

$$1a) \hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{20.58}{211} = 0.0975$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = 2.8278 - 0.0975(21.5556) = 2.8278 - 2.1017 = 0.7261$$

SRF model is $\hat{y}_i = 0.7261 + 0.0975 x_i$ meaning that the intercept is 0.7261 and the slope of this function is 0.0975

$$1b) r^2 = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{0.5781}{2.5849} = 1 - 0.2237 = 0.7763$$

r^2 is the measure of goodness to fit, in this case 77.63% can be explained by the regression model

$$1c) \text{ From } \hat{y}_i = 0.7261 + 0.0975 x_i$$

Plugging $x_i = 30$ in

$$\hat{y}_i = 0.7261 + 0.0975(30)$$

$$\hat{y}_i = 0.7261 + 2.925$$

$$\hat{y}_i = 3.6511$$

When $x_i = 30$ the average of $\hat{y}_i$ will be 3.6511

$$1d) \text{ Var}(u_i) = \frac{\sum_{i=1}^n u_i^2}{n-k} = \frac{0.5781}{18-2} = \frac{0.5781}{16} = 0.0361 = \sigma^2$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i)^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2 = \frac{9620}{118721} (0.0361) = 2.5329 (0.0361) = 0.0914$$

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{0.0361}{211} = 0.0002$$

1e) Find CI - Using t-test because σ^2 is unknown

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

90% CI ; $\alpha = 0.1$

$$t_{\frac{\alpha}{2}} = t_{\frac{0.1}{2}} = t_{0.05, 16} = 1.746$$

$$\text{Upper limit} = \hat{\beta}_2 + t_{\frac{\alpha}{2}} \cdot se \hat{\beta}_2 = 0.0975 + (1.746 \cdot 0.0141) = 0.0975 + 0.0246 = 0.1221$$

$$\text{Lower limit} = \hat{\beta}_2 - t_{\frac{\alpha}{2}} \cdot se \hat{\beta}_2 = 0.0975 - (1.746 \cdot 0.0141) = 0.0975 - 0.0246 = 0.0729$$

90% CI of $\hat{\beta}_2$ is $[0.0729, 0.1221]$.

It means that there are 90% probability that $\hat{\beta}_2$ will be in the range between 0.0729 and 0.1221

1f) Test β_2 ;

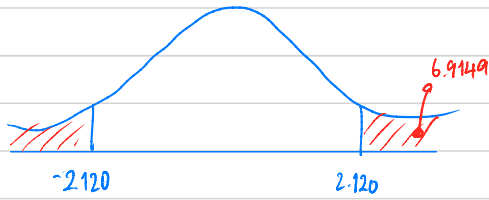
Assume that

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

$$\text{Compute } t_{\text{cal}} = \frac{\hat{\beta}_2 - \beta_2}{\text{se}(\hat{\beta}_2)} = \frac{0.0975 - 0}{0.0141} = 6.9149$$

Find CV when $\alpha = 0.05$, $t_{\frac{\alpha}{2}} = t_{0.025, 16} = \pm 2.120$



T_{cal} is in the rejection region, therefore reject H_0

We can say that β_2 is not zero with 95% CI.

2a) Test β_1

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$\alpha = 0.05,$$

$$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} = \frac{0.4280 - 0}{0.0140} = 30.5714$$

find CV at $\alpha = 0.05, n-k = 27,884$,

$$t_{0.025, 27,884} = \pm 1.960$$

t_{cal} is more than $t_{0.025, 27,884}$, Reject H_0 .

β_1 is not equal to zero with 95% CI

Test β_2

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

$$\text{Compute } t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{\text{se}(\hat{\beta}_2)} = \frac{0.0313 - 0}{0.0023} = 13.6087$$

$$\text{Find CR}; t_{\frac{\alpha}{2}, n-k} = t_{0.025, 27,884} = \pm 1.960$$

t_{cal} lies beyond the CR, reject H_0

We say that β_2 is not zero with 95% CI

2b) Yes, because higher in age it means that there is higher probability to get sick or going more to hospital

2C) When out_t changes to $\ln$ scale, we can reinterpret that when the age increase by 1 unit, the output will increase by

$$100 \cdot \hat{\beta}_2 \text{ percent [Approximately]}, (100 \cdot e^{\hat{\beta}_2}) - 100 \text{ percent [Exact measure]}$$

2d) When age_i divide by 10, There is no difference when we interpret the coefficients

2e) Mean Prediction $X_0 = 50$

$$\hat{y}_0 = 0.0031 + 0.4260(50) = 0.0031 + 21.4 = 21.4031$$

$$se_{\hat{y}_0} = \sqrt{\text{Var}(\hat{y}_0)} = \sqrt{0.0002} = 0.0045$$

Find 99% CI ($t_{\frac{\alpha}{2}} = 2.576$) ($\alpha=0.01, 1-\alpha=0.99$)

$$\Pr [21.4031 - (2.576(0.0045)) \leq y_0 \leq 21.4031 + (2.576(0.0045))] = 0.99$$

$$\Pr [21.4031 - 0.0116 \leq y_0 \leq 21.4031 + 0.0116] = 0.99$$

$$\Pr [21.3915 \leq y_0 \leq 21.4147] = 0.99$$

3) The reason is due to when the further away from the mean is, the larger of the variance will be. As explained by the formula of both mean and individual prediction. As both formula utilised Variance as the part in the calculation the further it is, the more error it can be.