Andrew Hillard

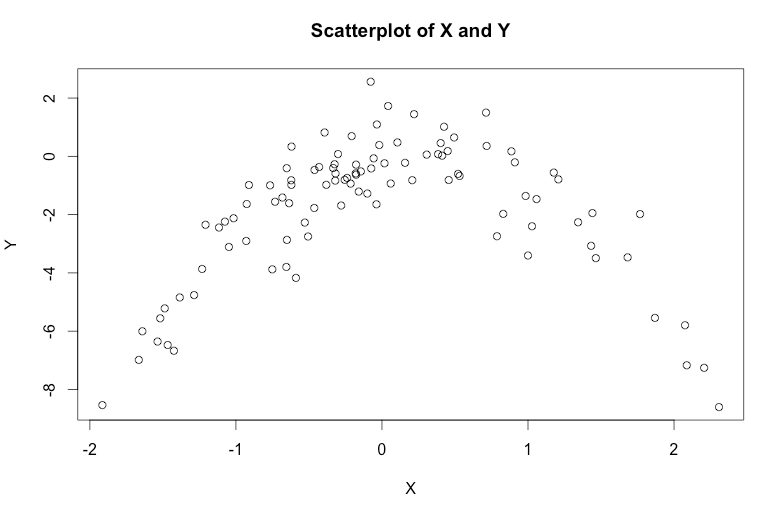
Data 3

Homework 2

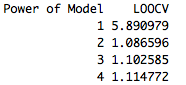
**5.8)**

a. n, the sample size, equals 100 and p, the number of parameters, equals 2. The model is where is normally distributed with mean equal to 0 and standard deviation equal to 1.

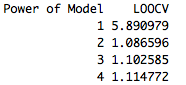
b. The scatterplot shows a non-linear relationship between X and Y.



c.



d. The LOOCV results are the same because the LOOCV procedure does not depend on the seed. The LOOCV procedure leaves out each data point since there are k=n folds.



e. Model 2 has the lowest LOOCV, which is expected because the model used to simulate this data was quadratic.

f. The statistical significance of the coefficient estimates for each model supports the conclusions of the cross-validation results. The linear model had no significant predictors, which makes sense because the data is non-linear. Both predictors of the quadratic model are significant. The coefficients for and are significant in the cubic and quartic models, but not the coefficients for and . This supports the conclusion that a second-degree model is the best model since third and fourth degree models provide no additional information and a first-degree model has insignificant predictors.

**5.9)**

a.

b. 0.40886

c. Estimate of the SE of the Mean using Bootstrap = 0.4086137. The bootstrap estimate is slightly lower than the estimate in part b.

d. The bootstrap estimate of the 95% confidence interval is (21.71558, 23.35004). The t-test 95% confidence interval is (21.72953, 23.33608). The two confidence intervals are very close.

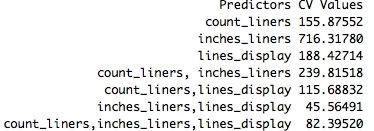
e.

f. Estimate of the SE of the Median using Bootstrap = 0.3778578. The SE of the Median Estimate is similar to the SE of the Mean Estimate.

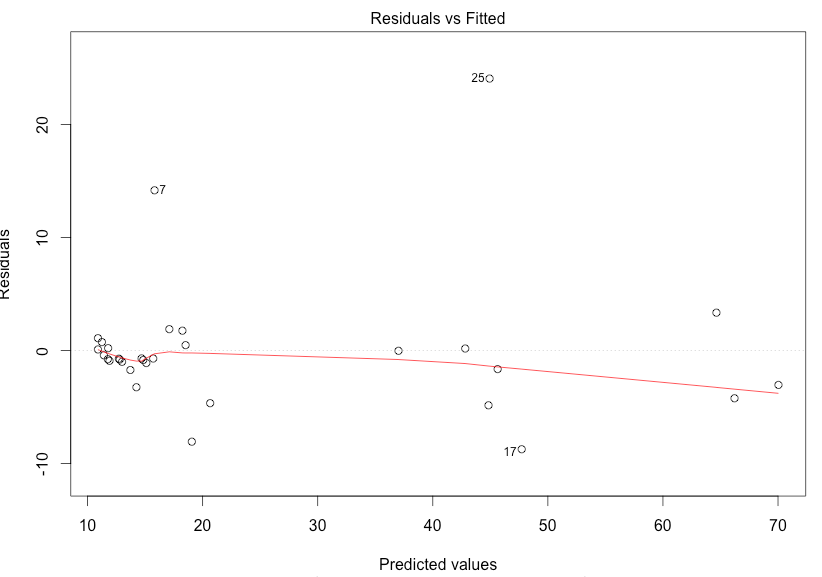
g. 12.75

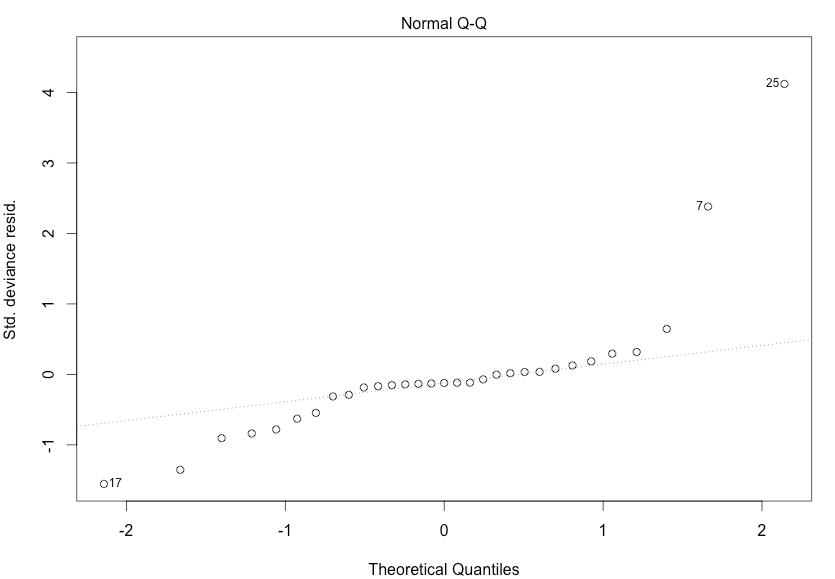
h. Estimate of the SE of the Tenth Percentile using Bootstrap = 0.5009769. The Bootstrap estimate for the SE of for the 10th Percentile is greater than the Bootstrap estimate of the SE for the mean and median.

**3)** The CV values for each model are reported below. The “best model” is the two variable model with inches liners and lines display as predictors.

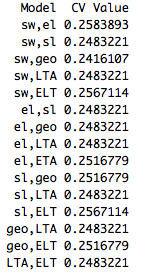
****

The residual plots suggest problems with the normal error regression assumption. The residual versus fitted plot (see below) displays a residual pattern that does not appear random, indicating non-linearity or lack of independence. The normality plot (see below) does not support a hypothesis of normality, and the leverage plot (not shown) indicates outliers.





**4)** The two variable models and CV value is given below. The best two variable model had sw (column 5, southwestness) and geo (column 8, geology). The best two variable model has a CV value of 0.2416107, which can be interpreted as a classification error rate of 24.16%.



When any combination of predictors could be used, the only model that performed better than the best two variable model was the four variable model with southwestness, slope, geology, and LTA. The four variable model has a CV value of 0.2348993. However, the results of this cross validation should be taken with a grain of salt. When the cross validation was run again with different random folds, the resulting best model was different.