STAT 420 – Homework 8

1. Productivity Data

An operations analyst in an electronics firm studied factors affecting production in a piecework operation where earnings are based on the number of pieces produced. Two employees each were selected from various age groups, and the data on their productivity last year was obtained. Note that *X* is the age of employee, in years; *Y* is employee's productivity. The dataset `productivity` contains the values.

a. Create a scatterplot of the data. Fit a simple linear model with *Y* as the response and add the best-fit line to the plot. Does a simple linear model seem appropriate here?

b. Fit a quadratic model with *Y* as the response. Does this model provide a better fit? Explain. Use a level of $\alpha = 0.05$.

c. Add the best fit parabola to the plot from part a.

d. Create a new vector with centralized values for the *X* variable. Create a brand new scatterplot of the data. Fit a new simple linear model and quadratic model and overlay their lines on the new scatterplot. Does the quadratic model with centralized *X* values with *Y* as the response provide a better fit than the model of part b?

e. Fit a cubic (third-order) model with *Y* as the response. Does this model provide a better fit than the one in part d? Use a level of $\alpha = 0.05$.

f. Add the best fit third-order line from the model in part e to the scatterplot.

2. Admissions Data

An educational foundation awards scholarships to high school graduates to help them with expenses during their first year at a major university. The foundation would like to consider a student for scholarship only if that student will earn a GPA of at least 2.80 (on a 4.0-point scale) during his/her first year at the university. Since the scholarship is awarded *before* the student enters the university, the first-year GPA must be *predicted.* Each applicant for a scholarship must take an achievement test, the result of which is used to predict his/her first-year GPA. The director of the foundation also wants to determine whether knowing which major one wants to pursue has an effect (0 means major not known, 1 means major known). The director has access to the records of the 2,000 students who applied for scholarships during the past five years and who completed their first year of college. A simple random sample of 20 students was selected from these records. Their scores on the achievement test and their first-year GPAs are in the `admissions` data set.

a. Plot GPA vs. Score using different symbols for those who know their major and those who don't. Comment on whether Major should be included as a covariate for predicting GPA.

b. Fit a model with GPA as the response and the other two variables as the predictors as main effects only (that is, do not include an interaction term). What proportion of the variation in GPA is explained by this model?

c. Give an practical interpretation of what the coefficient for the Score term means, and an interpretation of what the coefficient for the Major term means.

d. Fit a model with GPA as the response and the other two variables as the predictors as main effects including an interaction term. What proportion of the variation in GPA is explained by this model?

e. Run an *F*-test to determine which of the two models in parts b and d is preferable. Use 0.05 as the level of significance.

3. Prostate Data

For the `prostate` data, fit a model with `lpsa` as the response and the other variables as predictors.

```
> library(faraway)
> data(longley)
```

a. Implement backward elimination variable selection method to determine the "best" model. Use $\alpha = 0.10$.

b. Implement backward AIC variable selection to determine the "best" model.

c. Compare the values of Adjusted R^2 for the full model, the "best" model from part a, and the "best" model from part b. Which model is the "best" model out of the three? Explain.