# Stat 3502, Final Assignment

***Instructions: Due 11:59 pm Dec 12. Open book. Do not discuss with each other. Submit on Canvas.***

**The final in-class exam is closed book, and is on Dec 13 from 10:30 to 12:30 pm. You are allowed to bring one formula sheet.**

1. (3 points) Consider the beer price and sales data set "beer.csv". This data set contains 52 weeks of price and sales data for 3 carton sizes of beer at a small chain of supermarkets. The price and quantity-sold variables have all been converted to a per-24-can basis. For example, the value of \$19.98 for the price of 12-packs in week 1 means that a 12-pack sold for \$9.99 in that week, and the value of 223.5 for cases of 12-packs sold in that week means that 447 12-packs were sold.

Take logarithm transformation of the cases sold of 18-packs and denote it as the response $y$. Take logarithm transformation of the price of 12-packs, price of 18-packs, and price of 30-packs. Denote them as $x_1$, $x_2$ and $x_3$ respectively. Fit a multiple linear regression with $y$ as the response and $x_1$, $x_2$, $x_3$ as the predictors.

a) Do the predictors have significant effects on the response? Comment on the signs of estimated regression coefficients.

b) The first column of the data set denotes the week of the year. Plot the residual from the model above v.s. the variable week. Do you think we should add this variable to the model? Do you expect week to have a positive or a negative effect on the cases sold of 18-packs ?

c) Denote week as $x_4$. Fit a multiple linear regression with $y$ as the response and $x_1$, $x_2$, $x_3$, $x_4$ as the predictors. Perform a formal test to see if $x_4$ has a positive effect on $y$.

2. (4 points) Consider the auto mpg data set "auto.data", and the description of the data set is at

`https://archive.ics.uci.edu/ml/datasets/Auto%2BMPG`

Use R command *read.table* to read the data into R.

a) Draw the scatter plot between *mpg* (miles per gallon) v.s. *weight* (in pounds). Do you think linear regression with *mpg* as the response and *weight* as the predictor will be a good fit? Why or why not?

b) Create a new variable $y$ (gallons per 100 miles) and another new variable $x_1$ (weight in 1000 pounds). Draw the scatter plot between $y$ and $x_1$. Do you think linear regression will be a good fit?

c) Create a new variable $x_2$ that corresponds to the model year of the automobile. Fit a multiple linear regression with $y$ as the response, $x_1$ and $x_2$ as the predictors. Write the fitted regression model, comment on the signs of the estimated regression coefficients, and explain the meanings of $\hat{\beta}_1$ and $\hat{\beta}_2$.

d) Create a new variable $\tilde{y}$ (gallons per mile). Fit a multiple linear regression with $\tilde{y}$ as the response, $x_1$ and $x_2$ as the predictors. Compare the $R^2$ of this model with the model in part c). Then compare the estimated regression coefficients of this model with the model in part c).

3. (3 points) Consider the data set "fish.txt" and suppose we want to predict the weight of a fish. Use R command *read.table* to read the data into R. The description of the data set is at

`http://jse.amstat.org/datasets/fishcatch.txt`

Remove the 14th observation where weight is missing and remove the 47th observation where weight is 0.

a) Suppose we know the weight and length of a fish roughly follow the power curve equation

$$weight = a \times length^b \text{ with b close to 3.}$$

What transformations should we use before we fit a linear regression model? Implement the transformation(s) needed and then fit a linear regression model. Comment on the estimated regression coefficient. Here length should be the length from the nose to the end of the tail.

b) Create an indicator variable $z$ such that $z = 1$ if the fish species is smelt or pike, and $z = 0$ otherwise. Fit a multiple linear regression model with $z$ as an additional predictor to the simple linear regression model in part a). Formally test if $z$ has a significant effect on the response.

c) Denote the continuous predictor in part a) as $x$. Write the population level interaction model where we have the interaction between $x$ and $z$. Do you expect the interaction effect to be significant? Why or why not? Perform an analysis to formally test your conjecture.