Assignment 1
Assigned on Feb 8th, 2022. Due date Saturday, Feb 26th, 2022 before midnight.

Instructions

(1) Please read the instruction carefully. Also take this habit with you into the exam room.
(2) Please read each question carefully and answer the questions straightforwardly. Always provide economic reasons at lease a paragraph for your analysis, or a graph when necessary, even when the question does not indicate so.
(3) Handing and submitting assignments are only available via BE Moodle.

Answering the questions and preparing answer sheets

(1) Answers are to be handwritten, in either digital or analog form, in a blank canvas or any clean paper. Make sure that your handwriting is clearly visible and readable.
(2) There is no need to rewrite the question. Just indicate the question number clearly for each of the answer, such as 1.a).
(3) Default decimal point is 4.
(4) Choose precise wordings, especially when you want to interpret the meaning of a test, confidence interval, or coefficients.
(5) When done, for the digital case, collage all the pages into a single PDF file. For those who write on sheets of paper, take photo of all pages then convert all of them into a single PDF file as well.
(6) Name your PDF file as StudentID_YourNickname, such as 640123456_Bo.

Submitting your answers

(1) Make sure your file does not exceed 10MB. This is the maximum file size for BE Moodle upload.
(2) Login to BE Moodle, head into the course, then the assignment topic.
(3) Choose your file to submit. Done. There will be timestamp for your upload date and time, so please make sure to not submit later than that.

Assignment 1
Assigned on Feb 8th, 2022. Due date Saturday, Feb 26th, 2022 before midnight.

For all questions, answer up to 4 decimal places

Question 1. (15 points) Given this information

$$n = 18 \qquad \sum_{i=1}^{n} X_i = 388.00 \qquad \sum_{i=1}^{n} Y_i = 50.90$$

$$\sum_{i=1}^{n} (X_i)^2 = 9{,}620.00 \qquad \sum_{i=1}^{n} X_i Y_i = 1{,}254.90$$

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = 211.00 \qquad \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 2.5844$$

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 20.58 \qquad \sum_{i=1}^{n} \hat{u}_i^{\,2} = 0.5781$$

Use the above sample information to answer all the following questions. Show explicitly all formulas and calculations.

a) From regression model: $Y_i = \beta_1 + \beta_2 X_i + u_i$, $u_i \sim NIID\left(0, \sigma^2\right)$, **find the estimators** of β_1 and β_2 with OLS method. Interpret the intercept and slope coefficients.

b) Compute the value of R^2 and explain its meaning.

c) If $X_i = 30$, estimate the value of $\hat{Y}_i$ and explain its meaning.

d) Calculate the estimators of $\text{var}\left(u_i\right)$, $\text{var}\left(\hat{\beta}_1\right)$ and $\text{var}\left(\hat{\beta}_2\right)$.

e) What are the 90-percent confident intervals for β_2 ? Interpret the meaning.

f) Test the hypothesis whether the slope coefficients are different from zero at 0.05 level of significance.

Question 2. Using the 2015 Health and Welfare Survey from the National Statistical Office, a simple linear regression is modeled as follows,

$$outp_i = \beta_1 + \beta_2 age_i + u_i$$

where $outp_i$ is how many times person i has visited hospital in 2015, from 0 to 7 times

age_i is how old is person i, from 0 to 97 years.

We assume that both $outp_i$ and age_i are continuous, the estimation results in the following table. Answer the following questions and show your work.

```
      Source |       SS           df       MS      Number of obs   =    27,886
-------------+----------------------------------   F(1, 27884)     =    186.96
       Model |  77.5444409         1  77.5444409   Prob > F        =    0.0000
    Residual |  11565.0627    27,884  .414756231   R-squared       =    0.0067
-------------+----------------------------------   Adj R-squared   =    0.0066
       Total |  11642.6072    27,885  .417522223   Root MSE        =    .64402

------------------------------------------------------------------------------
        outp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0031338   .0002292                      .0026846     .003583
       _cons |   .4279898   .0140339      Omitted         .4004828    .4554969
------------------------------------------------------------------------------
```

a) Test if both parameters are significantly different from zero or not. Use $\alpha = 0.05$.

b) Interpret the meaning of $\hat{\beta}_2$. Does the sign of $\hat{\beta}_2$ make economic sense? Explain.

c) If $outp_i$ is turned into natural logarithmic scale (ln), how would you reinterpret the relationship between $\hat{\beta}_2$ and $\widehat{outp}_i$, assumed that the given coefficient given in the table above can be used to interpret this new functional form.

d) If age_i variable is divided by 10, how does it affect both the coefficients, standard errors, and confidence intervals? Answer the changes of both the constant and slope (if there is).

e) Find the confidence interval of mean prediction at the age of 50 years old, given that $var(\hat{Y}_0) = 0.00002$ and $\alpha = 0.01$.

Question 3. Discuss in a short paragraph why the confidence interval for both the mean prediction and individual prediction get larger as the X_0 is further away from $\bar{X}$.

Question 1. (15 points) Given this information

$n = 18$ $\quad \sum_{i=1}^{n} X_i = 388.00$ $\quad \sum_{i=1}^{n} Y_i = 50.90$

$\sum_{i=1}^{n} (X_i)^2 = 9{,}620.00$ $\quad \sum_{i=1}^{n} X_i Y_i = 1{,}254.90$

$\sum_{i=1}^{n} (X_i - \bar{X})^2 = 211.00$ $\quad \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 2.5844$

$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 20.58$ $\quad \sum_{i=1}^{n} \hat{u}_i^2 = 0.5781$

Use the above sample information to answer all the following questions. Show explicitly all formulas and calculations.

a) From regression model: $Y_i = \beta_1 + \beta_2 X_i + u_i$, $u_i \sim NIID(0, \sigma^2)$, **find the estimators** of β_1 and β_2 with OLS method. Interpret the intercept and slope coefficients.

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

$$\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

$$\min_{\beta_1, \beta_2} \sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

Find $\hat{\beta}_1$ $\quad \frac{\partial \sum \hat{u}_i^2}{\partial \beta_1} = \sum 2(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) \cdot -1$

F.O.C. $\quad -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_1 - \sum_{i=1}^{n} \hat{\beta}_2 X_i = 0$$

$$\sum_{i=1}^{n} Y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^{n} X_i = 0$$

$$n\hat{\beta}_1 = \sum_{i=1}^{n} Y_i - \hat{\beta}_2 \sum_{i=1}^{n} X_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i - \hat{\beta}_2 \sum_{i=1}^{n} X_i}{n} = \bar{Y} - \hat{\beta}_2 \bar{X}$$

$\bar{Y} = \frac{50.9}{18} = 2.8278$

$\bar{X} = \frac{388}{18} = 21.5556$

Find $\hat{\beta}_2$

$$\min_{\beta_1,\beta_2} \sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_2} = \sum 2(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) \cdot -X_i = 0$$

$$-2 \sum X_i (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\sum X_i (Y_i - (\bar{Y} - \hat{\beta}_2 \bar{X}) - \hat{\beta}_2 X_i = 0$$

$$\sum X_i (Y_i - \bar{Y} + \hat{\beta}_2 \bar{X} - \hat{\beta}_2 X_i) = 0$$

$$\sum X_i (Y_i - \bar{Y} - \hat{\beta}_2 X_i + \hat{\beta}_2 \bar{X}) = 0$$

$$\sum X_i (Y_i - \bar{Y} - \hat{\beta}_2 (X_i - \bar{X})) = 0$$

$$\sum X_i (Y_i - \bar{Y}) - \hat{\beta}_2 \sum X_i (X_i - \bar{X}) = 0$$

$$\hat{\beta}_2 = \frac{\sum X_i (Y_i - \bar{Y})}{\sum X_i (X_i - \bar{X})}$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i (Y_i - \bar{Y}) - \bar{X} \sum (Y_i - \bar{Y}) = \sum X_i (Y_i - \bar{Y})$$

$$\sum (X_i - \bar{X})(X_i - \bar{X}) = \sum X_i (X_i - \bar{X}) - \bar{X} \sum (X_i - \bar{X}) = \sum X_i (X_i - \bar{X})$$

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})(X_i - \bar{X})}$$

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{20.68}{211} = 0.0975$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 2.8278 - 0.0975(21.5556) = 0.7261$$

$\beta_1 \approx 0.7261$ is the intercept. It means when $X_i = 0$, Y_i is around 0.7261

$\hat{\beta}_2 \approx 0.0975$ is the slope of this line. It means when X increases by 1 unit Y will increase by approximately 0.0975 unit.

b) Compute the value of R^2 and explain its meaning.

$$TSS = ESS + RSS$$

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

$$\left(\div \sum (Y_i - \bar{Y})^2\right)$$

$$1 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} + \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}$$

$$1 = r^2 + \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}$$

$$r^2 = 1 - \frac{0.5781}{2.5844} = 0.7763$$

, When we devide all the equation by TSS or $[\sum (Y_i - \bar{Y})^2]$, we will see that $r^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{ESS}{TSS}$, it is a proportion of explained sum of square to total sum of square. $R^2 \in (0, 1)$, the more R^2 implies the less proportion of residual sum of square (error).

c) If $X_i = 30$, estimate the value of $\hat{Y}_i$ and explain its meaning.

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$$= 0.7261 + 0.0975 (30)$$

$$= 3.651$$

When X_i is 30, Y_i is around 3.651.

We use the word "around" because it is just an estimation of Y_i

d) Calculate the estimators of $\text{var}(u_i)$, $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_2)$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k} = \frac{RSS}{d.f.} = \frac{0.5781}{18-2} = 0.0361$$

$$\sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n\sum(X_i-\bar{X})^2} \cdot \sigma^2 = \frac{9620}{18(211)} \cdot 0.0361 = 0.0914$$

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum(X_i-\bar{X})^2} = \frac{0.0361}{211} = 1.71 \times 10^{-9}$$

e) What are the 90-percent confident intervals for β_2 ? Interpret the meaning

$\alpha = 0.1$ $\quad d.f. = 16$ $\quad se_{\hat{\beta}_2} = \sqrt{1.71 \times 10^{-9}} = 0.0131$

$t_{16,0.05} = 1.746$

$\hat{\beta}_2 = 0.0975$

$$P(\hat{\beta}_2 - t_{16,0.05}(se_{\hat{\beta}_2}) \leq \beta_2 \leq \hat{\beta}_2 + t_{16,0.05}(se_{\hat{\beta}_2})) = 1-\alpha$$

$$P(0.0975 - 1.746(0.0131) \leq \beta_2 \leq 0.0975 + 1.746(0.0131)) = 1-\alpha$$

$(0.0746 \leq \beta_2 \leq 0.1204)$ → it means that 90 out of 100 time we run the test, β_2 will be in between $0.0746 \leq \beta_2 \leq 0.1204$.

f) Test the hypothesis whether the slope coefficients are different from zero at 0.05 level of significance.

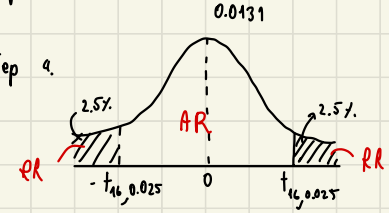
Step 1. $H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

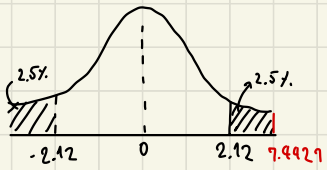
Step 2. $\alpha = 0.05$

Step 3. $t = \frac{0.0975 - 0}{0.0131} = 7.4427$

Step 4.



Step 5 Conclude



Since our $t_{cal} > t_{16,0.025}$. It falls into rejection region (RR). So, we can reject the null hypothesis and conclude that with 95% confident interval β_2 is not equal to zero.

Question 2. Using the 2015 Health and Welfare Survey from the National Statistical Office, a simple linear regression is modeled as follows,

$$outp_i = \beta_1 + \beta_2 age_i + u_i$$

where $outp_i$ is how many times person i has visited hospital in 2015, from 0 to 7 times

age_i is how old is person i, from 0 to 97 years.

We assume that both $outp_i$ and age_i are continuous, the estimation results in the following table. Answer the following questions and show your work.

```
      Source |       SS           df       MS      Number of obs   =    27,886
-------------+----------------------------------   F(1, 27884)     =    186.96
       Model |  77.5444409         1  77.5444409   Prob > F        =    0.0000
    Residual |  11565.0627    27,884  .414756231   R-squared       =    0.0067
-------------+----------------------------------   Adj R-squared   =    0.0066
       Total |  11642.6072    27,885  .417522223   Root MSE        =    .64402

------------------------------------------------------------------------------
        outp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0031338   .0002292                      .0026846    .003583
       _cons |   .4279898   .0140339      Omitted         .4004828    .4554969
------------------------------------------------------------------------------
```

a) Test if both parameters are significantly different from zero or not. Use $\alpha = 0.05$.

For β_1 : constant

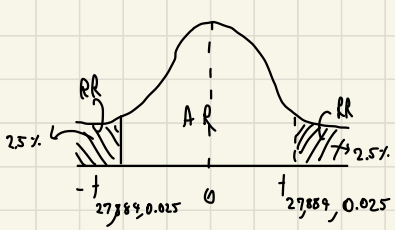
Step 1 : H_0 : $\beta_1 = 0$

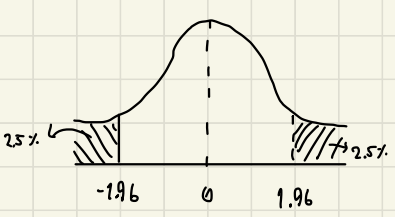
H_1 : $\beta_1 \neq 0$

Step 2 : $\alpha = 0.05$

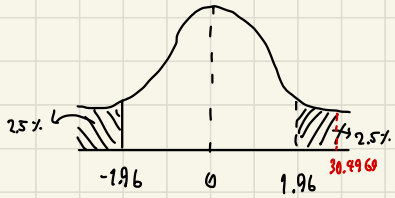
Step 3 : $t_{cal} = \frac{0.4279898 - 0}{0.0140339} = 30.4969$

Step 4 :





Step 5 : Conclude



According to our result, t_{cal} falls into the rejection region. This means we can reject the null hypothesis. Therefore, we could say that from 95% confident interval β_1 (constant) is not equal to 0.

For β_2 = age

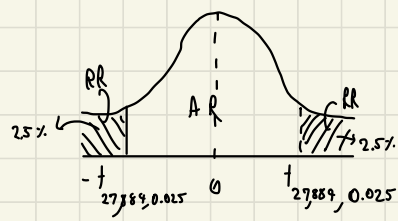
Step 1 : $H_0 : \beta_2 = 0$

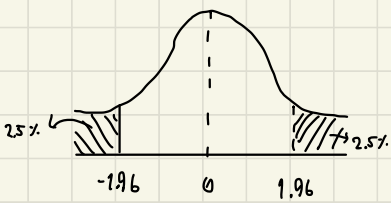
$H_1 : \beta_2 \neq 0$

Step 2 : $\alpha = 0.05$

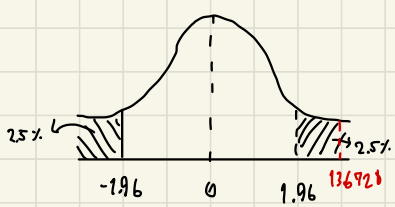
Step 3 : $t_{cal} = \dfrac{0.0031338 - 0}{0.0002292} = 13.6728$

Step 4 :





Step 5 :



According to our result, t_{cal} falls into the rejection region. This means we can reject the null hypothesis. Therefore, we could say that from 95% confident interval β_2 (age) is not equal to 0.

b) Interpret the meaning of $\hat{\beta}_2$. Does the sign of $\hat{\beta}_2$ make economic sense? Explain.

$\hat{\beta}_2$ is the slope of this function, it shows that when you getting older (x increases), you will visit the hospital more often.

The positive sign is quite make sense. As we can see in general that as people getting older and older, they face the health problem. Therefore, when age increase you will visit hospital more often is make economic sense.

c) If $outp_i$ is turned into natural logarithmic scale (ln), how would you reinterpret the relationship between $\hat{\beta}_2$ and $\widehat{outp}_i$, assumed that the given coefficient given in the table above can be used to interpret this new functional form.

$\ln(\widehat{outp}_i) = \hat{\beta}_1 + \hat{\beta}_2(age)$

$\widehat{outp}_i : Y \qquad age : X$

$\ln(Y_i) = \hat{\beta}_1 + \hat{\beta}_2 X_i$

$\frac{d \ln(Y_i)}{dX} = \hat{\beta}_2$

$\hat{\beta}_2 = \frac{dy/y}{dx} = 0.0031338$, assumed

$\frac{dy/y}{dx} = \hat{\beta}_2$

As people getting older by 1 years, $\widehat{outp}_i$ or Y will increase by $0.0031338 \cdot 100$ percent

$\frac{dy}{dx} \cdot \frac{1}{y} = \hat{\beta}_2$

$\frac{dy}{dx} = y\hat{\beta}_2$ → Slope of the line

$\times \frac{X}{y}$

$\frac{dy}{dx} \cdot \frac{x}{y} = x\hat{\beta}_2$ → Elasticity

d) If age_i variable is divided by 10, how does it affect both the coefficients, standard errors, and confidence intervals? Answer the changes of both the constant and slope (if there is).

For constant β_1 : there will be no change on it. This is because Y_i is not change, so when $x=0$ intercept still be the same

For variable $\hat{\beta}_2$: the $\hat{\beta}_2$, $se_{\hat{\beta}_2}$, $CI_{\hat{\beta}_2}$ will change, it will be multiply by 10. Since as we deviide the data, scale on the x-axis is decreased, so the slope is now steeper as the unit change

new -> $\hat{\beta}_2 = 0.031338$ $se_{\hat{\beta}_2} = 0.002292$ CI: $0.026846 \leq \beta_2 \leq 0.03583$

e) Find the confidence interval of mean prediction at the age of 50 years old, given that $var(\hat{Y}_0) = 0.00002$ and $\alpha = 0.01$.

$$se_{\hat{Y}_0} = \sqrt{var(\hat{Y}_0)} = \sqrt{0.00002} = 4.4721 \times 10^{-3}$$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2$$

$$\hat{Y}_i = 0.4279898 + 0.0031337\, X_i$$

at age : $X = 50$ $\hat{Y}_{50} = 0.4279898 + 0.0031338(50) = 0.5846798$

$$P\left(\hat{Y}_{50} - t_{2777,\,0.005} \cdot se_{\hat{Y}_0} \leq Y_{50} \leq \hat{Y}_{50} + t_{2777,\,0.005} \cdot se_{\hat{Y}_0}\right) = 1 - 0.01$$

$$P\left(0.5846798 - 2.576(4.4721 \times 10^{-3}) \leq Y_{50} \leq 0.5846798 + 2.576(4.4721 \times 10^{-3})\right) = 0.99$$

$$CI : \left(0.5732 \leq Y_{50} \leq 0.5962\right)$$

Question 3. Discuss in a short paragraph why the confidence interval for both the mean prediction and individual prediction get larger as the X_0 is further away from $\bar{X}$.

For mean prediction : $$\text{var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

$$Pr\left[\hat{Y}_0 - (t_{\frac{\alpha}{2}} \cdot se_{\hat{Y}_0}) \leq Y \leq \hat{Y}_0 + (t_{\frac{\alpha}{2}} \cdot se_{\hat{Y}_0}) \right] = 1 - \alpha$$

For individual prediction : $$\text{var}(fe) = \text{var}(\hat{Y}_0 - Y_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

$$Pr\left[\hat{Y}_0 - (t_{\frac{\alpha}{2}} \cdot se_{fe}) \leq Y_0 \leq \hat{Y}_0 + (t_{\frac{\alpha}{2}} \cdot se_{fe}) \right] = 1 - \alpha$$

By looking into the variance equation, distance between X_0 and $\bar{X}$ will increase the variance. As variance increases, standard error (se) will increase. This is because se is coming from square root of variance. Finally, at the same level of confident interval, we need a larger distance between lower bound and upper bound.

In other words, we can roughly say that as the distance between X_0 and $\bar{X}$ increases. the data dispersion is wider, so at the same level of confident interval we need a larger distance between lower bound and upper bound.