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.

$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{20.58}{211} = 0.0975$

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

$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{50.90}{18} = 2.8278$

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{388}{18} = 21.5556$

(expect)

Thus, $[\hat{\beta}_1 = 0.7261]$ when $X = 0$, $Y = 0.7261$

$[\hat{\beta}_2 = 0.0975]$ $X_i \uparrow$ 1 unit, $Y \uparrow$ on average by 0.0978 unit.

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.

b) $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = 1 - \frac{0.5781}{2.5844} = 0.7763$

$R^2 \rightarrow 0.7763$: X variable can explain 77.63 % of variation in Y

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

$= 0.7261 + 0.0975(30)$

$\hat{Y}_i = 3.6511$ on average when $X_i = 30$

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

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.

0.0 6 0 9 7 1 0 9

d) $\text{var}(u_i) = \sigma^2 = \frac{\sum \hat{u}_i^2}{n-k} = \frac{0.5781}{16} = 0.0361$

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

$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} = \frac{0.0361}{211} = 0.0002$

e) $se(\hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_2)} = \sqrt{0.0002} = 0.0141$

$\alpha = 0.1 \rightarrow t_{16, 0.05} = 1.746 \quad : \hat{\beta}_2 = 0.0975$

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

$P(0.0975 - (1.746 \cdot 0.0141) \leq \beta_2 \leq 0.0975 + (1.746 \cdot 0.0141)) = 0.9$

$\hookrightarrow P(0.0729 \leq \beta_2 \leq 0.1221) = 0.9$

Thus, β_2 will be $0.0729 \leq \beta_2 \leq 0.1221$

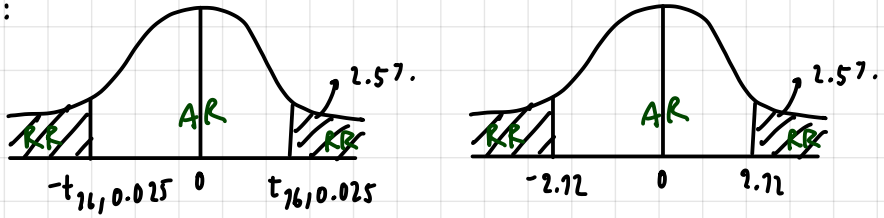
f) Test the hypothesis whether the slope coefficients are different from zero at 0.05 level of significance.

step 1: $H_0 : \beta_2 = 0$
$H_1 : \beta_2 \neq 0$

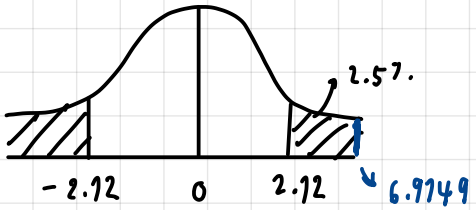
step 2: $\alpha = 0.05$

step 3: $t = \frac{0.0975 - 0}{0.0141} = 6.9149$

step 4:



step 5:



since $t_{cal} > t_{16,0.025}$, reject the null hypothesis and can conclude that with 95% confident interval, the $\beta_2 \neq 0$

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 $\alpha = 0.05$.

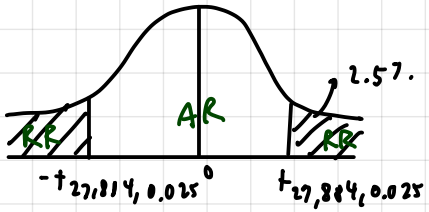
b) Interpret the meaning of $\hat{\beta}_2$. Does the sign of $\hat{\beta}_2$ make economic sense? Explain.

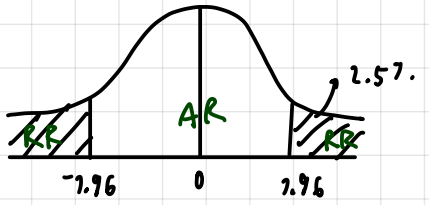
step 1: $H_0 : \beta_1 = 0$ (β_1 = constant)
$H_1 : \beta_1 \neq 0$

step 2: $\alpha = 0.05$

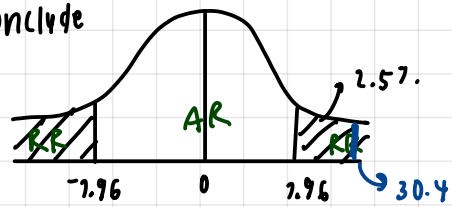
step 3: $t_{cal} = \frac{0.4279898 - 0}{0.0140339} = 30.4969$

step 4:





step 5: conclude



t_{cal} falls in rejection region (RR)
Thus, We can reject null hypothesis
so, with 95% confident interval
β_1 (constant) $\neq 0$

β_2 = age

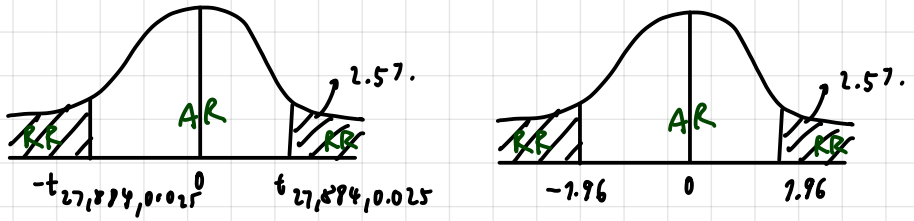
step 1 = H_0 = $\beta_2 = 0$
H_1 : $\beta_2 \neq 0$

step 2 = $\alpha = 0.05$

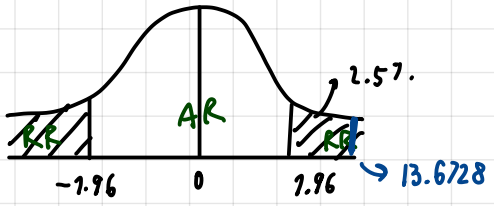
step 3 =

$$t_{cal} = \frac{0.0031338 - 0}{0.0002292} = 13.6728$$

step 4 =



step5



t_{cal} falls in RR → we can reject null hypothsis.
so, with 95% confident interval, β_2(age) ≠ 0

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

b) when the age of people ↑ 1 yr, we can expect that the visit / yr will increase 0.0031 time on average.
This make economics sense! because we trend to use more medical service as we get older.

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.

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

$\ln(y_i) = \hat{\beta}_1 + \hat{\beta}_2 x_i$

$$\rightarrow \frac{d\ln(y_i)}{dx} = \hat{\beta}_2$$

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

$$\frac{dy}{dx} \cdot \frac{1}{y} = \hat{\beta}_2 \rightarrow \frac{dy}{dx} = y\hat{\beta}_2 \rightarrow \text{slope of the line}$$

$$\frac{dy}{dx} \cdot \frac{x}{y} = x\hat{\beta}_2 \rightarrow \text{elasticity}$$

when age increase by 1 year, visiting hospital by people will increase by $\hat{\beta}_2 \cdot 100 = 0.31334\%$.

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 $\alpha = 0.01$.

d) constant $\hat{\beta}_1$: no change → y_i does not change, and thus if $x = 0$, the intercept still the same
variable $\hat{\beta}_2$: if $\hat{\beta}_2$, $se_{\hat{\beta}_2}$, $CI_{\hat{\beta}_2}$ change, they will be multiply by 10.
as the data is divided, the x axis's scale will then decrease [slope steeper]

new one: $\hat{\beta}_2 = 0.031338$
$se_{\hat{\beta}_2} = 0.002292$
CI = $0.026846 \leq \beta_2 \leq 0.035833$

e) $se_{\hat{y}_0} = \sqrt{Var(\hat{y}_0)} = \sqrt{0.0002} = 4.4721359\underline{55} \times 10^{-3}$

$\alpha = 0.01$

$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

$\hat{y}_1 = 0.4279898 + 0.00331338\,(50) = 0.5846798$

$P(\hat{y}_{50} - t_{21884,\,0.005} \cdot se_{\hat{y}_0} \leq y_{50} \leq \hat{y}_{50} + t_{21884,\,0.005} \cdot se_{\hat{y}_0}) = 1-\alpha$

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

so → CI: $(0.5732 \leq y_{50} \leq 0.5962) = 0.99$

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 $\bar{X}$.

mean prediction : $Var(\hat{y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_1 - \bar{x})^2}{\Sigma (x_i - \bar{x})^2} \right]$

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

on the other hand, individual prediction : $Var(f_e) = Var(\hat{y}_0 - y_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma (x_i - \bar{x})^2} \right]$

$Pr\left[\hat{y}_0 - (t_{\frac{\alpha}{2}} \cdot se_{fe}) \leq y_0 \leq \hat{y}_0 + (t_{\frac{\alpha}{2}} \cdot se_{fe}) \right] = 1 - \alpha$

as X_0 is farther away from $\bar{X}$ [larger in $X_0 - \bar{x}$], the $se(\hat{y}_0)$ will be larger because the variance get larger.

→ when x_0 move further away from $\bar{x}$, the lesser data points of x_i. Therefore the CI must get larger to tackle with the unknown.