Homework 6

Due Tuesday, November 5

Show all your work. Data files available from `alr4` package or previous homework problems.

1. Consider a simple weighted regression model, $y_i = \beta_0 + \beta_1 x_i + e_i$, for $i = 1, \ldots, n$, where the errors e_i are independent with $E(e_i) = 0$ and $\text{Var}(e_i) = \sigma^2 w_i^{-1}$, and $w_i > 0$ known. Starting with $RSS(\beta_0, \beta_1) = \sum w_i (y_i - \beta_0 - \beta_1 x_i)^2$, and setting both $\frac{\partial}{\partial \beta_0} RSS(\beta_0, \beta_1) = 0$ and $\frac{\partial}{\partial \beta_1} RSS(\beta_0, \beta_1) = 0$, show that

$$\hat{\beta}_1 = \frac{\sum w_i x_i (y_i - \bar{y}_w)}{\sum w_i x_i (x_i - \bar{x}_w)} = \frac{\sum w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum w_i (x_i - \bar{x}_w)^2}$$

and

$$\hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w$$

where

$$\bar{y}_w = \frac{\sum w_i y_i}{\sum w_i} \quad \text{and} \quad \bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

2. Consider a model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, with $E(\mathbf{e}) = \mathbf{0}$ and $\text{Var}(\mathbf{e}) = \sigma^2 \mathbf{W}^{-1}$, where $\mathbf{W}$ is a diagonal matrix of weights with known positive entries. We have already established that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$ (see page 157). From this, derive $E(\hat{\boldsymbol{\beta}})$ and $\text{Var}(\hat{\boldsymbol{\beta}})$.

3. Refer to HW 4, Problem 1. If we have a model of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, with $E(\mathbf{e}) = \mathbf{0}$ and $\text{Var}(\mathbf{e}) = \sigma^2 \mathbf{W}^{-1}$, where $\mathbf{W} = \text{diag}(4, 3, 1, 2, 5)$, then find $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$ and $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$, using matrix computation in R (check with `lm()` with `weights=c(4,3,1,2,5)`).

4. Consider the dataset `galtonpeas`.
   (a) Draw the scatterplot of *Progeny* (Y) versus *Parent* (X).
   (b) Compute the weighted regression of *Progeny* on *Parent*, with the *SD* as the weight (the same way as the "Physics" example in lecture). Add a regression line to the plot in part (a). Use `lm()` function in R and report the summary.
   (c) Verify the numbers in part (b) by matrix computation in R.
   (d) Provide a residual plot.
   (e) Interpret the results and comment on any interesting observations.

5. Refer to HW 4, Problem 3.
   (a) Produce a residual plot of the (original) model and comment.
   (b) Fit a weighted regression model, with weights = 1/(Hour+1), using `lm()` function in R. Report the summary and produce a residual plot of this model.
   (c) Comment and interpret your results, including any comparison with HW 4, Problem 3.

6. Consider the dataset `baeskel`. The predictor *Sulfur* (X) is the weight percent sulfur, and the response is *Tension* (Y), the decrease in surface tension in dynes per cm.

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

   (b) Fit the OLS regression with *Tension* as the response and *1/Sulfur* as the predictor. Report the summary and provide a residual plot. Comment.

   (c) Replace *Sulfur* by its logarithm. Report the summary and provide a residual plot. Comment.

   (d) Starting with part (c), consider transforming the response *Tension*. Try different transformations on *Tension* and see if it makes any difference. (You can limit your transformations to log, inverse, square root, and power Y^b). Report any appropriate or interesting findings.

7. The dataset `stopping` give stopping times for $n = 62$ trials of various automobiles traveling at Speed (X) miles per hour and the resulting stopping Distance (Y) in feet.

   (a) Draw the scatterplot of Distance versus Speed. Add the simple regression mean function to your plot. What problems are apparent?

   (b) Find an appropriate transformation for Distance that can improve this regression. As usual, please report appropriate results/plots and interpret/comment on your findings.

8. Consider the `pipeline` dataset. The goal is to decide if the field measurement can be used to predict the more accurate lab measurement.

   (a) Draw the scatterplot of *Lab* versus *Field*, and comment on the applicability of the simple linear regression model.

   (b) Fit the simple regression model, and get the residual plot. Compute the score test for nonconstant variance, and summarize your results.

   (c) If the result suggests nonconstant variance, please run the analysis again with transformed data (your choice of which transformation to use and whether to transform Y and/or X).

9. Consider the `caution` dataset. Find an appropriate transformation for Y that can improve this regression. As usual, please report appropriate results/plots and interpret/comment on your findings.