STAT 3008
Exercise 9

Problems refer to the problem sets in the textbook: Applied Linear Regression, 3rd edition by Weisberg.

1. Problem 7.1

   The data in the file `baeskel.txt` were collected in a study of the effect of dissolved sulfur on the surface tension of liquid copper (Baes and Kellogg, 1953). The predictor Sulfur is the weight percent sulfur, and the response is Tension, the decrease in surface tension in dynes per cm. Two replicate observations were taken at each value of *Sulfur*. These data were previously discussed by Sclove (1968).

   (a) Draw the plot of *Tension* versus *Sulfur* to verify that a transformation is required to achieve a straight-line mean function.

   (b) Set $\lambda = -1$, and fit the mean function

   $$E(Tension|Sulfur) = \beta_0 + \beta_1 Sulfur^{\lambda}$$

   using OLS; that is, fit the OLS regression with *Tension* as the response and $1/Sulfur$ as the predictor. Let *new* be a vector of 100 equally spaced values between the minimum value of *Sulfur* and its maximum value. Compute the fitted values from the regression you just fit, given by $Fit.new = \hat{\beta}_0 + \hat{\beta}_1 new^{\lambda}$. Then, add to the graph you drew in Problem(a) the line joining the points (*new*, *Fit.new*). Repeat for $\lambda = 0, 1$. Which of these three choices of λ gives fitted values that match the data most closely?

   (c) Replace *Sulfur* by its logarithm, and consider transforming the response *Tension*. To do this, draw the inverse fitted value plot with the fitted values from the regression of *Tension* on $loglog(Sulphur)$ on the vertical axis and *Tension* on the horizontal axis. Repeat the methodology of Problem(b) to decide if further transformation of the response will be helpful.

2. Problem 7.2. (For 7.2.3, only compare two models qualitatively by looking at the fitted curve. F-test cannot compare two models with different weight.)

The (hypothetical) data in the file `stopping.txt` give stopping times for $n = 62$ trials of various automobiles traveling at *Speed* miles per hour and the resulting stopping *Distance* in feet (Ezekiel and Fox, 1959).

(a) Draw the scatterplot of *Distance* versus *Speed*. Add the simple regression mean function to your plot. What problems are apparent? Compute a test for lack of fit, and summarize results.

(b) Find an appropriate transformation for *Distance* that can linearize this regression.

(c) Hald (1960) has suggested on the basis of a theoretical argument that the mean function $E(Distance|Speed) = \beta_0 + \beta_1 Speed + \beta_2 Speed^2$, with $Var(Distance|Speed) = \sigma^2 Speed^2$ is appropriate for data of this type. Compare the fit of this model to the model found in Problem(b). For *Speed* in the range 0 to 40 mph, draw the curves that give the predicted *Distance* from each model, and qualitatively compare them.

3. The data file `UN3.txt` contains data described in Table 7.5 of the Textbook. There are data for $n = 125$ localities, mostly UN member countries, for which values are observed for all the variables recorded.

   Consider the regression problem with *ModernC* as the response variable and the other variables in the file as defining terms.

   (a) Use Box-Cox transformation to select a model (transform response only). Select among $\lambda = (-1, -1/2, 0, 1/3, 1/2, 1, 2)$.

   (b) Use modified power transformation to select a model (transform predictors only). (There are a lot of possibilities, just target at one transformation that you feel reasonable.)

   (c) Compute the RSS for the best fit models in i) and ii). Which method gives a smaller RSS?

   Caution: Some predictors are negative!