Stat 425 Hw 4

Question 1 LPGA (10 points)

Load the data file `lpga2009.csv` .

a. (0.75 point) Fit a SLR model to predict `prize` with `percentile` in tournaments (Same as Hw 3 #5a). Use the Box Cox method on this model, and identify and apply an appropriate transformation to the response variable. Fit a model using the transformed response.

b. (0.25 pts) Comment on the normality of the transformed residuals.

c. (0.75 pts) Check for high leverage points, outliers, and influential points.

Fit a OLS model of `ln.prize.` vs all the other predictors except `prize` . We will consider this model the **full model**.

d. (0.25 pts) Look at the VIFs – do you think collinearity is an issue?

Tip: (Tip:**) *When you run the `step` function in R, you can save the results into an object. This will store the final model, which can be used later.*

e. (0.5 pts) Starting with the *full model*, Use Backwards Elimination to select a model based on **AIC**. What model do we end up with?

f. (0.5 pts) Starting with the *full model*, Use Backwards Elimination to select a model based on **BIC**. What model do we end up with?

g. (0.5 pts) Starting with the *intercept only model*, Use Forward Selection to select a model based on **AIC**. What model do we end up with?

Tip: (Tip:**) *It may be easier to first define both models separately before using them in the `step` function*

h. (0.5 pts) Starting with the *intercept only model*, Use Forward Selection to select a model based on **BIC**. What model do we end up with?

i. (5 points) For **each** of the 4 models in part (e)-(h), use 10-fold cross validation and calculate the Cross-validation SSE. *(Using each fold as the test set, find the SSE between the predictions on the test set and the actual response. Sum the SSEs from all the folds to get the cross-validation SSE)*. Which model does the cross validation approach prefer?

j. (1 point) For **each** of the 4 models, report the value of adjusted R^2. Which model or models does adjusted R^2 prefer?

Question 2 Variance Stabilizing Transformation (3 points)

Suppose $Var[Y]$ is proportional to $(E[Y])^3$. Find a variance stabilizing transformation $h(y)$. Show your work.

Question 3 Permutation Test (4 points)

Dustin has a fancy machine and wants to test to see if most people's reaction times improve in another dimension called the *Upside Down*. He collects the following paired data *in ms* for 20 subjects in the real world and in the *Upside Down*. For the sake of this experiment, improvement is defined as whether or not an individual's reaction time is **lower** in the *Upside Down*. (Suppose there is no variability in an individual's reaction times)

Download and load the file `reaction.csv` .

a. (1 point) Run a paired t-test on this dataset to compare the means. What is your conclusion? Note: you can use the **R** command `t.test` .

b. (0.5 points) Calculate the number of observations who have better (lower) reaction times in the *Upside Down*.

c. (2.5 point) Code and perform a permutation test with 1000 repetitions to determine whether most people's reaction times improve in the upside down. This test should permute the reaction times in the regular world, compare with the result in part (b), and return a p-value for the permutation test. Is there significant evidence at α = 0.05 to conclude that most people have faster reaction times in the upside down?

Question 4 Polynomial (3 points)

Download and load the file `polydata.csv` . Start by fitting a 5th order polynomial using `y` as the response.

a. (1 point) Determine if the 5th order polynomial term is significant. If not, remove it and test to see if the 4th order polynomial term is significant…Continue on and stop until you find a polynomial term that is significant. What order polynomial best suits this data?

b. (2 points) Make a scatterplot of x vs y. Then, use the predict function, predict the value of the response at all given values of x. Overlay the following 3 items as lines onto your scatterplot:

- Prediction Line (curve)
- Lower bound of 95% Prediction interval at each point
- Upper bound of 95% PI at each point.

(you are essentially plotting 95% prediction bands and the curve of best-fit)