STAT 3A03 Applied Regression With SAS
Fall 2017

Term Test 1 Solution Set

Q. 1 **a)** The sum of squared errors is

$$S(\beta_1) = \sum_{i=1}^{n}(y_i - \beta_1 x_i)^2$$

[2 marks]

To minimize this we take the derivative with respect to β_1

$$\frac{dS(\beta_1)}{d\beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_1 x_i)$$

[2 marks]

Now we set that equal to 0 when $\beta_1 = \hat{\beta}_1$ and solve.

$$\begin{aligned}
\left.\frac{dS(\beta_1)}{d\beta_1}\right|_{\beta_1=\hat{\beta}_1} = 0 \quad &\Rightarrow \quad \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_1 x_i) = 0 \\
&\Rightarrow \quad \sum_{i=1}^{n} x_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \\
&\Rightarrow \quad \hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}
\end{aligned}$$

[4 marks]

b) First we note that we can write the estimator

$$\hat{\beta}_1 = \sum_{i=1}^{n}\left(\frac{x_i}{\sum x_j^2}\right) Y_i$$

Hence we can write

$$\mathrm{E}\left(\hat{\beta}_1 \mid x_1, \ldots, x_n\right) = \sum_{i=1}^{n}\left(\frac{x_i}{\sum x_j^2}\right) \mathrm{E}\left(Y_i \mid x_i\right)$$

[2 marks]

Next we note that

$$\mathrm{E}\left(Y_i \mid x_i\right) = \beta_1 x_i + \mathrm{E}\left(\varepsilon_i \mid x_i\right) = \beta_1 x_i$$

[2 marks]

Hence we have

$$\begin{aligned} \mathrm{E}(\hat{\beta}_1 \mid x_1, \ldots, x_n) &= \sum_{i=1}^{n} \left(\frac{x_i}{\sum x_j^2} \right) \beta_1 x_i \\ &= \beta_i \frac{\sum x_i^2}{\sum x_j^2} \\ &= \beta_1 \end{aligned}$$

[2 marks]

c) From the form of $\hat{\beta}_1$ given above we have

$$\mathrm{Var}\left(\hat{\beta}_1 \mid x_1, \ldots, x_n\right) = \sum_{i=1}^{n} \left(\frac{x_i}{\sum x_j^2} \right)^2 \mathrm{Var}\left(Y_i \mid x_i\right)$$

[2 marks]

The assumption in the question gives us that

$$\mathrm{Var}\left(Y_i \mid x_i\right) = \mathrm{Var}\left(\varepsilon_i \mid x_i\right) = \sigma^2$$

[2 marks]

Hence we get

$$\begin{aligned} \mathrm{Var}\left(\hat{\beta}_1 \mid x_1, \ldots, x_n\right) &= \sum_{i=1}^{n} \left(\frac{x_i}{\sum x_j^2} \right)^2 \sigma^2 \\ &= \frac{\sum x_i^2}{\left(\sum x_j^2\right)^2} \sigma^2 \\ &= \frac{\sigma^2}{\sum x_i^2} \end{aligned}$$

[2 marks]

Q. 2 **a)**

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{274.786}{\sqrt{472.076 \times 731.961}} \\ &= 0.4675 \end{aligned}$$

[4 marks]

b)

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\ &= \frac{274.786}{472.076} \\ &= 0.5821 \end{aligned}$$

[3 marks]

$$\begin{aligned} \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\ &= 59.47 - 0.5821 \times 165.52 \\ &= -36.8759 \end{aligned}$$

[3 marks]

c)

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \overline{y} + \hat{\beta}_1 \overline{x} - \hat{\beta}_1 x_i \right)^2 \\ &= \sum_{i=1}^{n} \left((y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x}) \right)^2 \\ &= \sum_{i=1}^{n} (y_i - \overline{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) + \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= S_{yy} - 2 \left(\frac{S_{xy}}{S_{xx}} \right) S_{xy} + \left(\frac{S_{xy}}{S_{xx}} \right)^2 S_{xx} \\ &= S_{yy} \left(1 - \frac{S_{xy}^2}{S_{xx} S_{yy}} \right) \\ &= S_{yy} \left(1 - r^2 \right) \end{aligned}$$

[6 marks]

Hence we have

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\text{SSE}}{n-2} \\ &= \frac{S_{yy}(1-r^2)}{n-2} \\ &= \frac{731.961(1-0.4675^2)}{8} \\ &= 71.5017\end{aligned}$$

[2 marks]

d) The proportion of variability in weight which can be explained by the model is the coefficient of determination and that is equal to the square of the correlation coefficient between height and weight.

$$R^2 = r^2 = 0.4675^2 = 0.2185$$

[2 marks]

Q. 3 a)

$$\begin{aligned} a &= p = 4 \\ b &= n - p - 1 = 25 - 5 = 20 \\ c &= n - 1 = 24 \\ d &= 2853.05 \times a = 2853.05 \times 4 = 11412.2 \\ e &= 26833 - d = 26833 - 11412.2 = 15420.8 \\ f &= e/b = 15420.8/20 = 771.04 \\ g &= 2853.05/f = 2853.05/771.04 = 3.70 \end{aligned}$$

[7 marks]

Here is the completed ANOVA table.

	DF	Sum of Squares	Mean Square	F Value
Model	4	11412.2	2853.05	3.70
Source	20	15420.8	771.04	
Corrected Total	24	26833		

b) The F Value is the test statistic to test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad \text{V} \quad H_1 : \text{At least one of } \beta_1, \beta_2, \beta_3, \beta_4 \text{ is non-zero.}$$

[2 marks]

The 5% critical value for the $F_{4,20}$ distribution is 2.87 (from the table) and since the observed F Value is $F = 3.70 > 2.87$ we can reject H_0 at the 5% level. **[2 marks]**

c) This does not contradict the results of the t tests in the Parameter Estimates table since those are testing each covariate individually assuming that the other three covariates are in the model whereas the F-test is testing if **all** of the covariates can be removed from the model **simultaneously**. It is quite possible to reject the F-test and not reject any of the individual tests as happens in this example. **[4 marks]**

d) The test statistic to test if we can remove both height and weight from the model is the F-test comparing the reduced model (without height or weight) to the full model (with all 4 covariates).

$$\begin{aligned} F &= \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(\text{df}_{\text{red}} - \text{df}_{\text{full}})}{\text{MSE}_{\text{full}}} \\ &= \frac{(16345 - 15420.8)/(22 - 20)}{771.04} \\ &= \frac{462.1}{771.04} \\ &= 0.5993 \end{aligned}$$

[2 marks]

If the reduced model is the true model then F has an F distribution with 2 and 20 degrees of freedom.

Using the $F_{(2,20;0.05)} = 3.49$ critical value we see that $F < 3.49$ and so the p-value is $p > 0.05$. Hence we will not reject the null hypothesis and can conclude that height and weight are not needed in the model when age and tlc are in the model. **[3 marks]**