

The maximum cumulative GPA prior was 3.93 while the minimum GPA was 0.857. The average GPA for the class prior to the term was 2.59%

The maximum ACT score was 32 while the minimum ACT score was 13. The average ACT score of the class was 22.51%

(ii) Estimate the model

$$atndrte = \beta_0 + \beta_1\ priGPA + \beta_2\ ACT + u,$$

and write the results in equation form. Interpret the intercept. Does it have a useful meaning?

Code

```{r}
model = lm(attend$V6~attend$V3 + attend$V4)
summary(model)

```

The equation form is given by atndrte = 75.7 + 17.261(priGPA) - 1.717(ACT). The intercept means that given a GPA of 0 and ACT score of 0 students attend 75.7% of classes.

(iii) Interpret the estimated slope coefficients. Are there any surprises?

Something that was surprising about the produced model is that ACT score decreased the school attendance. Generally, one would expect that someone that attends more class would acquire a better score on the ACT exam

(iv) What is the predicted $atndrte$ if *priGPA* = 3.65 and *ACT* = 20? What do you make of this result? Are there any students in the sample with these values of the explanatory variables?

Code

```{r}
similar_ACT = nrow(subset(attend, V3==3.65))
similar_GPA = nrow(subset(attend, V4==20))

```

The predicted value of atndrte was 104.36265 which, technically, should be above the percentage of classes possible to attend. There is about 1 student with the same GPA, but 71 students with a similar ACT

(v) If Student A has *priGPA* = 3.1 and *ACT* = 21 and Student B has *priGPA* = 2.1 and *ACT* = 26, what is the predicted difference in their attendance rates?

The student with a priGPA = 3.1 and ACT of 21 was found to have attendance percentage of 93.1521% while the student with priGPA = 2.1 and ACT = 26 had an attendance percentage of 25.846%. The predicted difference in the two attence rates is 25.846%

5. Use the data in HPRICE1.TXT to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

where *price* is the house price in thousands of dollars.

(i) Write out the results in equation form.

Code

```{r}
hprice1 <- read.table("hprice1.txt")

attach(hprice1)
price = V1
sqrtft = V5
bdrms = V3
model = lm(price ~ sqrtft + bdrms)
summary(model)
```

The predictive modeled equation is defined as
Price = -19.315 + 0.12844(Sqrft) + 15.19819 (bdrms)

(ii) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

The estimated increase in price for a house with one more bedroom is $15198.19. This is given by the modeled coefficient from the multiple linear regression model.

(iii) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).

Code

```{r}
x_1 = predict(model, data.frame(bdrms = 2, sqrtft = 140))
x_2 = predict(model, data.frame(bdrms = 3, sqrtft = 140))
x_2 - x_1
```

The estimated increase in price for a house with an additional bedroom that is 140 sqrft in size is $15198.19. This is the same as the price increase found from the coefficient in the previous equation as the square footage is independent of the increase in price from the increase in the number of bedrooms.

(iv) What percentage of variation in price is explained by square footage and number of bedrooms?

The variation in price that is explained by the square footage and number of bedrooms is seen through the Adjusted R^2 which is 62.33%. This means that these two explanatory variables explain 62.33% variation of price.

(v) The first house in the sample has *sqrft* = 2,438 and *bdrms* = 4. Find the predicted selling price for this home from the OLS regression line.

Code

```{r}
predict(model, data.frame(bdrms = 4, sqrtft = 2438))
```

The house is predicted to have a selling price of $354,605.20.

(vi) The actual selling price of the first house in the sample was $300,000 (so *price* = 300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

The residual contribution by this specific data point is given by the
(observation – predicted) as a result this value is -54.6052. The directionality of this value indicates that the buyer underpaid in respect to the model. This is because he is paying less than the model would predict.

6. Use the data set in WAGE2.TXT for this problem. As usual, be sure all of the following regressions contain an intercept.

(i) Run a simple regression of *IQ* on *educ* to obtain the slope coefficient, say,$\tilde{\delta}_1$.

Code

```
```{r}
wage2 <- read.table("wage2.txt")

attach(wage2)
IQ = V3
education = V5
model1 = lm(IQ ~ education)
summary(model1)
```
```

The value of the coefficient $\tilde{\delta}_1$. is found to be 3.5338.

(ii) Run the simple regression of log(*wage*) on *educ*, and obtain the slope coefficient, $\tilde{\beta}_1$.

Code

```
```{r}
wage = V1
model2 = lm(log(wage) ~ education)
summary(model2)
```
```

The coefficient found from the fitted simple regression was found to be 0.059839

(iii) Run the multiple regression of log(*wage*) on *educ* and *IQ*, and obtain the slope coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

Code

```
```{r}
model3 = lm(log(wage) ~ IQ + education)
summary(model3)
```
```

The coefficient assigned to education and IQ given by the coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ the values were 0.0391199 and 0.0058631

(iv) Verify that $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$.

0.059839 = 0.0058631 + 0.0391199 (3.5338)

The expressions are equivalent.

Question #1

Show that the R^2 from the regression of Y on X is the same as the R^2 from the regression of X on Y.

$$SSE = \sum_{i=1}^{n} (\hat{X}_i - \bar{X})^2$$

predicted model simple linear regression: $\hat{X} = \hat{\beta}_0 + \hat{\beta}_1 Y$ * plug back into equation

$$SSE = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 Y_i - \bar{X})^2 \qquad \beta_0 = \bar{x} - \hat{\beta}_1 \bar{Y}$$

$$SSE = \sum_{i=1}^{n} (\bar{X} - \hat{\beta}_1 \bar{Y} + \hat{\beta}_1 Y_i - \bar{X})^2$$

$$SSE = \sum_{i=1}^{n} (\hat{\beta}_1 (Y_i - \bar{Y}))^2$$

$$* \hat{B}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

$$SSE = \hat{\beta}_1^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SSE = \left(\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \right)^2 \sum (Y_i - \bar{Y})^2$$

$$SSE = \frac{\left(\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) \right)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

$$R^2 = \frac{SSE}{SST} = \frac{\left(\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) \right)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \Big/ \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \frac{\left(\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) \right)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$R^2 = \left(\frac{\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}} \right)^2 = \left(\frac{s_{yx}}{s_y s_x} \right)^2$$

$R^2 = r_{yx}^2$ this means that the regression on Y on X has the same R^2 from the regression of X on Y.

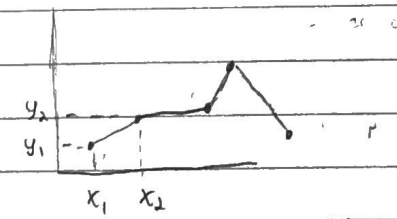
Question #2

a) Geometric representation of $\tilde{\beta}_1$

$$\tilde{\beta}_1 = \frac{1}{n-1} \sum_{i=2}^{n} \frac{y_i - y_{i-1}}{x_i - x_{i-1}}$$

$$\tilde{\beta}_1 = \frac{1}{n-1} \left[\frac{y_2 - y_1}{x_2 - x_1} + \frac{y_3 - y_2}{x_3 - x_2} + \dots + \frac{y_n - y_{n-1}}{x_n - x_{n-1}} \right]$$

Slope



- average of slopes show the geometric representation
- there are $n-1$ total slopes

b) $$\tilde{\beta}_1 = \frac{1}{n-1} \sum_{i=2}^{n} \frac{y_i - y_{i-1}}{x_i - x_{i-1}}$$

$$\tilde{\beta}_1 = \frac{1}{n-1} \sum_{i=2}^{n} (\beta_0 + \beta_1 x_i + u_i) - (\beta_0 + \beta_1 x_{i-1} + u_{i-1})$$

$$\tilde{\beta}_1 = \frac{1}{n-1} \sum_{i=2}^{n} \frac{\beta_1 (x_i - x_{i-1})(u_i - u_{i-1})}{(x_i - x_{i-1})}$$

$$\tilde{\beta}_1 = \frac{1}{n-1} \sum_{i=2}^{n} \beta_1 + \frac{u_i - u_{i-1}}{x_i - x_{i-1}} \to 0$$

* $E(u_i) = 0$ assumption

$E(u_{i-1}) = 0$

$$\tilde{\beta}_1 = \frac{n-1}{n-1} \beta_1$$

$$\tilde{\beta}_1 = \beta_1$$

Question #3

a) $r_{x,y} = \frac{cov(x,y)}{s_x s_y}$

$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{cov(x,y)}{s_x^2}$

$= \frac{r_{x,y} s_x s_y}{s_x^2} = \frac{r_{x,y} s_y}{s_x}$

$\hat{\beta}_1 = r_{x,y} \frac{s_y}{s_x}$

b) $\hat{\beta}_0 = \bar{y} - \hat{\beta} \bar{x}$ given x $\bar{y} = \beta_0 + \beta_1 \bar{x}$

$E(\hat{\beta}_0) = E(y) - E(\hat{\beta}_1 \bar{x})$

$= E(y) - \bar{x} E(\hat{\beta}_1)$

$= E(\beta_0 + \beta_1 \bar{x}) - \bar{x} E(\hat{\beta}_1)$

$= E(\hat{\beta}_0) + E(\beta_1 \bar{x}) - \bar{x} E(\hat{\beta}_1)$

$= \beta_0 + \bar{x} E(\beta_1) - \bar{x} E(\hat{\beta}_1)$,

$= \beta_0 + \bar{x} \beta_1 - \bar{x} \beta_1$ Unbiased estimator these cancel

$E(\hat{\beta}_0) = \beta_0$