STAT 3A03 Applied Regression With SAS
Fall 2017

Term Test 2 Solution Set

Q. 1 **a)** The assumptions of the linear model are

Linearity The true model should be linear in the parameters

$$\boldsymbol{Y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

[1 mark]

Homoscedasticity The errors $\boldsymbol{\varepsilon}$ have mean 0 and the same variance, σ^2, for all values of the covariate vector. **[1 mark]**

Normality The errors $\boldsymbol{\varepsilon}$ are independently and normally distributed. **[1 mark]**

We assess violations of these assumptions as follows:

Linearity Look for deviations from linearity in the plot of the fitted values $\hat{y}_i$ against the observed values y_i $(i = 1, \ldots, n)$ and/or in the added variable plots. **[2 marks]**

Homoscedasticity Examine the plot of the studentized residuals r_i against the fitted values $\hat{y}_i$. This should be a random scatter about 0, major differences in vertical spread across the plot indicate a violation of this assumption. **[2 marks]**

Normality Examine a normal quantile-quantile plot of the studentized residuals. This should be a straight line, any marked curvature in the plot indicates a violation of this assumption. **[2 marks]**

b) **(i)** An outlier is a point which deviates from the linear model that describes the rest of the dataset. Such points will usually have a large studentized residual in absolute value $(|r_i| > 2)$. **[3 marks]**

(ii) A highly influential point is one whose inclusion in the model changes the fitted model significantly. Such points usually have a large leverage or they can be found using jackknifing in which the fitted models with and without including the point are compared. **[3 marks]**

c) From the Delta method we have

$$\mathrm{E}\big(\sqrt{Y} \mid X\big) \;\approx\; \sqrt{\mathrm{E}(Y \mid X)} \;=\; \sqrt{\beta_1}\,\sqrt{X}$$

$$\mathrm{Var}\big(\sqrt{Y} \mid X\big) \;\approx\; \left(\frac{1}{2\sqrt{\mathrm{E}(Y \mid X)}}\right)^2 \mathrm{Var}(Y \mid X) \;=\; \left(\frac{1}{2\sqrt{\beta_1 X}}\right)^2 X\sigma^2 \;=\; \frac{\sigma^2}{4\beta_1}$$

[4 marks]

Hence if we set $Y' = \sqrt{Y}$, $X' = \sqrt{X}$, $\beta_1' = \sqrt{\beta_1}$ and $\sigma_1^2 = \sigma^2/(4\beta_1)$ then we can write

$$Y' \;=\; \beta_1' X' + \varepsilon$$

where $\mathrm{Var}(\varepsilon) \;=\; \mathrm{Var}(Y' \mid X') \;=\; \sigma_1^2$ which is a linear model with constant variance as required. **[1 mark]**

Q. 2 **a)** β_0 is the expected blood coagulation time for animals eating Diet A. **[2 marks]**

$\hat{\beta}_0 = 61$ **[1 mark]**

β_1 is the difference in expected blood coagulation times between animals eating Diet B and those eating Diet A. **[2 marks]**

$\hat{\beta}_1 = 5.0$ **[1 mark]**

σ^2 is the variance in blood coagulation time among all animals eating the same diet. **[2 marks]**

$\hat{\sigma}^2 = 5.6$ **[1 mark]**

b) The null hypothesis being tested is $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ against the alternative that one of these is not 0. **[1 mark]**

The interpretation of this test is that it is testing if expected blood coagulation time is identical for all four tested diets against the alternative that there is some difference in expected blood coagulation time across diets. **[2 marks]**

Since $p < 0.0001$ we reject the null hypothesis and conclude that there is very strong evidence that the expected blood coagulation time differs across these diets. **[2 marks]**

c) This model has only one covariate which is defined by

$$\text{DietBC}_i = \begin{cases} 1 & \text{if animal } i \text{ is given Diet B or Diet C} \\ 0 & \text{if animal } i \text{ is given Diet A or Diet D} \end{cases}$$

so the fitted model assumes that there is no difference in blood coagulation time between animals fed diets B or C and also that there is no difference between animals fed diets A or D although there may be a difference in blood coagulation time between these two sets of animals. It is a reduced model relative to the full model because it adds the restrictions

$$\beta_1 = \beta_2, \qquad \beta_3 = 0$$

to the original model. **[2 marks]**

A test of this reduced model is done by calculating

$$\begin{aligned} F &= \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(df_{\text{red}} - df_{\text{full}})}{\text{MSE}_{\text{full}}} \\ &= \frac{(124 - 112)/(22 - 20)}{5.6} \\ &= 1.07 \end{aligned}$$

[2 marks]

Since $F = 1.07 < 3.49 = F(0.05; 2, 20)$ there is insufficient evidence to reject the null hypothesis and so we conclude that the reduced model is sufficient to explain the data. **[2 marks]**

Q. 3 **a)** Let P denote price and let HP denote horsepower then the fitted models are:

Germany: $\hat{P} = 0.893 + 0.142HP$ **[2 marks]**

Japan: $\hat{P} = -6.127 + 0.161HP$ **[2 marks]**

USA: $\hat{P} = -8.806 + 0.185HP$ **[2 marks]**

Other: $\hat{P} = -10.882 + 0.237HP$ **[2 marks]**

b) Let $\alpha_1, \alpha_2, \alpha_3$ be the coefficients of the product of the dummy variables for country and the horsepower. Then the null hypothesis is $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$. In other words we need to test if there is a significant interaction effect on price between horsepower and country of origin.

[2 marks]

We can test this by looking at the Type III sum of squares from which we get $F = 1.46$ and this should have be distributed according to an $F_{3,82}$ distribution if the parallel lines model is sufficient. **[2 marks]**

Since we have p-value of $p = 0.2204$ we conclude that there is insufficient evidence against H_0 and so it appears that the parallel lines model is sufficient to explain the variability seen in the dataset. **[2 marks]**

c) To test this assumption we would need to fit a new model which only included horsepower as a covariate. Let SSE_2 denote the error sum of squares from this model then we would need to calculate

$$F = \frac{(\text{SSE}_2 - 1319.8460)/6}{16.0957}$$

[4 marks]

If the model without country of origin is suffient to describe the data then this should be an observation from an $F_{6,82}$ distribution and so we could use SAS to calculate a p-value or we could compare the observed value with $F(0.05; 6, 82) < F(0.05; 6, 60) = 2.25$ and if the observed value is greater than 2.25 then we reject the null hypothesis. **[2 marks]**