STAT 3A03 Applied Regression With SAS

Assignment 5 Due at 5pm on Monday December 4, 2017

Dropboxes for assignment submission are outside HH-105. Your assignment **MUST** be deposited in the appropriate dropbox for your lab section.

N.B. Late assignments will not be accepted

Q. 1 The dataset in the file `Wool.txt` come from an experiment to understand the strength of wool under load stress as a function of three factors – The length of the test specimen, the amplitude of the loading cycle to which the wool was subjected and the amount of the load. Each of these factors were under control of the experimenter and three settings for each factor was used in the experiment. In this question, we shall not be interested in the actual values used but will assume that each is a categorical (class) variable with three categories. The response was the number of loading cycles before the specimen failed.

a) Fit a linear model with just main effects for the three class variables and show that this is not a good fit to the data.

b) Add in all interactions between pairs of the class variables and refit the model. Summarize the results of this model. Does it fit better than the model in (a)?

c) Using the model without interactions, look for a transformation of the response which gives a good fit.

d) Show that, in the transformed scale, the model with the interactions is no better than the model with main effects only.

Q. 2 Consider the oil production data in your textbook (Table 6.17 in the 4th edition and 6.18 in the 5th edition). The data are in the file `Oil.txt` on the website.

a) Draw a scatterplot of barrels of oil produced against year. Does a linear relationship seem reasonable? Repeat this using the log of barrels.

b) Fit a linear regression using Year as covariate and log(Barrels) as the response. Conduct standard regression diagnostics and comment on any problems with the model.

c) The pattern of oil production changed after the 1973 oil crisis. Construct an indicator variable that indicates if the year is later than 1973 or not. Extend the model from part (b) to one that allows for intercepts and slopes in the two time periods. Give the estimates of the two lines and your conclusions from the analysis. Are there any issues with the assumptions underlying this model?

Q. 3 Consider the weighted simple linear regression with responses $y_1, \ldots, y_n$, covariate values $x_1, \ldots, x_n$ and known weights $w_1, \ldots, w_n$.

a) Define the weighted means of x and y to be

$$\overline{x}_w = \frac{\sum w_i x_i}{\sum w_i} \qquad \overline{y}_w = \frac{\sum w_i y_i}{\sum w_i}.$$

Show that $\sum_{i=1}^{n} w_i(x_i - \overline{x}_w) = 0$, and hence that $\sum_{i=1}^{n} w_i x_i(x_i - \overline{x}_w) = \sum_{i=1}^{n} w_i(x_i - \overline{x}_w)^2$ and that $\sum_{i=1}^{n} w_i x_i(y_i - \overline{y}_w) = \sum_{i=1}^{n} w_i(x_i - \overline{x}_w)(y_i - \overline{y}_w)$

b) Derive the weighted least squares estimates in the weighted simple linear regression setting:

$$\begin{aligned} \hat{\beta}_0 &= \overline{y}_w - \hat{\beta}_1 \overline{x}_w \\ \hat{\beta}_1 &= \frac{\sum w_i(x_i - \overline{x}_w)(y_i - \overline{y}_w)}{\sum w_i(x_i - \overline{x}_w)^2} \end{aligned}$$

c) Show that these estimators can be written as

$$\begin{aligned} \hat{\beta}_1 &= \sum_{i=1}^{n} \frac{w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} Y_i \\ \hat{\beta}_0 &= \sum_{i=1}^{n} \left(\frac{w_i}{\sum w_j} - \frac{w_i \overline{x}_w (x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} \right) Y_i \end{aligned}$$

and hence show that the estimators are unbiased.

Q. 4 The data in `Chromatography.txt` was used to calibrate a gas chromatograph. Five measurements were taken from each of four specimens with different amounts of a substance. The amount of substance in each specimen was carefully determined before the experiment.

a) Fit a linear model with the log of chromatograph output as response and the log of the amount of substance as the predictor and verify that there is a problem with heteroscedasticity.

b) Construct appropriate weights for the observations based on the residual variances at different amounts.

c) Using the weights you determined in (b) fit a weighted regression. Write a report explaining your analysis and the results making sure to comment on the appropriateness of the model assumptions.