Question 1. (15 points) Given this information

$n = 18$	$\sum_{i=1}^{n} X_i = 388.00$	$\sum_{i=1}^{n} Y_i = 50.90$
$\sum_{i=1}^{n} (X_i)^2 = 9{,}620.00$	$\sum_{i=1}^{n} X_i Y_i = 1{,}254.90$	
$\sum_{i=1}^{n} (X_i - \bar{X})^2 = 211.00$	$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 2.5844$	
$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 20.58$	$\sum_{i=1}^{n} \hat{u}_i^2 = 0.5781$	

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.

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

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

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

Find β_1

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

FOC $$-2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

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

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

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

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

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

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

Find β_2

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

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

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

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

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

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

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

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

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

The intercept is β_1 which mean that when $X_i = 0$ $Y_i = 0.7261$

The slope is β_2 which mean that when X increases by 1 unit Y increases by 0.0975 unit

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

TSS = ESS + RSS

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

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

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

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

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

If we devide everything by TSS

$r^2 = \frac{ESS}{TSS}$ which is a proportion of explained sum of sq. to total sum of sq.

r^2 can be from (0, 1) which imply that the more r^2, the less the proportion of residual sum of square.

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

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

$$\hat{Y}_i = 0.7261 + 0.975(30)$$

$$\hat{Y}_i = 3.651$$

When X_i is equal 30, Y_i is equal to 3.651

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

$$Var(u_i) = \hat{\sigma}^2 = \frac{RSS}{d.f.} = \frac{\sum \hat{U}_i^2}{n-k} = \frac{0.5781}{18-2} = 0.0361$$

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

$$Var(\hat{\beta}_2) = \hat{\sigma}^2_{\hat{\beta}_2} = \frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2} = \frac{0.0361}{211} = 1.7109 \times 10^{-4}$$

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

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

$\alpha = 0.1$ $\quad \hat{\beta}_2 = 0.0475$

$se_{\hat{\beta}_2} = \sqrt{1.7104 \times 10^{-4}}$ $\quad d.f. = 16$

$= 0.0131$

$$P[0.0475 - 1.746(0.0131) \leq \beta_2 \leq 0.0475 + 1.746(0.0131)]$$

$$P(0.0746 \leq \beta_2 \leq 0.1204)$$

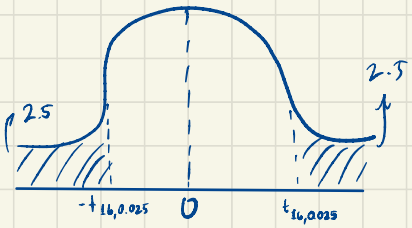
90% of confident interval means that 90% of the time β_2 will fall in between $0.0746 \leq \beta_2 \leq 0.1204$

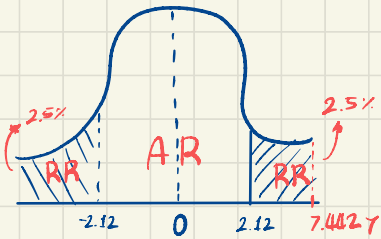
f) Test the hypothesis whether the slope coefficients are different from zero at 0.05 level of significance.

$H_0 : \beta_2 = 0$; $\alpha = 0.05$

$H_1 : \beta_2 \neq 0$

$$t = \frac{0.0475 - 0}{0.0131} = 7.4427$$





t_{cal} falls between the rejection region RR.
Therefore, reject the null hypothesis.
β_2 is not equal to 0 with 90% confident interval.

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

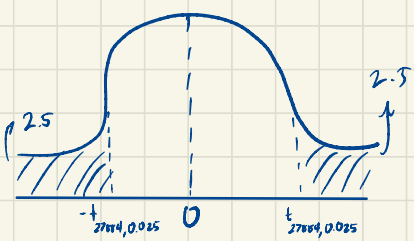
outp	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	.0031338	.0002292	Omitted		.0026846	.003583
_cons	.4279898	.0140339			.4004828	.4554969

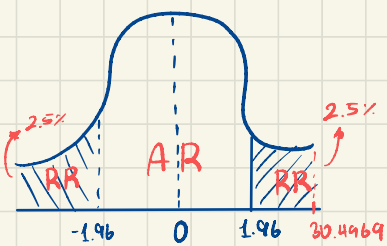
a) Test if both parameters are significantly different from zero or not. Use $\alpha = 0.05$.

β_1

$H_0: \beta_1 = 0$ $\quad \alpha = 0.05$

$H_1: \beta_1 \neq 0$ $\quad t_{cal} = \frac{0.4279898 - 0}{0.0140339} = 30.4469$





t_{cal} falls into rejection region (RR)

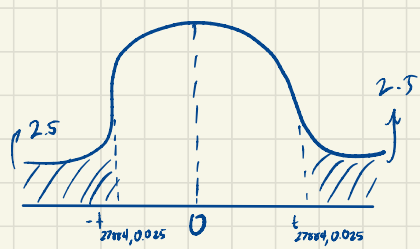
Therefore, reject null hypothesis.

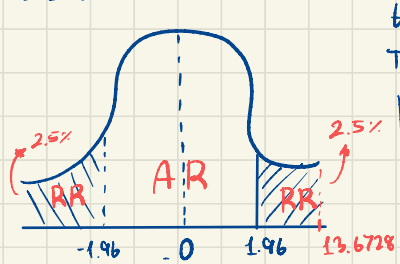
β_1 is not equal to 0 with 95% confident interval

β_2 : age

$H_0: \beta_2 = 0$ $\quad \alpha = 0.05$

$H_1: \beta_2 \neq 0$ $\quad t_{cal} = \frac{0.0031338 - 0}{0.0002292} = 13.6728$





t_{cal} falls into rejection region (RR)

Therefore, reject null hypothesis.

β_2 is not equal to 0 with 95% confident interval.

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 the function. It does make economic sense since as you are getting older. You are likely to encounter with health issue more often which result to visiting hospital more often.

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.

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

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

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

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

Slope

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

Elasticity

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

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

The interpretation is as people get older by 1yr $Outp_i$ will increase by 0.0031338

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).

$\hat{\beta}_1$: The value will remain the same since y_i hasn't changes when $x = 0$

$\hat{\beta}_2$: $\hat{\beta}_2$, std error and confidence interval will be multiply by 10

$\hat{\beta}_2 = 0.031338$

$se_{\hat{\beta}_2} = 0.002242$

CI : $0.026646 \leq \beta_2 \leq 0.03583$

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

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

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

$$\hat{Y}_i = 0.4279898 + 0.0031338(X_i)$$

$$\hat{Y}_{50} = 0.4279898 + 0.0031338(50)$$
$$= 0.5846798$$

$$P(\hat{Y}_{50} - t_{29884, 0.005} \cdot se_{\hat{Y}_0} \leq Y_{50} \leq \hat{Y}_{50} + t_{29884, 0.005} \cdot se_{\hat{Y}_0}) = 0.99$$

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

Confident interval

$$= (0.5732 \leq Y_{50} \leq 0.5962)$$

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

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

The variance equation suggests that as the distance of X_0 and $\bar{X}$) is further away. The data dispersion is more spread which result in larger distance of lower bound and upper bound.