Homework 1

Instructions

Due: 5:00pm on Wednesday, September 15th

1. Add your name between the quotation marks on the author line in the YAML above.
2. Compose your answer to each problem between the bars of red stars.
3. Commit your changes frequently.
4. Be sure to knit your .Rmd to a .pdf file.
5. Push both the .pdf and the .Rmd file to your repo on the class organization before the deadline.

Theory

Problem 1

For each of parts (a) through (d), indicate whether we would generally expected the performance of a complex statistical learning method to be better or worse than a low complexity method. Justify your answer.

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

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

(c) The relationship between the predictors and response is highly non-linear.

(d) The variance of the error term $\text{Var}(\epsilon)$ is extremely high.

(a) With large sample size, a complex model is not as susceptible to overfitting, and with few predictors, may be better able to match the true relationship between predictors and response.

(b) With few observations, a simple model will often outperform a more complex model, since random samples of the data are likely to exhibit considerable variation, and so the complex model is liable to overfit.

(c) A simple model is likely to have large degree of bias when modeling extremely non-linear relationships. More complex models are likely more appropriate, especially in presence of large amounts of data.

(d) In general, more complex models are more susceptible to high variance between samples from the data set, as a result of either small sample size or large $\text{Var}(\epsilon)$. In this setting, simple models may be more appropriate.

Problem 2

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

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

(b) We are considering launching a new product and wish to know whether it will be a *success* or *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.

(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 teh British market, and the % change in the German market.

(a) This is a *regression* problem, since CEO salary is quantitative. The sample size is $n = 500$ and the number of predictors is $p = 3$.

(b) This is a *classification* problem, since the response (success) is categorical. However, we could also perform logistic regression. The sample size is $n = 20$ and the number of predictors is $p = 13$.

(c) This is a *regression* problem, since % change is quantitative. The sample size is $n = 52$ and the number of predictors is $p = 3$.

Problem 3

The following problem asks you to think of some real-life applications for statistical learning.

(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 answers.

(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 answers.

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

Answers may vary considerably.

(a) Examples:

- Spam filters; the response is whether an email is spam or not spam; predictors could include key words or phrases, whether the sender is in recipients contact lists, presence of links; the goal is almost entirely prediction.
- Digital character recognition; the response is the letter displayed ; predictors could include complete pixel color map, or at a higher level, character size, ratio of black/white pixels, presence of loops; the goal is almost entirely prediction.
- voter sampling; the response could the individual's presidential candidate preference; predictors could include party affiliation, geographic location, income level, race; the goal is inference, to learn what factors may influence candiate preference.

(b) Examples:

- Weather; the response could be the next day's temperature; predictors could include the past week's temperature and pressure, presence of high/low pressure systems, precipitation over the past week; the goal may be both inference (to determine factors influencing weather) and prediction (determine what the weather will be)

- House price; the response could be the list price for a house; predictors could include square footage, lot size, number of bedrooms, number of bathrooms, geographic location; the goal could be inference to assess features that influence price, as well as prediction, in the case of automated home pricing websites, like Zillo.
- Grades; the response could be a student's final grade in a course; predictors could include midterm scores; attendance rate; scores on homework; the goal is likely inference, to better understand how work undertaken throughout the term influences an overall course grade.

(c) Examples

- Movies: determine groupings within Academy Award best picture nominations from 1990 - 2020.
- Image processing: Given 8x8 pixilated images, determine the linear combination of predictors that explain the most amount of variability in the data.
- Surveys: Analyze responses to multiple choice questionnaires for similarities among many responses.

Problem 4

The table below provides a training data set containing six observations, three predictors, one categorical response variable, and one quantitative response variable.

Obs.	X_1	X_2	X_3	Y_1	Y_2
1	0	3	0	Red	5
2	2	0	0	Red	3
3	0	1	3	Red	1
4	0	1	2	Green	2
5	-1	0	1	Green	3
6	1	1	1	Red	4

Suppose we wish to use the data set to make a prediction for Y_1 and for Y_2 when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

a. Compute the Euclidean distance between each observation and the test point $X_1 = X_2 = X_3 = 0$.

b. What are the predictions for Y_1 and for Y_2 when $K = 1$? Explain.

c. What are the predictions for Y_1 and for Y_2 when $K = 3$? Explain.

d. What is the prediction for Y_1 and for Y_2 when $K = 6$? Explain.

e. If the Bayes decision boundary in this problem is highly non-linear, would we expect our predictions to be most accurate when K is large or when K is small? Explain.

---

a. The euclidean distance between $X_0 = (0, 0, 0)$ and an observation $x = (a, b, c)$ is

$$\sqrt{a^2 + b^2 + c^2}$$

Applying this formula to the six observations, we see

Observation	Distance
x_1	3
x_2	2
x_3	$\sqrt{10}$

Observation	Distance
x_4	$\sqrt{5}$
x_5	$\sqrt{2}$
x_6	$\sqrt{3}$

Ordering of observations from nearest to furthest: x_5, x_6, x_2, x_4, x_1, x_3.

b. The nearest observation is x_5, and so KNN with $K = 1$ predicts $Y_1 =$ Green and $Y_2 = 3$.

c. The three nearest observations are x_5, x_6, x_2. Of these, two are red and one is green, so $Y_1 =$ Red. The average of Y_2 for these observations is $\frac{10}{3}$.

d. When $K = 6$, the prediction for Y_1 is red (the most common level) and the prediction for Y_2 is 3 (the average of all values of Y_2).

e. If the Bayes decision boundary is non-linear, a flexible KNN model is more appropriate, and so we might expect better prediction accuracy with small values of K.

Applied

Problem 5

The `hw1_p5.csv` file contains 20 test data points for a predictor X and a quantitative response Y.

Three models were fit on a separate training data set consisting of 60 observations; a linear model, a quadratic model, and a septic model (i.e. polynomial of degree 7). The predictions made from these models on the test data are included in `hw1_p5.csv` as well.

a. Load the `hw1_p5` data set using the `read_csv(file = "...")` function (where ... should be replaced with the file path from your project directory to the csv file).

b. Plot the test data, along with color-coded curves for each model (hint: you can use `geom_line` to create a curve in `ggplot2` which interpolates between points in a data frame).

c. Based on inspection of the graph, which model seems to best fit the test data?

d. Calculate the MSE for each of the three models. (Use R to assist with calculation, don't do this by hand.)

e. Which model had the highest test MSE? Which had the lowest?

f. Which model do you expect had the highest MSE on the training set? Which had the lowest?

g. Suppose the true relationship between X and Y is quadratic. Which model do you think would be most accurate?

---

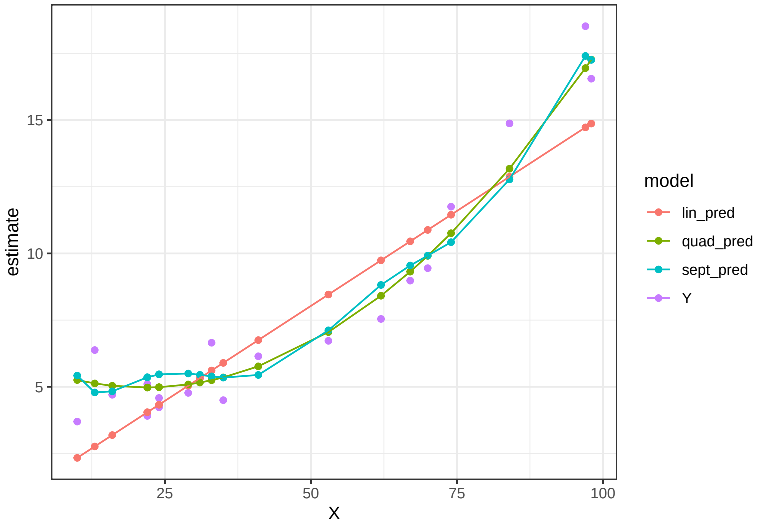
a.

```r
hw1_p5 <- read_csv(file = "hw1_p5.csv")
```

b.

```r
data_long <- hw1_p5 %>% pivot_longer(!X, names_to = "model", values_to = "estimate")


data_long %>%
  ggplot(aes(x = X, y = estimate, color = model))+
  geom_point()+
  geom_line(data = data_long %>%filter(model != "Y")  )+
  theme_bw()
```



c. Based on the graph, the quadratic model appears to best predict the response overall.

d.

```
sq_err <- function(y,x){
  (y - x)^2
}
```

```
hw1_p5 %>% mutate(lin_err = sq_err(Y,lin_pred),
                  quad_err = sq_err(Y,quad_pred),
                  sept_err = sq_err(Y,sept_pred)) %>%
  summarize(lin_mse = mean(lin_err),
            quad_mse = mean(quad_err),
            sept_err = mean(sept_err))
```

```
## # A tibble: 1 x 3
##   lin_mse quad_mse sept_err
##     <dbl>    <dbl>    <dbl>
## 1    2.76    0.850     1.18
```

e. The linear model had the largest MSE, while the quadratic model had the lowest MSE.

f. Since the septic model has the greatest flexibility, it likely has the lowest training MSE. The linear model likely has the greatest training MSE.

g. Assuming the true relationshop is quadratic, we would indeed expect to see the quadatic model to be most accurate.

Problem 6

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`. Read about the data set. By running the following code chunk. Note that the code options include `echo = F` so that the code chunk isn't printed in the .pdf output, and include `eval = F` so that the code is not run when knitting to .pdf.

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

(b) Make some (2-3) pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

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

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

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

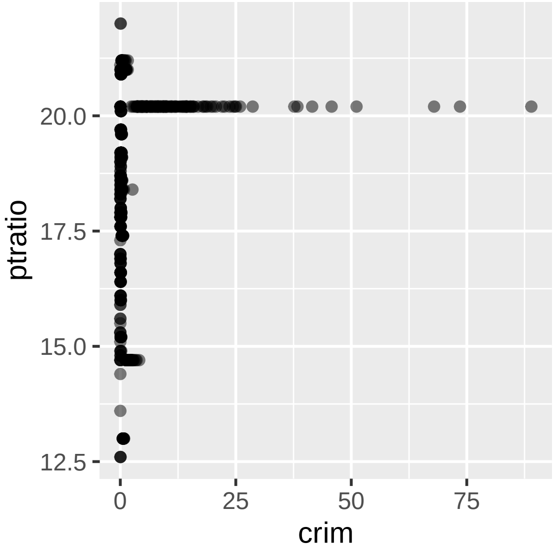
(f) What is the median pupil-teacher ratio among the towns in this data set?

(g) If you want to build a model to predict the average value of a home based on the other variables, what is your output/response? What is your input?

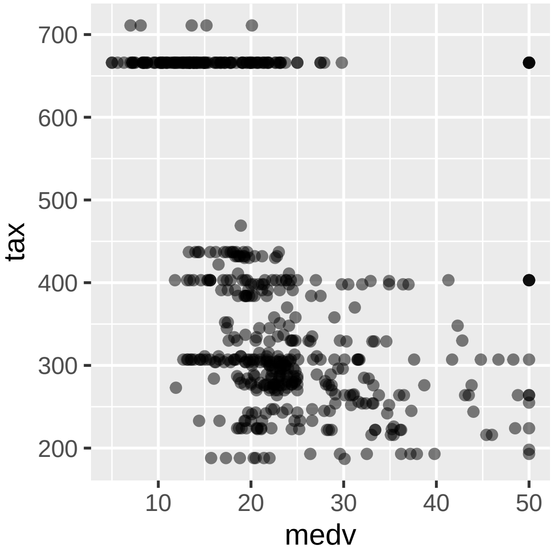
(a) The data set has 506 rows and 14 columns. Each row represents the data for a particular Boston town, while each column represents a variable measured.

(b) Descriptions may vary.

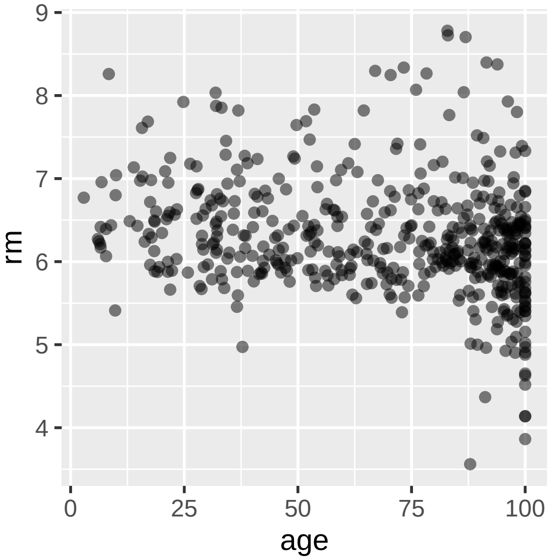
```
Boston %>%
  ggplot(aes(x = crim, y = ptratio)) + geom_point(alpha = .5)
```



```
Boston %>%
  ggplot(aes(x = medv, y = tax)) + geom_point(alpha = .5)
```



```
Boston %>%
  ggplot(aes(x = age, y = rm)) + geom_point(alpha = .5)
```



(c) We compute correlation for each predictor paired with crime rate:

```
cor(Boston)[1,]
```

```
##        crim          zn       indus        chas         nox          rm 
##  1.00000000 -0.20046922  0.40658341 -0.05589158  0.42097171 -0.21924670 
##         age         dis         rad         tax     ptratio       black 
##  0.35273425 -0.37967009  0.62550515  0.58276431  0.28994558 -0.38506394 
##       lstat        medv 
##  0.45562148 -0.38830461
```

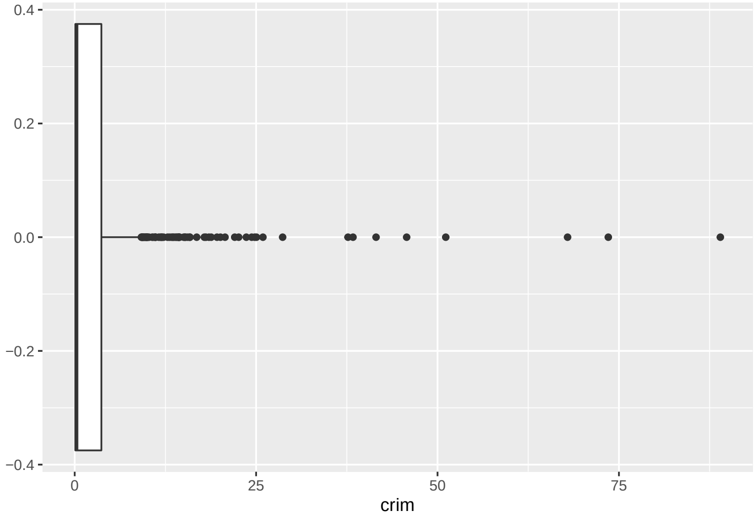
We could also compute scatterplots for each pair, but doing so and having the results display nicely with 14 predictors is somewhat tedious.

Base on our correlation vector, it appears the following predictors are positively correlated (r > .25) with Crime Rate

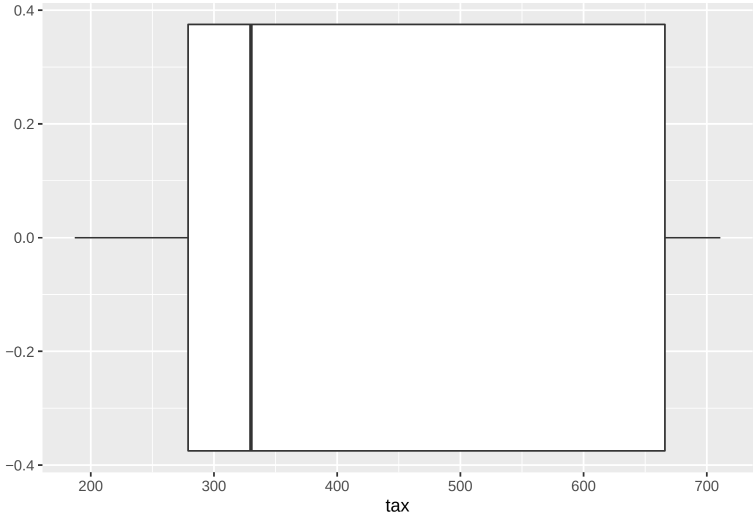
`indus`, `nox`, `age`, `rad`, `tax`, `ptratio`, `lstat`

(d)

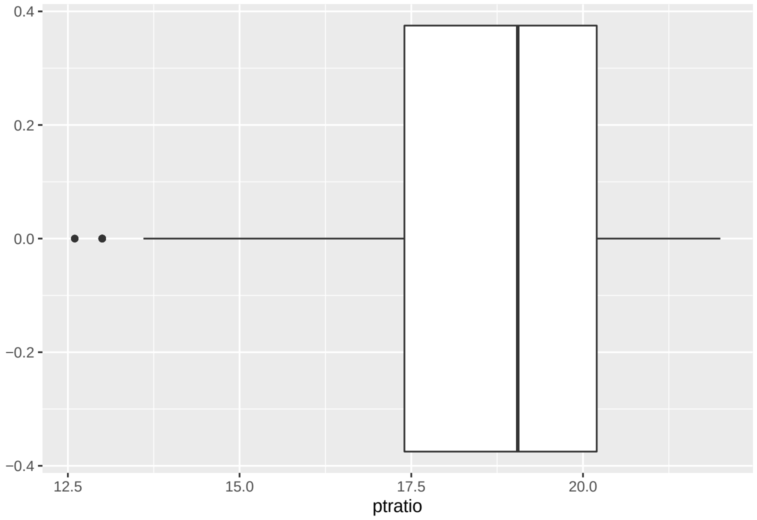
```
Boston %>%
  ggplot(aes( x = crim))+geom_boxplot()
```



```
Boston %>%
  ggplot(aes( x = tax))+geom_boxplot()
```



```
Boston %>%
  ggplot(aes( x = ptratio))+geom_boxplot()
```



Based on the boxplots, there seem to be a number of suburbs with crime rate 1.5 x IQR above the 3rd quartile. There do not appear to be noticeable (large) outliers in tax rate or pupil ration, although there are a few suburbs with noticeably low pupil-teacher ratio.

```
Boston %>%summarise_at(c("crim", "tax", "ptratio"), lst(quantile) )
```

```
##   crim_quantile tax_quantile ptratio_quantile
## 1      0.006320          187            12.60
## 2      0.082045          279            17.40
## 3      0.256510          330            19.05
## 4      3.677083          666            20.20
## 5     88.976200          711            22.00
```

Crime rate Most suburbs have crime rate between .08 and 3.67.

Tax Most suburbs have tax rate between 279 and 666.

Pupil Teacher Ratio Most suburbs have ptratio between 17.4 and 20.2

(e)

There are 35 suburbs bordering on the charles river.

```
Boston %>% summarize( n = sum(chas == 1))
```

```
##    n
## 1 35
```

(f)

The median ptratio is 19.05

```
Boston %>% summarize(M = median(ptratio))
```

```
##       M
## 1 19.05
```

(g) In a model predicting the average value of a home based on the other variables, the output for the linear function would be home value and the input would be the data vector for a particular suburb.