Spring 16 – AMS256 Homework 6

1. (Faraway 2005) For the prostate data (available online – or use data in R (http://cran.r-project.org/web/packages/faraway/faraway.pdf)), fit a model with `lpsa` as the response and the other variables as predictors.

   (a) Compute 90% and 95% CIs for the parameter associated with age. Based on these intervals, what could you have deduced about the p-value for age in the regression summary?

   (b) Compute and display a 95% joint confidence region for the parameters associated with age and lbph. Plot the origin on this display. The location of the origin on the display tells us the outcome of a certain hypothesis test. State that test and its outcome.

   (c) Suppose a new patient arrives with the following values:
   `lcavol=1.447, lweight=3.623, age=65, lbph=0.3, svi=0.0, lcp=-0.799, gleason=7, pgg45=15`
   Predict the `lpsa` for this patient with an appropriate 95% CI.

   (d) Repeat the last question for a patient with the same values, except that he/she is age 20. Why is the CI wider?

   (e) Now, remove all the predictors that are not significant at the 5% level. Recompute the predictions of the previous two questions. Are the intervals wider or narrower? Which predictions would you prefer? Test this model against the previous model. Which one is the preferred model?

   (f) For the model with `lpsa` as the response and all the other variables as predictors perform regression diagnostics answering the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say.

   - Check the constant variance assumption for the errors.
   - Check the normality assumption.
   - Check for large leverage points.
   - Check for outliers.
   - Check for influential points.
   - Check the structure of the relationship between the predictors and the response.

2. (Faraway 2005) Consider the `pipeline` data (do `library(faraway)` in `R`). Researchers of the National Institutes of Standard and Technology (NIST) collected pipeline data on ultrasonic measurements of the depths of defects in the Alaska pipeline in the field. The depths of the defects were then remeasured in the laboratory. These measurements were performed in six different batches. It turns out that this batch effect is not significant and so can be ignored in the analysis that follows. The laboratory measurements are more accurate than the in-field measurements, but more time consuming and expensive. We want to develop a regression equation for correcting the in-field measurements.

   (a) Fit a regression model with `Lab` as a response variable and `Field` as the explanatory variable. Check for nonconstant variance.

   (b) Find transformations on `Lab` and/or `Field` so that in the transformed scale the relationship is approximately linear with constant variance. You may restrict your choice on transformation to square root, log and inverse.

3. (Faraway 2005) For the `ozone` data (do `library(faraway)` in `R`), fit a model with `O3` as the response and `temp`, `humidity` and `ibh` as predictors. Use the Box-Cox method to determine the best transformation

4. Determine whether there are differences in the weights of `chickens` according to their feed in the `chickwts` data (available in `R`). Perform all the necessary model diagnostics.

5. Analyze `warpbreaks` data (available in `R`) as a two-way ANOVA. Which factors are significant? Now check for a good transformation on the response and see whether the model can be simplified. Now form a six-level factor from all combinations of `wool` and `tension` factors. Which combinations are significantly different?