DAT 500S Module 1

Quiz:

Exercise 2.4 Q1&Q2

Group Assignment:

Exercise 2.4 Q9

Individual Assignment:

Exercise 2.4 Q10

Module 1 – Purpose of Machine Learning

- Estimate unknown (True) function *f*.
- Machine-Learning "learns" about *f* from data.
- **Why do we care about estimating *f*?**
- 2 reasons for estimating *f*,
  - **Prediction** and
  - **Inference**

Module 1 – How do we estimate the unknown function?

- **Training data**

- **Use the training data** and **a learning method** to estimate *f*.
- Learning Methods:
  - **Parametric Methods (pre-determined structure)**
  - **Non-parametric Methods (flexible)**

Module 1 – Terminology

- **Supervised Learning vs. Unsupervised Learning**
- **Regression vs. Classification**
- **Training Data vs. Test Data**
- **Y variable: Response, Target, Outcome, and Dependent Variable.**
- **X variable: Independent and Predictor**
- **Machine-Learning: Statistical Learning, Data Mining, Artificial Intelligence, etc.**

Exercise 2.4 Q1 (questions in the note)

- (a) The sample size n is extremely large, and the number of predictors p is small.

Better. A flexible method will fit the data closer and with the large sample size, would perform better than an inflexible approach.

- (b) The number of predictors p is extremely large, and the number of observations n is small.

Worse. A flexible method would overfit the small number of observations.

- (c) The relationship between the predictors and response is highly non-linear. Better.
- (d) The variance of the error terms, i.e. $\sigma^2 = Var(\varepsilon)$, is extremely high.

Worse. A flexible method would fit to the noise in the error terms and increase variance.

Exercise 2.4 Q2

- (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Regression and inference with n=500 and p=3

Inference: a conclusion reached on the basis of evidence and reasoning

Exercise 2.4 Q2

- (b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Classification and accuracy with n=20 and p=13

Exercise 2.4 Q2

- (c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Regression and accuracy with n=52 and p=3

Exercise 2.4 Q9

- (a) Which of the predictors are quantitative, and which are qualitative?

Look Auto dataset documentation.->library(ISLR) help(Auto)

Remove NAs if any. -> Auto = na.omit(Auto)

-> fix(Auto) / View(Auto)

-> summary(Auto)

Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numerical variables (e.g. how many; how much; or how often).

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

Exercise 2.4 Q9

- (a) Which of the predictors are quantitative, and which are qualitative?

We can find out that among 9 variables “name” is a qualitative variables. Also intuitively we can judge that “origin” is a qualitative variable, which is understood by R as a quantitative. We need to transfer “origin” to a factor if we want to run regression on this variable.

-> Auto$origin = factor(Auto$origin)

year, cylinder, name, origin

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

- -> range(Auto$mpg)

- -> sapply(Auto[,-c(8,9)], range)
- -> summary(Auto)

(c) What is the mean and standard deviation of each quantitative predictor?

- -> sapply(Auto[, -c(8, 9)], mean)
- -> sapply(Auto[, -c(8, 9)], sd)

or

- -> summary(Auto)

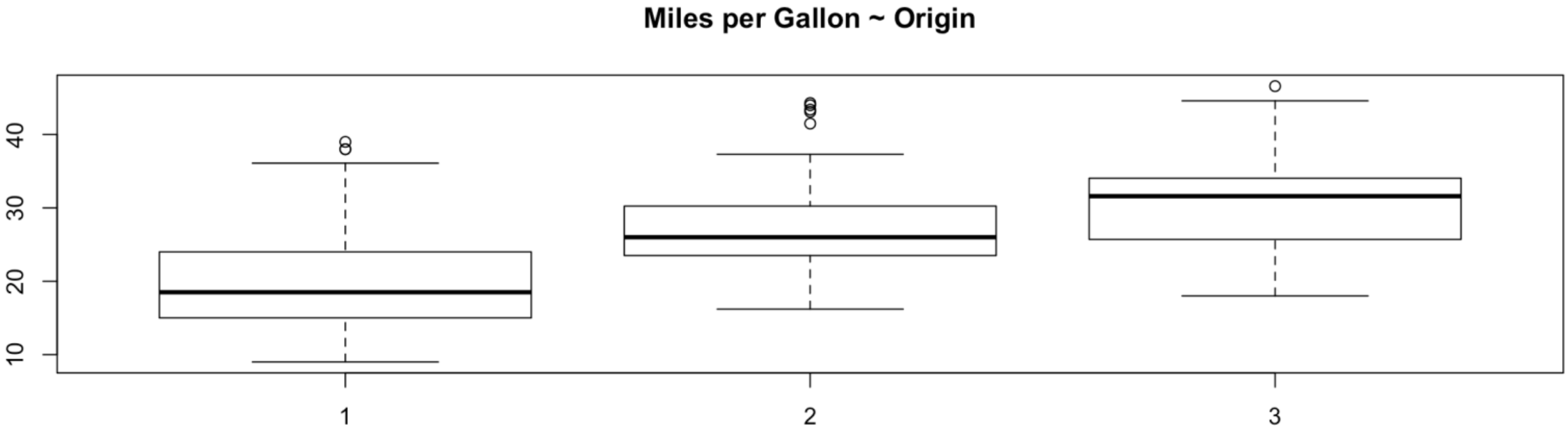
(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

- subset <- auto[-c(10:85), -c(4,9)]

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

- Scatterplot: -> pairs (Auto)

-> plot(mpg, horsepower, data = Auto, xlab = "", ylab = "" , main = "Miles per Gallon ~ Origin")



(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

- -> round(cor(Auto[, -9]), digits = 3)
- I think acceleration and year are not good

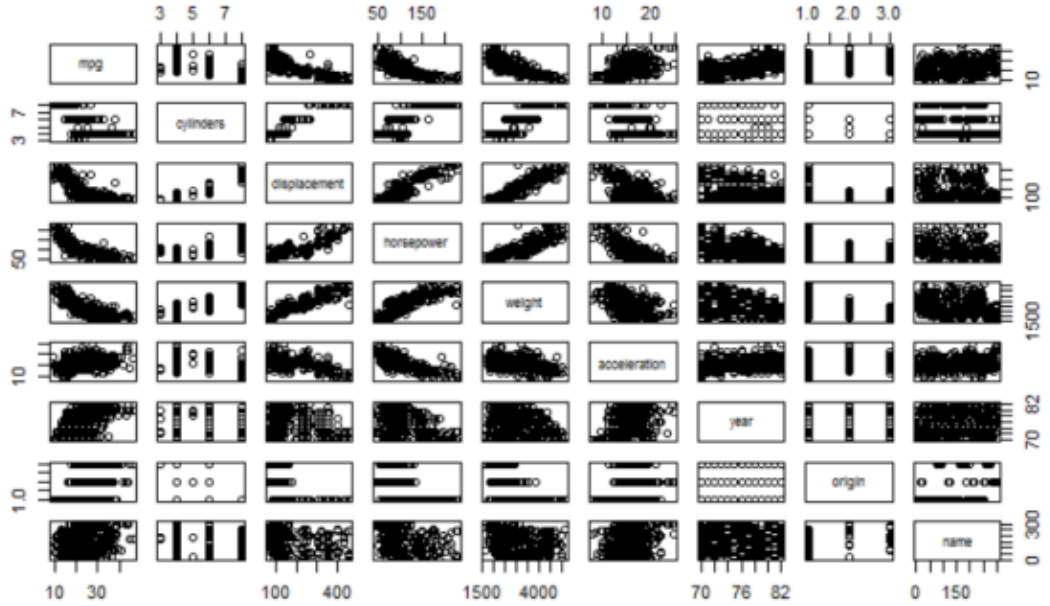
by looking at the value.

- What do you think? Any other factors to consider?

	mpg
mpg	1.0000000
cylinders	-0.7776175
displacement	-0.8051269
horsepower	-0.7784268
weight	-0.8322442
acceleration	0.4233285
year	0.5805410
origin	0.5652088

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

- look at scatterplot
- do a regression;
- look at the p -value




```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.218435   4.644294  -3.707  0.00024 ***
cylinders    -0.493376   0.323282  -1.526  0.12780
displacement  0.019896   0.007515   2.647  0.00844 **
horsepower   -0.016951   0.013787  -1.230  0.21963
weight       -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration  0.080576   0.098845   0.815  0.41548
year          0.750773   0.050973  14.729  < 2e-16 ***
origin        1.426141   0.278136   5.127 4.67e-07 ***
```