Analytics 512: Homework # 1

Arif Ali

January 28th, 2016

please note, all the plots are located in https://github.com/arifyali/Statistical-Learning-Analytics-512-Spring-2016/tree/master/HW1 because a pdf was required for submission and there is an issue rendering the plots when making the pdf.

Exercise 2.4 #2

Part A

This is a regression problem because the objective is to figure out how the different explanatory variables affect CEO salary. Salary is not binned into categories, but rather is a number.

n is the 500 firms since they are the observations and *p* is profit, number of employees, and industry.

Part B

Since the response variable is binary, this is a classification problem.

n is the observation of the 20 previous products and *p* is price charged for the product, marketing budget, competition price, and ten other variables (not the success/failure variable).

Part C

This is a regression problem because the percent change is trying to determine to exact change in percentage.

n is the number of weeks during 2012 and *p* the % change in the dollar, the % change in the British market, and the % change in the German market.

Exercise 2.4 #4

Part A

1. Determining whether a person has cancer or not. The Response would be a binary response of whether someone does or doesn't have cancer. This could be expanded to checking the specific type of cancer (ie lung, breast, skin, brain, etc). The predictors could be health measurements of an individual's vitals. An individual cancer history rating, etc. Goal is prediction; however, changes in vitals could be used as a measure of inference.
2. Predicting what elected official wins. The Response is a candidate. Predictors could be money raises, whether a candidate is an incumbent, and the party the candidate is affiliated. County Statistics like average income or Census Data could be used. While the goal would be to predict, inference could be used to study with counties, a party should invest in campaigning in.
3. Determining the reason behind whether a TV Series is cancel or renewed. This could be based on nielsen ratings, DVD sales, the cost of production. This seems more of an inference goal, trying to figure out what makes a show worth renewing.

Part B

1. Determining the amount a indivdual makes. The response would be the income amount of a person. The Predictor would be credit history, payement, debt, assets. Goal would probably be more along the lines of inference because observing changes in the predictors might lead into insights as how much money an indivdual makes.
2. Determing the amount of rain fall in a region. The response would the amount of rain and predictors could be time, amount of overcast, precipitation. The goal is prediction; the reason inference is probably not needed is because the relationships have probably already been uncovered.
3. Determining the change in stock market. The Response would the amount the market increases or decreases. The Predictors could be currency rate changes, consumer sentiment, etc. This is a predicton goal.

Part C

1. Using cluster analysis to looking at build of blood cells could be used to identitify blood clothes.
2. Cluster analysis could be used to find vehicle built-ups which could be useful in planning highways and extra roadlanes in certain areas.
3. In a war zone, population clusters could be used by peacekeeping forcing to set up safe zones for refugees. Using cluster analysis could effectively identitify clusters.

Exercise 2.4 #7

Part A

Observation	Euclidean Distance
1	3
2	2
3	$\sqrt{10}$
4	$\sqrt{5}$
5	$\sqrt{2}$
6	$\sqrt{3}$

Part B

Green, because observation 5 is the the 1^{st} nearest neighbor and the response for 5 is green.

Part C

Red, 2 out of the 3 closest neighbors has a response of red.

Part D

Small, it would be more flexible whereas a bigger value of K would result in trying to accomodate more points.

Exercise 2.4 #9

Part A

In [2]:
```
library("ISLR")
data(Auto)
Auto = na.omit(Auto)
Auto = Auto[!is.null(Auto),]
summary(Auto)
```

Out[2]:
```
      mpg          cylinders      displacement     horsepower        weight
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140

  acceleration        year           origin                      name
 Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
 1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
 Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
 Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
 3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
 Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
                                                 (Other)           :365
```

The quantitative predictors: mpg, displacement, horsepower, weight, and acceleration.

The qualitative predictors: cylinders, year, origin, and name.

Part B

In [3]:
```
quantitative = c("mpg", "displacement", "horsepower", "weight", "acceleration")

partb = mapply(range, Auto[, quantitative])
rownames(partb) = c("lower", "upper")
partb
```

Out[3]:

	mpg	displacement	horsepower	weight	acceleration
lower	9	68	46	1613	8
upper	46.6	455.0	230.0	5140.0	24.8

Part C

In [4]:
```
partc = cbind(mapply(mean, Auto[, quantitative]),mapply(sd, Auto[, quantitative]))
colnames(partc) = c("mean", "sd")
partc
```

Out[4]:

	mean	sd
mpg	23.445918	7.805007
displacement	194.412	104.644
horsepower	104.46939	38.49116
weight	2977.5842	849.4026
acceleration	15.541327	2.758864

Part D

In [5]:
```
Auto_minus10_to_85 = Auto[,-(10:85)]
partd = cbind(t(mapply(range, Auto_minus10_to_85[, quantitative])),
      mapply(mean, Auto_minus10_to_85[, quantitative]),
      mapply(sd, Auto_minus10_to_85[, quantitative]))
colnames(partd) = c("lower range","upper range", "mean", "sd")
partd
```

Out[5]:

	lower range	upper range	mean	sd
mpg	9.000000	46.600000	23.445918	7.805007
displacement	68.000	455.000	194.412	104.644
horsepower	46.00000	230.00000	104.46939	38.49116
weight	1613.0000	5140.0000	2977.5842	849.4026
acceleration	8.000000	24.800000	15.541327	2.758864

Part E

In [6]:
```
png("HW1E9e.png")
pairs(Auto[,quantitative])
dev.off()
```

Out[6]: **pdf:** 2

For mpg, there seems to be a negative relation between it compared to displacement, horespower, and weight. The relation betwwen acceleration and mpg is not as well defined but seemingly more positive.

The is a positive relation between displacement compared to horsepower and weight. There seems to be a negative relationship between displacement and acceleration.

Horsepower and weight has a postive relationship and horsepower and acceleration have a negative relationship.

Relation weight and acceleration is less defined, but there seems to be a slightly negative relationship

Part F

The relation between mpg and horsepower seems to fit a line resembling $y = e^{-x}$. Displacement follows this line as well; however, the points in the middle are more spread out compared to mpg vs. horsepower.

Exercise 2.4 #10

Part A

In [3]:
```
library(MASS)
```

In [8]:
```
dim(Boston)
names(Boston)
```

Out[8]: 506 14

Out[8]: 'crim' 'zn' 'indus' 'chas' 'nox' 'rm' 'age' 'dis' 'rad' 'tax' 'ptratio' 'black' 'lstat' 'medv'

The rows represents observation in Bostonian suburbs; there are 506 observations, which seem to correspond to each Boston suburb. The columns are different attributes for the suburbs; there are 14 columns.

Part B

In [9]:
```
png("HW1E10b.png")
pairs(Boston)
dev.off()
```

Out[9]: **pdf:** 2

Chas is obviously a binary variable based on how the scatter plots look. RM and medv seem to have a postive relationship; however, RM and lstat seems to have a negative relationship. Nox and dis have a relationship best describe by $y = e^{-x}$.

Part C

In [10]:
```
png("HW1E10c.png")
par(mfrow = c(4,4))
for(i in names(Boston)){
    plot(Boston$crim, Boston[, i], xlab="crim", ylab = i)
    }
dev.off()
```

Out[10]: **pdf:** 2

It seems that the higher the age, the more crime. It could be that the older houses could be in more rundown neighborhoods; therefore, crime is more rampant. The lower the distance variable, the more crime, which is interesting because it seems counter-intuitive. In areas with higher tax rates, there is more crime. This could be because lower economic classes are taxed higher, and there is a relationship between poverty and crimes.

Part D

In [4]:
```
partd = cbind(quantile(Boston$crim, seq(from = 0, to = 1, by = .1)),
      quantile(Boston$tax, seq(from = 0, to = 1, by = .1)),
      quantile(Boston$ptratio, seq(from = 0, to = 1, by = .1)))
colnames(partd) = c("crim","tax","ptratio")
partd
png("HW1E10d.png")
par(mfrow = c(3,1))
hist(Boston$crim, breaks = 25)
hist(Boston$tax, breaks = 25)
hist(Boston$ptratio, breaks = 10)
dev.off()
```

Out[4]:

	crim	tax	ptratio
0%	6.32e-03	1.87e+02	1.26e+01
10%	0.038195	233.000000	14.750000
20%	0.06417	273.00000	16.60000
30%	0.099245	289.000000	17.800000
40%	0.15038	307.00000	18.40000
50%	0.25651	330.00000	19.05000
60%	0.55007	398.00000	19.70000
70%	1.72844	437.00000	20.20000
80%	5.58107	666.00000	20.20000
90%	10.753	666.000	20.900
100%	88.9762	711.0000	22.0000

Out[4]: **pdf:** 2

The crime rate is pretty low throughout boston, with less than 40% of the suburb having a crime rate of greater than one. Based on the quantile, there seems to be a right tail as indicated. Boston seems to be a pretty safe place. However, there appears to be one suburb that has an abnormally high crime rate.

Two of the suburbs have very high Tax rates. There is a significant gasp, which could indicate income inequality.

There are an unusual number of suburbs with a ratio between 20 and 21. However no indication that there is any skew. Therefore none of the suburbs seem to have a particularily high pupil-teacher rate.

Part E

In [14]:
```
sum(Boston$chas == 1)
```

Out[14]: 35

Part F

In [15]:
```
median(Boston$ptratio)
```

Out[15]: 19.05

Part G

In [22]:
```
Boston[Boston$medv == min(Boston$medv),]

t(mapply(function(x) round(mean(x), digits = 4), Boston))
```

Out[22]:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
399	38.3518	0	18.1	0	0.693	5.453	100	1.4896	24	666	20.2	396.9	30.59	5
406	67.9208	0	18.1	0	0.693	5.683	100	1.4254	24	666	20.2	384.97	22.98	5

Out[22]:

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
3.6135	11.3636	11.1368	0.0692	0.5547	6.2846	68.5749	3.7950	9.5494	408.2372	18.4555	356.6740	12.6531	22.532

There are two suburbs with the lowest median house value referred to as min suburbs. In terms of crime, the min suburbs have significiantly higher rates. There are no zoned lots. Both have the same proportion of non-retail business acres per town, which was higher than the average. Neither min suburb borders the Charles River. There less average number of rooms in the min suburbs than overall. It looks like both suburbs have 100% of owner-occupied units built prior to 1940. The tax rate is higher than the average. The min suburbs are more accessible from the radial highways than the average suburbs. There is a slightly higher proportion of blacks in the min suburbs. The pupil teacher ratio is slightly higher in the min suburbs than overall. The weighted mean of distances to five Boston employment centres is shorter for the min suburbs.

Part H

In [17]:
```
sum(Boston$rm>7)
sum(Boston$rm>8)
```

Out[17]: 64

Out[17]: 13

In [38]:
```
t(mapply(function(x) round(mean(x), digits = 4),Boston[Boston$rm>8,]))
```

Out[38]:

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0.7188	13.6154	7.0785	0.1538	0.5392	8.3485	71.5385	3.4302	7.4615	325.0769	16.3615	385.2108	4.3100	44.2000

In [39]:
```
t(mapply(function(x) round(mean(x), digits = 4),Boston))
```

Out[39]:

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
3.6135	11.3636	11.1368	0.0692	0.5547	6.2846	68.5749	3.7950	9.5494	408.2372	18.4555	356.6740	12.6531	22.532

The suburbs where $rm > 8$ has a lower crime rate compared to the rest of the city. These suburbs are on average further than the average of the suburbs. Slight higher proportion are bordering the Charles River compared to the rest of the city. The Average age and median house value are higher. The tax rate is lower.