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

NOTE: If you use up to 4 decimal places for your calculation and the values you get are different, you will still receive full mark.

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

- $\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{20.58}{211} = \mathbf{0.0975}$
- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$, First, we have to find $\bar{Y}$ and $\bar{X}$ that is
- $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{50.90}{18} = 2.8278$ and $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{388}{18} = 21.5556$ then
- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 2.8278 - 0.0975(21.5556) = \mathbf{0.7261}$
- To interpret the coefficients: when $X_i = 0$, we expect that $\hat{Y}$ is 0.7261 and when X_i increases by 1 unit, we can expect that $\hat{Y}$ will increase on average by 0.0978 unit.

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

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

- $R^2 = 0.7763$ means that X variable can explain 77.63 percent of variation in Y.

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

- From the SRF $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_i = 0.7261 + 0.0975X_i$, we can plug $X_i = 30$ then
- $E(\hat{Y}|X_i = 30) = 0.7261 + 0.0975(30) = \mathbf{3.6511}$ or when $X_i = 30$ we expect that the value of $\hat{Y}$ will be **3.6511** on average.

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

- $var(u_i) = \hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k} = \frac{0.5781}{18-2} = \mathbf{0.0361}$
- $var(\hat{\beta}_1) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2 = \frac{9{,}620}{18(211)} \cdot 0.0361 = \mathbf{0.0914}$
- $var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} = \frac{0.0361}{211} = \mathbf{0.0002}$

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

- Find $se(\hat{\beta}_2) = \sqrt{var(\hat{\beta}_2)} = \sqrt{0.0002} = 0.0141$
- $t_{0.05} = \mathbf{1.746}$ for $\alpha = 0.1$; then
- $\Pr(0.0975 - (1.746 * 0.0141) \leq \beta_2 \leq 0.0975 + (1.746 * 0.0141)) = 0.90$
- $\mathbf{Pr(0.0729 \leq \beta_2 \leq 0.1221) = 0.90}$

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

- $H_0: \beta_2 = 0$; $H_a: \beta_2 \neq 0$
- From e), $se(\hat{\beta}_2) = 0.0141$
- Find $t_{cal} = \frac{0.0975-0}{0.0141} = 6.9149$
- Critical t-value for $\alpha = 0.05$ and degrees of freedom of n-k or 18-2 = 16 is 2.120.
- Therefore, we can reject the null hypothesis, which means that we can make sure that **95 percent, β_2 is not 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    13.67   0.000     .0026846     .003583
       _cons |   .4279898   .0140339    30.50   0.000     .4004828    .4554969
------------------------------------------------------------------------------
```

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

- $H_0: \beta_1 = 0$; $H_a: \beta_1 \neq 0$
- Find $t_{cal} = \frac{0.4279898-0}{0.0140339} \approx 30.4969$ (or 30.5714 if only 4 decimal places are used)
- $H_0: \beta_2 = 0$; $H_a: \beta_2 \neq 0$
- Find $t_{cal} = \frac{0.0031338-0}{0.0002292} \approx 13.6728$ (or 15.5 if only 4 decimal places are used)
- Critical t-value for $\alpha = 0.05$ and degrees of freedom tends to infinity is 1.96.
- Therefore, we can reject null hypothesis for both tests, which means that we can make sure that **95 percent, β_1, β_2 are not zero.**

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

- When people age 1 more year, we can expect that visit per year will increase on average of 0.0031 times. This positive sign makes economic sense because as people age, we tend to have to rely more on medical services.

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.

- If $outp_i$ is turned into natural logarithmic scale, or

$$\ln \widehat{outp}_i = \hat{\beta}_1 + \hat{\beta}_2 age_i$$

- then, the interpretation of $\hat{\beta}_2$ becomes: if age_i increases by a year, we expect that people's visiting hospital will increase by $\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}} * \mathbf{100} = \mathbf{0.3134}$ percent.

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).

- If age_i variable is divided by 10, the coefficient, SE, and CI will be scaled up by 10 while all the values of the constant remain the same as displayed in this table here.

outp	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	**.031338**	**.002292**	13.67	0.000	**.026846**	**.03583**
_cons	.4279898	.0140339	30.50	0.000	.4004828	.4554969

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

- $\hat{Y}_0 = 0.428 + 0.0031(50) = \mathbf{0.583}$;
- $\text{se}(\hat{Y}_0) = \sqrt{var(\hat{Y}_0)} = \mathbf{0.0045}$;
- $t_{0.005} = \mathbf{2.576}$; then
- $\Pr\left(0.583 - (2.576 * 0.0045) \leq Y_0 \leq 0.583 + (2.576 * 0.0045)\right) = 0.99$
- $\mathbf{Pr(0.5714 \leq Y_0 \leq 0.5946) = 0.99}$

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 $\bar{X}$.

The variance used for calculating confidence intervals for the mean prediction is

$$var(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(x_i - \bar{X})^2}\right]$$

From the equation, if $X_0 - \bar{X}$ gets larger, or X_0 is farther away from $\bar{X}$, the variance gets larger leading to larger $se(\hat{Y}_0)$.

In other words, $\bar{X}$ is the central tendency of X_i, which means that the further away from $\bar{X}$, the less information or data points of X_i is available. Hence, the confidence interval must be larger to cope with more unknown.
