Homework 4

PSTAT Summer 2024

Due date:Aug 4th 2024 at 23:59 PT

‘

1. This question uses the Auto dataset available in the ISLR package. The dataset under the name *Auto* is automatically available once the ISLR package is loaded.

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.2.3
```

```
data(Auto)
```

The dataset *Auto* contains the following information for 392 vehicles:

- mpg: miles per gallon
- cylinders: number of cylinders (between 4 and 8)
- displacement: engine displacement (cu.inches)
- horsepower: engine horsepower
- weight: vehicle weight (lbs)
- acceleration: time to accelerate from 0 to 60 mph (seconds)
- year: model year
- origin: origin of the vehicle (numerically coded as 1: American, 2: European, 3: Japanese)
- name: vehicle name

Our goal is to analyze several linear models where *mpg* is the response variable.

(a) **(2 pts)** In this data set, which predictors are qualitative, and which predictors are quantitative?
(b) **(2 pts)** Fit a MLR model to the data, in order to predict mpg using all of the other predictors except for name. For each predictor in the fitted MLR model, comment on whether you can reject the null hypothesis that there is no linear association between that predictor and mpg, conditional on the other predictors in the model.
(c) **(2 pts)** What mpg do you predict for a Japanese car with three cylinders, displacement 100, horsepower of 85, weight of 3000, acceleration of 20, built in the year 1980?
(d) **(2 pts)** On average, holding all other predictor variables fixed, what is the difference between the mpg of a Japanese car and the mpg of an European car?
(e) **(2 pts)** Fit a model to predict *mpg* using origin and horsepower, as well as an interaction between origin and horsepower. Present the summary output of the fitted model, and write out the fitted linear model.
(f) **(2 pts)** If we are fitting a polynomial regression with mpg as the response variable and weight as the predictor, what should be a proper degree of that polynomial?

(g) **(4 pts)** Perform a backward selection, starting with the full model which includes all predictors (except for name). What is the best model based on the AIC criterion? What are the predictor variables in that best model?

2. Use the *fat* data set available from the *faraway* package. Use the percentage of body fat: *siri* as the response, and the other variables, except *bronzek* and *density* as potential predictors. Remove every tenth observation from the data for use as a test sample. Use the remaining data as a training sample, building the following models:

(a) **(5 pts)** Linear regression with all the predictors.
(b) **(5 pts)** Linear regression with variables selected using AIC and BIC. Include comparison plots and comment on your findings.
(c) **(5 pts)** Ridge regression.
(d) **(5 pts)** Use the models you fit to predict the response in the test sample (provide point and interval estimate). How do the models compare in terms of prediction?