MATH 4322 Homework 1

Phu Nguyen

9/4/2021

Problem 1

Explain whether each scenario is a classification or regression problem,and indicate whether we are most interested in inference or prediction. Finally, provide n number of observations and p number of variables.

(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.

- **Answer**: This is a regression problem because the response 'output' is the CEO salary and it's a quantitative value. We are most interested in inference because the factors that affect CEO salary is number of employees, and industry. We are interested in the way the Y which is the CEO salary is affected as X changes. $n = 500$ and $p = 3$.

(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.

- **Answer**: This is a classification problem because the response is a qualitative value, which means no arithmetic operation can be done. Here we are interested in predicting the $\hat{Y} = \hat{f}(X)$, where $\hat{f}$ is often treated as a *blackbox* The response is success or failures $n = 20$ and $p = 13$.

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

- **Answer**: This is a regression problem because the response is the USD/Euro exchange rate. Variables are US market, British market, and German market, therefore $p = 3$. The number of observation is weekly, so in year 2012 there is $365/7 = 52$, therefore $n = 52$.

Problem 2

You will now think of some real-life applications for statistical learning. Think of ones other than what your friends have.

(a) Describe three real-life applications in which $classification$ might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

- (1) Check the weather if today is rainy or not rainy day. This can be predict by inputs of different type data, and we can get the data from the weather database. This application is inference because $\hat{f}$ is known.
- (2) Analyze images if a person is young or old. We can build a machine learning and feed it information of young and people. So that the machine can classify the difference between the two. This application is inference because $\hat{f}$ is known.
- (3) Speech recognition, we can have a machine learning that learn how human talk and learn the different frequency that each person produce. It can analyze if you're actually who you are when speaking. This is a prediction application because $\hat{f}$ is a *blackbox* here.

(b) Describe three real-life applications in which *regression* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

- (1) House price, we can predict the house price by when it was build, the area around it, number of floors, materials, etc... This is a inference because $\hat{f}$ is known.
- (2) Revenue, we can predict the revenue of a company by the number of sales. This is a inference because $\hat{f}$ is known.
- (3) Weight, we can predict a person's weight by their sex, heights and age. This is a prediction because $\hat{f}$ is not known here.

(c) Describe three real-life applications in which *cluster* analysis might be useful.

- (1) Volcano Studies, we studies different sites of volcano to see when it would erupted and determined the time and period to evacuated people.
- (2) Earthquakes Studies, we studies the geographic of a land and gather data to determine the danger zone and evacuated people.
- (3) City's houses, studies the values based on the geographic and locations.

Problem 3

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?
* **Answer**: If we are mainly interested in inference, then the less flexible are much more interpretable. For example, when inference is the goal, the linear model may be good choice since it's easier to understand the relationship between Y and $X_1, X_2, ..., X_p$. On the other hand, the flexible approach, such as the splines and the boosting methods can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.

Problem 4

This exercise involves the *mtcars* data set looked at in class.

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

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

- **Quantitative**: mpg, disp, hp, drat, wt, qsec, gear, carb.
- **Qualitative**: cyl, vs, am.

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

```
range(mtcars)
```

```
## [1]   0 472
```

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

```
## [1] 20.09062
```

```
## [1] 230.7219
```

```
## [1] 146.6875
```

```
## [1] 3.596563
```

```
## [1] 3.21725
```

```
## [1] 17.84875
```

```
## [1] 3.6875
```

```
## [1] 2.8125
```

Standard Deviation

```
## [1] 6.026948
```

```
## [1] 123.9387
```

```
## [1] 68.56287
```

```
## [1] 0.5346787
```

```
## [1] 0.9784574
```

```
## [1] 1.786943
```

```
## [1] 0.7378041
```

```
## [1] 1.6152
```

(d) Now remove the 10th through 32nd observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
mtcars.rm = mtcars[-c(10:32),]
```

```
range(mtcars.rm)
```

```
## [1]   0 360
```

Means

```
## [1] 20.09062
```

```
## [1] 230.7219
```

```
## [1] 146.6875
```

```
## [1] 3.596563
```

```
## [1] 3.21725
```

```
## [1] 17.84875
```

```
## [1] 3.6875
```

```
## [1] 2.8125
```

Standard Deviation

```
## [1] 6.026948
```

```
## [1] 123.9387
```

```
## [1] 68.56287
```

```
## [1] 0.5346787
```

```
## [1] 0.9784574
```

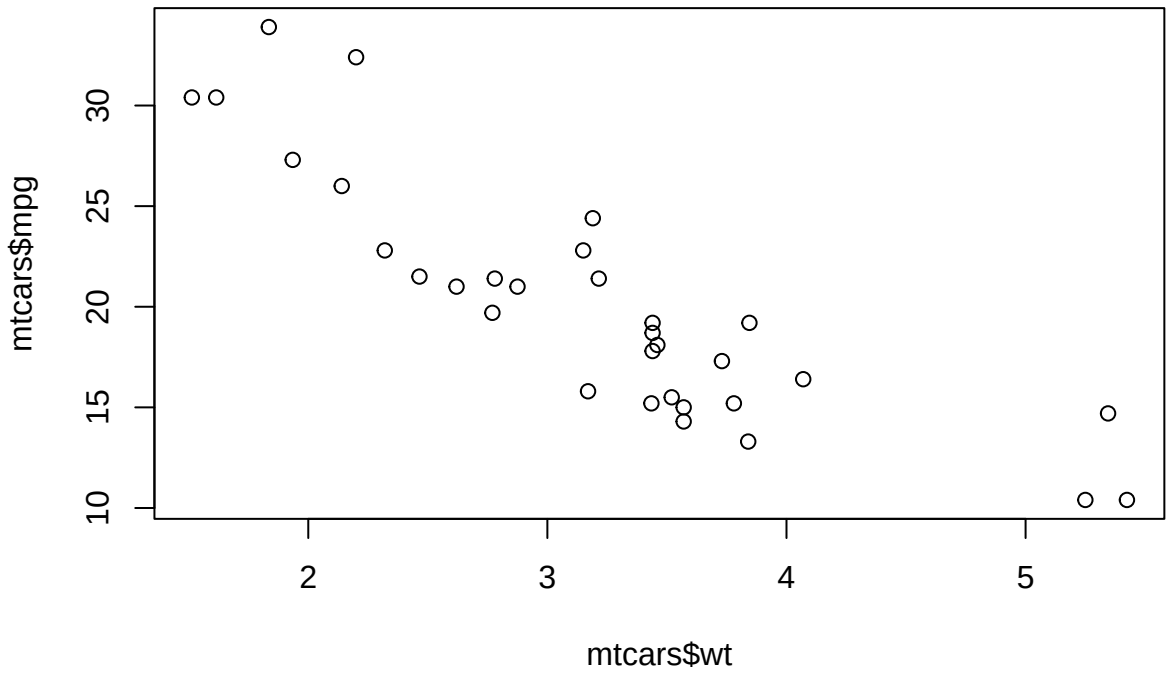
```
## [1] 1.786943
```

```
## [1] 0.7378041
```

```
## [1] 1.6152
```

(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.

```
plot(mtcars$wt, mtcars$mpg)
```



* The relationship between *wt* and *mpg* is neative linear relationsip

(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.

```
mpglm.lm = lm(mtcars$mpg ~ mtcars$disp + mtcars$hp + mtcars$drat + mtcars$wt)
summary(mpglm.lm)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$disp + mtcars$hp + mtcars$drat +
##     mtcars$wt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5077 -1.9052 -0.5057  0.9821  5.6883
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.148738   6.293588   4.631  8.2e-05 ***
## mtcars$disp  0.003815   0.010805   0.353  0.72675
## mtcars$hp   -0.034784   0.011597  -2.999  0.00576 **
## mtcars$drat  1.768049   1.319779   1.340  0.19153
```

```
## mtcars$wt   -3.479668   1.078371  -3.227  0.00327 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.602 on 27 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8136
## F-statistic: 34.82 on 4 and 27 DF,  p-value: 2.704e-10
```

- This model shows that beside from weight *wt*, Gross horsepower *hp* can be useful at predicting the *mpg*

Problem 5

This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the $MASS$ library in R.

```
library(MASS)
```

Now the data set is contained in the object Boston.

```
Boston
```

Read about the data set:

```
#?Boston
```

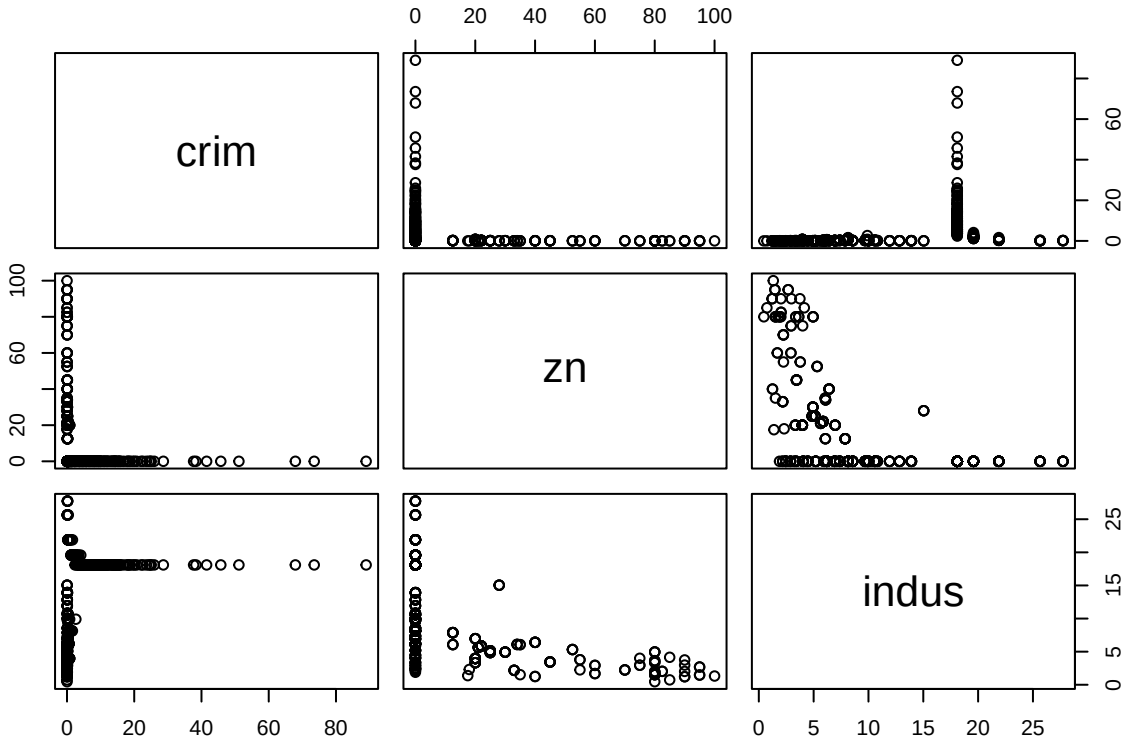
How many rows are in this data set? How many columns? What do the rows and columns represent?

```
dim(Boston)
```

```
## [1] 506  14
```

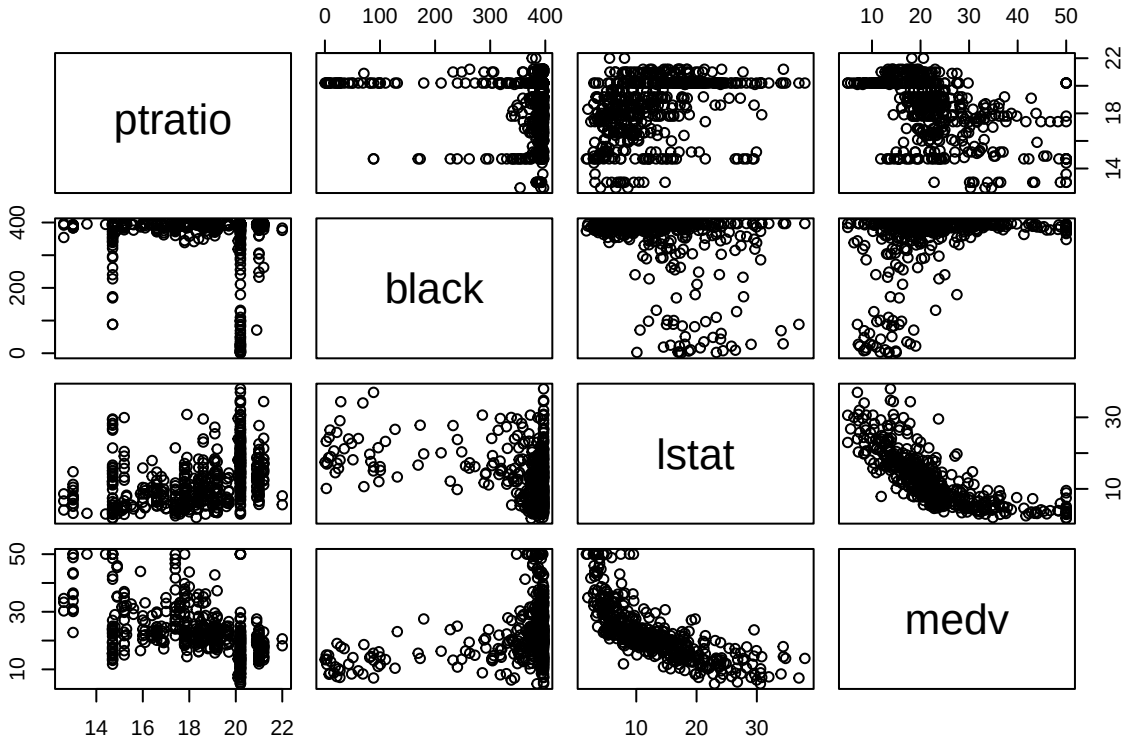
(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

```
pairs(Boston[,c(1:3)])
```



* This is a pairwise scatterplot between column 1 to 3. Which the variables are *crim*, *zn* and *indus*. If we look at the relationship between *zn* and *indus*, as the proportion of non-retail business acres per town increase, the proportion of residential land zoned for lots over 25,000 sq.ft decreases.

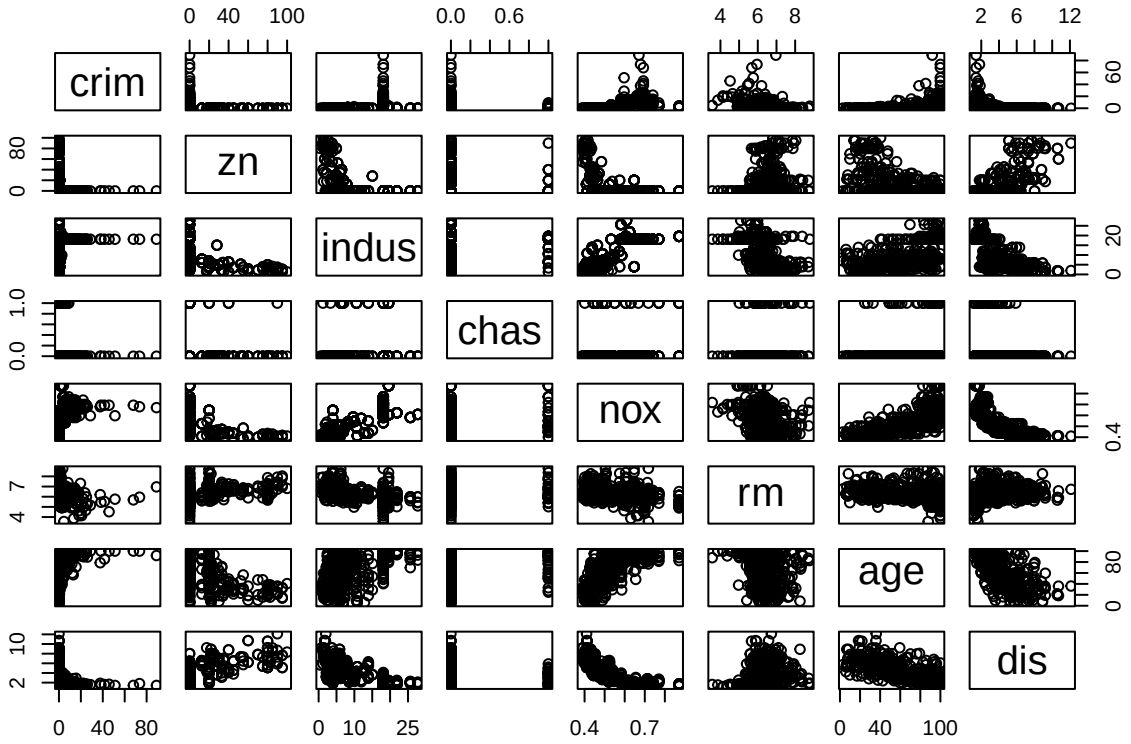
```
pairs(Boston [,c(11:14)])
```



* This is a pairwise scatterplot of *ptratio*, *black*, *lstat* and *medv*. If we take a look at the relationship between *medv* and *lstat* we can see that as the median value of owner-occupied homes increases, the lower status of the population decreases.

(c) Are any of the predictors associated with per capital crime rate? If so, explain the relationship.

```
pairs(Boston[,c(1:8)])
```



* Base on the correlation coefficients, there is an association between the per capita crime rate *crim* and the other predictors.

(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

```
summary(Boston$crim)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00632   0.08204   0.25651   3.61352   3.67708 88.97620
```

```
selection = subset(Boston, crim>10)
nrow(selection)/nrow(Boston)
```

```
## [1] 0.1067194
```

- 11% of the neighborhood's have crime rates that is above 10

```
summary(Boston$tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   279.0   330.0   408.2   666.0   711.0
```

```
selection = subset(Boston, tax < 400)
nrow(selection)/nrow(Boston)
```

```
## [1] 0.6047431
```

- 60.4% of the neighborhood pay under $400.

```
summary(Boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

```
selection = subset(Boston, ptratio > 20)
nrow(selection) / nrow(Boston)
```

```
## [1] 0.3972332
```

(e) How many of the suburbs in this data set bound the Charles river?

```
nrow(subset(Boston, chas == 1))
```

```
## [1] 35
```

(f) What is the median pupil-teacher ratio among the towns in this data set?

```
summary(Boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

- The median pupil-teacher ratio among the towns is 19.

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

```
over7rooms = subset(Boston, rm > 7)
nrow(over7rooms)
```

```
## [1] 64
```

- There are 64 suburbs that have more than 7 rooms

```
over8rooms = subset(Boston, rm > 8)
nrow(over8rooms)
```

```
## [1] 13
```

```
summary(over8rooms)
```

```
##       crim                zn             indus              chas        
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm              age              dis        
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black       
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      lstat            medv      
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7  
##  Median :4.14   Median :48.3  
##  Mean   :4.31   Mean   :44.2  
##  3rd Qu.:5.12   3rd Qu.:50.0  
##  Max.   :7.44   Max.   :50.0
```

- There are 13 suburbs that have more than 13 rooms