Natthapat Worrawattanapreecha 6304640805

Assignment 1

(1a) $\hat{\beta}_2 = \frac{\sum(x_i-\bar{x})(Y_i-\bar{Y})}{\sum(x_i-\bar{x})^2} = \frac{20.58}{211} = 0.0975$ //

$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} = 2.7883 - 0.0975(21.5556)$

$= 2.7883 - 2.1017$

$= 0.6866$ //

$\hat{Y}_i = \bar{\hat{Y}} = \bar{Y} = \frac{50.19}{18} = 2.7883$

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

The SRF $\hat{Y}_i = 0.6866 + 0.0975X_i$

The intercept is 0.6866 and slope is 0.0975

(1b) $R^2 = 1 - \frac{\sum\hat{u}_i^2}{\sum(Y_i-\bar{Y})^2} = 1 - \frac{0.5781}{2.5844} \approx 0.1632$ //

The r^2 shows the goodness of fit of how well the data fits with the regression. In this case, $r^2 = 0.1632$, this means that 16.32 of Y fits by the regression model

(1c) $X_i = 30$ $\hat{Y}_i = ?$

$\hat{Y}_i = 0.6866 + 0.0975X_i$

$\hat{Y}_i = 0.6866 + 0.0975(30) = 3.6116$ //

This means that when X_i is 30 the $\hat{Y}_i$ will be 3.116

(1d) $var(u_i)$ $var(\hat{\beta}_1)$ $var(\hat{\beta}_2)$

$var(u_i) = \frac{\sum\hat{u}_i^2}{n-k} = \frac{0.5781}{18-2} = 0.0361 \approx \hat{\sigma}^2$ //

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

$$var(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2} = \frac{0.0361}{211} \approx 0.00017 \approx 0.0002$$

(1e) $P\left[\hat{\beta}_2 - t_{\frac{a}{2}} \leq \beta_2 \leq \hat{\beta}_2 + t_{\frac{a}{2}}\right] = 1-a$

$a = 0.1$ $\quad 1-a = 0.90$ CI or 90%

$t_{\frac{a}{2}} = t_{\frac{0.1}{2}} = t_{0.05} = 1.7459$

$df = n-k = 18-2 = 16$

$\hat{\beta}_2 = 0.0975$

$se_{\hat{\beta}_2} = \sqrt{var\,\hat{\beta}_2} = \sqrt{0.00017}$

$= 0.0131$

$P\left[0.0975 - 1.7459(0.0131) \leq \beta_2 \leq 0.0975 + 1.7459(0.0131)\right]$

$P\left[0.0746 \leq \beta_2 \leq 0.1204\right]$

This means that 90% of the time the true β_2 will be in the interval of 0.0746 to 0.1204.

(1f) Hypothesis test: β_2

$H_0 = 0$ null hypothesis

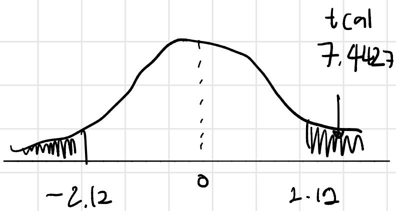
$H_a \neq 0$ alternative hypothesis

$a = 0.05$ $\quad 1-a = 0.95$ (95% CI)

$$t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{se\,\hat{\beta}_2} = \frac{0.0975 - 0}{0.0131} \approx 7.4427$$

$$t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025} \qquad df = \underset{n-k}{18-2} = 16$$

$t_{0.025} = \pm 2.120 \Rightarrow$ critical values



In this case tcal > critical value therefore, we reject the null hypothesis. This means that 95% of the time beta (β_2) will not be zero.

Hypothesis test : β_1

$H_0 = 0$ null $\qquad \alpha = 0.05$

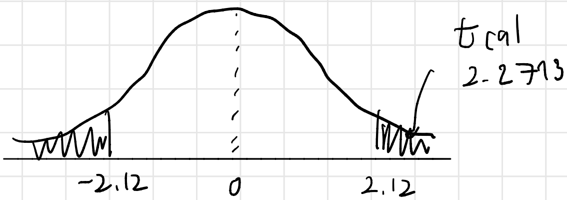
$1-\alpha = 95\%$ CI

$H_a \neq 0$ alternate

$$se(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = \sqrt{0.0914} = 0.3023$$

$$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se\,\hat{\beta}_1} = \frac{0.6866 - 0}{0.3023} \approx 2.2713$$

$$t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025} = \pm 2.120$$



Since tcal > critical value we reject the null hypothesis. This means that 95% of the time the true β_1 will not be zero.

(2a)

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

Y_i outp	Coefficient	Std. err. (SE)	t	P>|t|	[95% conf.	interval]
$\hat{\beta}_2$ age	.0031338	.0002292	Omitted		.0026846	.003583
$\hat{\beta}_1$ _cons	.4279898	.0140339			.4004828	.4554969

Hypothesis test (β_1)

$H_0 : \beta_1 = 0$ null

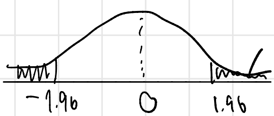
$H_a : \beta_1 \neq 0$ alternative

$\alpha = 0.05$

$1-\alpha = 0.95 = 95\%$ CI

$$t_{cal} : \frac{\hat{\beta}_1 - \beta_1}{se_{\hat{\beta}_1}} = \frac{0.4279898 - 0}{0.0140339} \approx 30.4969$$

$df = \infty \quad t_{\frac{\alpha}{2}} = t_{0.025} = 1.960$



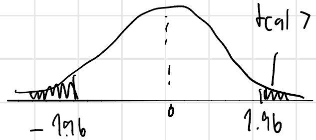
tcal > critical value

$H_0 : \beta_2 = 0$

$H_a \ \beta_2 \neq 0$

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

$t_{0.025} = 1.960$



tcal > critica valu

For both hypothesis test, the tvalue is more than the critical value
This means that we can be 95% confident β_1 & β_2 are not zero

(2b) $\hat{\beta}_2 = 0.0031888$ This means that if the age increases by 1 year the number of hospital visits will increase by 0.0031888 times.
It make sense because as people age, there are more diseases affected and is hard to recover by themselves, therefore, more aged more they go on hospital visits.

(2c) If there is a increase by 1 unit of year (age) it is associated with an income increase of the transformed $\hat{\beta}_2$ percent multiply by 100 in increase of hospital visits which is 0.3134%.

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

2d If the age variable is divided by 10 the confidence interval, the coefficient, standard error will be scaled by 10.

output	C	SE	95% CI	Interval
age	0.031338	0.002292	0.02684	0.03583

2e $\hat{Y}_0 = 0.4279898 + 0.0031338(50) = 0.5847$

$se(\hat{Y}_0) = \sqrt{var(\hat{Y}_0)} = 0.0045$

$t_{0.005} = 2.576$ $\qquad \alpha = 0.01 \quad \frac{\alpha}{2} = 0.005$

$$P\left[0.5847 - (2.576 \cdot 0.0045) \leq Y_0 \leq 0.5847 + (2.576 \cdot 0.0045)\right] \quad (1-\alpha)$$

$$P\left[0.5731 \leq Y_0 \leq 0.5963\right] = 0.99$$

3 As X_0 is getting further away from $\bar{x}$, the variance will be larger as stated from variance formula. The variance is used to calculate the standard error $\hat{Y}_0$. When the X_0 is further away from $\bar{x}$, the data would be more scattered with the large variance. The confidence interval must be larger too because the data are scattered away from $\bar{x}$.