Andrew Rothman
anr248@stanford.edu

STATS 202 – Homework #1

October 5th, 2018
SUID: 06325118

Problem #1:

Exercise 2 from section 2.4 of ITSL textbook

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

*** Note, that the problem asks that for pasts A through C we state whether we are* ***most*** *interested in inference or prediction. I will assume the option "we are interested in both inference and prediction" is not a suitable option. I would argue that "part C" possibly falls into this category, depending on the interpretation of the wording of the problem. That being said, for Part C I choose the response I thought most appropriate assuming I can only choose one.*

(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. The outcome of interest is CEO salary, which is a continuous variable.
- In this scenario we are more interested in inference. From the problem statements, we are most interested in understanding the way that CEO salary is affected as firm profit, firm employee count, and firm industry change. Our goal is not particularly to make predictions for CEO salary, but rather to understand how CEO salary changes as a function of the predictors (firm profit, firm employee count, firm industry)
- n = 500 observations *(500 firms)*
- p = 3 *(firm profit, firm number of employees, and firm industry)*

(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. The outcome of interest is "success" or "failure" of a new product in the market place, which is a categorical (and in this case a binary) variable.
- In this scenario we are more interested in prediction. We are treated the functional form of the model as a "black-box". We are not particularly interested in understanding exactly how the outcome changes as a function of the predictors. Rather we are most interested in being able to predict the outcome as well as possible *(i.e. whether the new product will be a success or failure)*.
- n = 20 observations *(20 products)*
- p = 13 variables *(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 the British market, and the % change in the German market.

Answer:

- This is a regression problem. The outcome of interest is % change in the USD/Euro exchange rate, which is a continuous variable.
- In this scenario we are more interested in prediction. We are treated the functional form of the model as a "black-box". We are not particularly interested in understanding exactly how the outcome changes as a function of the predictors. Rather we are most interested in being able to predict the outcome as well as possible *(i.e. the % change in the USD/Euro exchange rate)*.
- n = 52 observations *(weekly data (52) for all of 2012)*
- p = 3 variables *(% change in the US market, the % change in the British market, and the % change in the German market)*

Problem #2:

Exercise 5 from section 2.4 of ITSL textbook

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:

The advantage of using a very flexible modeling approach (i.e. a nonparametric method) over a less flexible approach (i.e. a parametric method) is that you are not supplementing the analysis with as many rigid parametric assumptions about what the functional form of $\boldsymbol{f}$ actually is. Therefore you are not relying on those underlying parametric assumptions to hold-true in order for the analysis to recover accurate results and accurate parameter estimates. In particular, if the true functional form of $\boldsymbol{f}$ is highly complex and/or highly non-linear, and one is going into the analysis with very little prior expert subject-matter knowledge, attempting to a-priori specify the functional form of $\boldsymbol{f}$ using a rigid parametric approach can be very challenging. It is in these scenarios leveraging a very flexible modeling approach can be beneficial in the interest of recovering accurate results.

The disadvantage of using a very flexible modeling approach over a less flexible approach is that the results and estimated parameters, while theoretically more likely to be accurate due to fewer parametric assumptions required to assume to hold to recover accurate results, are in turn less likely to recover results as precise as those recovered from a less flexible modeling approach. This is often referred to as the “bias-variance tradeoff”. Essentially, because we are supplementing the analysis with far more parametric assumptions with a parametric approach, we are not relying on the data as much to inform the model specification of $\boldsymbol{f}$. This in turn typically leads to more “precise” results than one would have with a non-parametric approach.

Circumstances where we may prefer a very flexible approach are when we believe the true functional form of $\boldsymbol{f}$ is highly complex and/or highly non-linear, but we are not confident in our understanding or ability to specify $\boldsymbol{f}$ a priori. Circumstances where we may prefer a less flexible approach are when we are more confident in our prior-understanding of the functional form of $\boldsymbol{f}$, and are more comfortable supplementing the analysis with hard parametric assumptions with confidence. We may also prefer a less flexible approach if we are highly interested in being able to give estimated model parameters a descriptive and actionable interpretation. The more complex and non-linear models become, typically the harder it is to interpret the resulting estimated model parameters in a descriptive or actionable manner.

Problem #3:

Exercise 4 from section 2.4 of ITSL textbook

You will now 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 answer.

Answer:

1) Market research on soda vs non-soda drinkers
   - Taking a representative random sample among the United States population, assess how demographics affect the likelihood of being a soda-drinker or not.
   - The outcome variable is binary (soda drinker: 1, non-soda drinker: 0)
   - Predictor variables: age, gender, race, level of education, diet, socioeconomic status, region of the country where they live in, type of statistical area they live in (rural, suburb, urban)
   - The goal in this case is inference. We would like to understand how the likelihood of being a soda drinker changes as a function of the demographic related predictor variables
2) Predicting risk of breast cancer among USA residents
   - Develop a prediction model to assess the risk of breast cancer among USA residents
   - The outcome variable is binary (breast cancer: 1, no breast cancer: 0)
   - Predictor variables: race, age, gender, genetics, BMI, family history of cancer, history of tobacco use, use of NSAIDS, use of HRT, history of cancer screenings, dietary factors
   - The goal in this case is prediction. We are most interested in predicting the outcome (likelihood of breast cancer) as well as possible. We are not as concerned with the descriptive interpretability of model parameters or interpretability of the functional form of $\boldsymbol{f}$.
3) Predicting if a credit card transaction is fraudulent or not
   - Develop a prediction model to assess if a credit card transaction is fraudulent or not.
   - The outcome variable is binary (fraudulent: 1, not fraudulent: 0)
   - Predictor variables: credit history, past purchases, timing and frequency of credit-card use, size of purchase in dollars, location of credit-card use.
   - The goal in this case is prediction. We are most interested in predicting the outcome (likelihood a credit card transaction is fraudulent) as well as possible. We are not as concerned with the descriptive interpretability of model parameters or interpretability of the functional form of $\boldsymbol{f}$.

(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.

Answer:

1) Research on relationship between student demographics and SAT scores
    - Taking a representative random sample among the United States population of high school students, assess how demographics affect SAT score performance
    - The outcome variable is SAT score (continuous variable)
    - Predictor variables: age, gender, race, high-school rank, private school indicator, level of education, socioeconomic status, region of the country where they live in, type of statistical area they live in (rural, suburb, urban)
    - The goal in this case is inference. We would like to understand how the predicted SAT performance changes as a function of the demographic related predictor variables
2) Research on track-athlete performance in the long-jump
    - Taking a representative random sample among the United States track long-jumpers, assess how predictors affect long-jump performance
    - The outcome variable is long-jump performance (continuous variable: distance in feet)
    - Predictor variables: age, gender, race, BMI, body fat percentage, blood pressure, height, years of practice, physical fitness regimen
    - The goal in this case is inference. We would like to understand how the predicted long-jump performance changes as a function of the demographic related predictor variables
3) Research on the relationship of demographics and income among white-collar workers
    - Taking a representative random sample among the United States white-collar workers, assess how predictors affect predict yearly income
    - The outcome variable is yearly income (continuous variable: US dollars)
    - Predictor variables: age, gender, race, level of education, degree category, industry, job status (part time, contractor, full-time employee), years of employment at present company
    - The goal in this case is inference. We would like to understand how the predicted income changes as a function of the demographic related predictor variables.

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

Answer:

1) Customer segmentation at Walmart
    - Clustering analysis on Walmart customers to identify distinct groups within Walmart's customer base. This could be informative to later determining where and how to spend targeted marketing dollars. Variables in the cluster analysis may include: age, gender, race, level of education, socioeconomic status, region of the country where they live in, type of statistical area they live in (rural, suburb, urban)
2) Cluster patients in a clinic
    - Clustering analysis of patients in a medical clinic to identify distinct groups with similar symptoms. This could be informative in later determining the most effective interventions to provide and how to provide said interventions in a targeted manner. Variables in the cluster analysis may include: race, age, gender, genetics, current symptoms, drug-use, medical history
3) Segment students in a school system
    - Clustering analysis of students in a school system to determine distinct clusters of students. This could help determine which clusters of students need special attention or special needs. Variables in the cluster analysis may include: age, gender, race, grades, socioeconomic status, region of the country where they live in, primary language.

Problem #4:

Exercise 7 from section 2.4 of ITSL textbook

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

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

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

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

Answer:

Obs.	X_1	X_2	X_3	Y	**Euclidean distance from $X_1 = X_2 = X_3 = 0$**
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	3.162278
4	0	1	2	Green	2.236068
5	-1	0	1	Green	1.414214
6	1	1	1	Red	1.732051

(b) What is our prediction with K=1? Why?

Answer:

The prediction for K=1 is **green**. We wish to predict the outcome “Y” for the point (0,0,0) using K-nearest neighbors with K=1. So among the “K” nearest points in the training set to (0,0,0), we assign the class “Y” with the highest percentage as the prediction. With K=1, this means we are only using 1 point. The point in the training set with the closest Euclidean distance to (0,0,0) is observation “5” shown below. Because the “Y” for observation 5 is green, this is what we specify as the prediction of the outcome for point (0,0,0).

Obs.	X_1	X_2	X_3	Y	**Euclidean distance from** $X_1 = X_2 = X_3 = 0$
5	-1	0	1	Green	1.414214

(c) What is our prediction with K=3? Why?

Answer:

The prediction for K=3 is **red**. We wish to predict the outcome “Y” for the point (0,0,0) using K-nearest neighbors with K=3. Among the “K” nearest points in the training set to (0,0,0), we assign the class “Y” with the highest percentage as the prediction. The three points in the training set with the closest Euclidean distance to (0,0,0) are observations “5, 6, and 2” shown below. Because the majority (2/3) of the “Y” for these observations are red, this is what we specify as the prediction of the outcome for point (0,0,0).

Obs.	X_1	X_2	X_3	Y	**Euclidean distance from** $X_1 = X_2 = X_3 = 0$
5	-1	0	1	Green	1.414214
6	1	1	1	Red	1.732051
2	2	0	0	Red	2

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the *best* value for K to be large or small? Why?

Answer:

In this case we would expect the best value of K to be small. If the true decision boundary is highly non-linear, we would want to implement K-nearest neighbors in a way that provides a great deal of flexibility. This corresponds to a small K. As K becomes larger in K-nearest neighbors, the estimated decision boundary becomes more linear and less flexible. The ability of K-nearest neighbors to be very flexible and perform well estimating highly non-linear decision boundaries when K is small is due to the prediction for any given point being based on only a small number of local observations in the training data as opposed to a lot or even a predominant percentage of the training data.

<u>R-code:</u>

```
### load libraries
library(dplyr)

### create dataset
ID = c(1,2,3,4,5,6)
x1 = c(0,2,0,0,-1,1)
x2 = c(3,0,1,1,0,1)
x3 = c(0,0,3,2,1,1)
y = c('red','red','red','green','green','red')

df = data.frame(ID,x1,x2,x3,y)

####################################
####################################
## Part A:                        ##
## Compute the Euclidean distance ##
## between each observation and   ##
## the test point, x1=x2=x3=0     ##
####################################
####################################
point = c(0,0,0)
df$EucDist = sqrt(((df$x1 - point[1])^2) + ((df$x2 - point[2])^2) + ((df$x3 - point[3])^2))

###########################################
###########################################
## Part B:                               ##
## What is our prediction with K=1? Why? ##
###########################################
###########################################
k=1
arrange(df, EucDist)[1:k,]

###########################################
###########################################
## Part C:                               ##
## What is our prediction with K=3? Why? ##
###########################################
###########################################
k=3
arrange(df, EucDist)[1:k,]
```

Problem #5:

Exercise 10 from section 2.4 of ITSL textbook

This exercise involves the **Boston** housing data set.

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

Answer:

The Boston dataset has 506 rows and 14 columns. The rows represent the 506 unique observations (unique suburbs), and the columns represent the 14 variables.

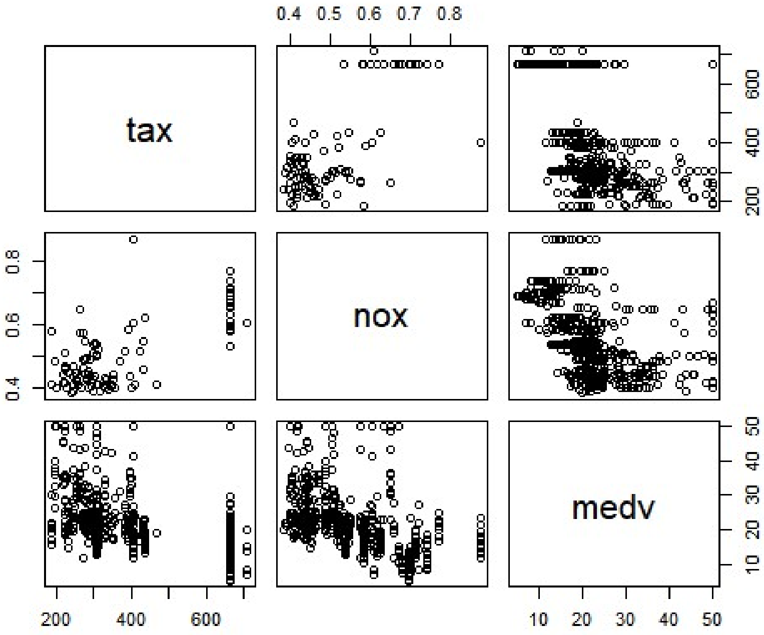
(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

Answer:

For the purposes of this question, I did not produce scatterplots for combinations of all 14 variables. From the wording of the problem, I assumed I was to choose just a few variables of interest. I choose three continuous variables:

- TAX – full-value property-tax rate per $10,000
- NOX - nitric oxides concentration (parts per 10 million)
- MEDV - Median value of owner-occupied homes in $1000's

From the three scatterplots below, we can see that TAX has a somewhat direct relationship with NOX, and TAX has an indirect relationship with MEDV. Additionally, NOX appear to have an indirect relationship with MEDV.



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

Answer:

Given the ambiguity and non-specificity of this question, I performed a very simple set of analyses. I ran a univariate OLS regression of CRIM (per capita crime rate) on each of the 13 other variables in the dataset. No multivariate analyses were performed. In terms of assessing whether there was an “association” in each model, I referred to the “statistical significance” (with α=0.05) of the associated p-value for the estimated regression coefficient in each model. Note that consideration as to whether effect sizes of statistically significant associations were large enough to be substantively relevant were not considered. Results of those univariate analyses are shown below:

Univariate OLS Regression results (outcome CRIM)

Variable	Estimated β	p-value
ZN	-0.07393	5.51E-06
INDUS	0.50978	< 2e-16
CHAS	-1.8928	0.209
NOX	31.249	< 2e-16
RM	-2.684	6.35E-07
AGE	0.10779	2.85E-16
DIS	-1.5509	<2e-16
RAD	0.61791	< 2e-16
TAX	0.029742	<2e-16
PTRATIO	1.152	2.94E-11
BLACK	-0.03628	<2e-16
LSTAT	0.5488	< 2e-16
MEDV	-0.36316	<2e-16

At an alpha threshold of 0.05, all univariate OLS analyses for outcome CRIM were statistically significant except for CHAS. Of the remaining 12 univariate analyses that produced statistically significant results:

- Variables ZN, RM, DIS, BLACK, and MEDV were all found to have in inverse relationship with CRIM
- Variables INDUS, NOX, AGE, RAD, TAX, PTRATIO, and LSTAT were all found to have a direct relationship with CRIM

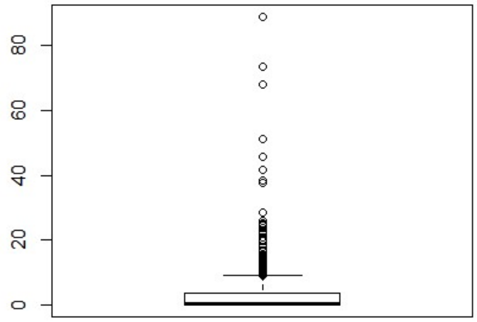
(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

Answer:

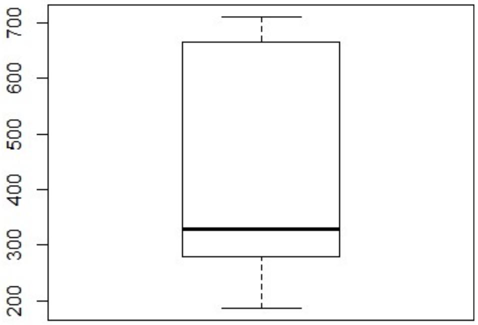
Descriptive statistics of the variables in question are below:

Variable	Min	25th percentile	Median	Mean	75th percentile	Max
CRIM	0.00632	0.08204	0.25651	3.61352	3.67708	88.9762
TAX	187	279	330	408.2	666	711
PTRATIO	12.6	17.4	19.05	18.46	20.2	22

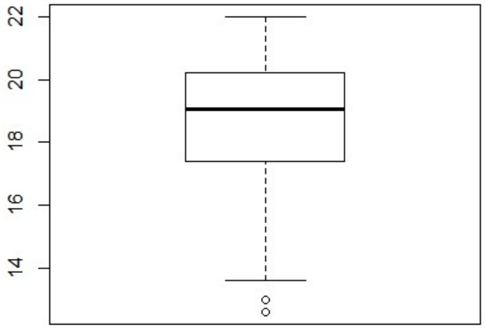
Box plots are below:



per capita crime rate



full-value property-tax rate per $10,000



pupil-teacher ratio by town

Looking at the table and box plots above, yes some of the towns have very high crime rates (in the case of a few observations, >60 per capita crime rate). The CRIM variable is very right skewed. The TAX variable on the other hand is left skewed, with the mean tax rate per $10,000 being 408, the median being 330, and the range being 187 – 711. The distribution of the Pupil-teacher ratio is fairly symmetrical, with the median being 19, mean 18.5, and range 12.6 – 22.

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

Answer:
There are 35 suburbs in the dataset that bound the Charles river.

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

Answer:
The median pupil-teacher ratio among the towns in this data set is 19.05

(g) Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

Answer:
There are actually two observations with the lowest median value of owner occupied homes (that being MEDV=5). The data for these two observations are shown below:

	observation 1	observation 2
CRIM	38.3518	67.9208
ZN	0	0
INDUS	18.1	18.1
CHAS	0	0
NOX	0.693	0.693
RM	5.453	5.683
AGE	100	100
DIS	1.4896	1.4254
RAD	24	24
TAX	666	666
PTRATIO	20.2	20.2
BLACK	396.9	384.97
LSTAT	30.59	22.98
MEDV	5	5

Compared to the overall ranges for the predictors, these two observations have:

- Higher than average values for CRIM, INDUS, NOX, AGE, RAD, TAX, PTRATIO, BLACK, LSTAT
- Lower than average values for ZN, CHAS, RM, DIS, MEDV

(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.

Answer:

- There are 64 suburbs that average more than seven rooms per dwelling
- There are 13 suburbs that average more than eight rooms per dwelling

To just touch on some comments, of the 13 suburbs that average more than eight rooms per dwelling, they have a low per capita crime rate (mean 0.71879), have older homes (mean AGE=71.54), and a very high median value of owner-occupied homes in $1000's (mean MEDV=44.2).

R-code:

```
### load libraries
library(dplyr)
library(MASS)
library(Hmisc)
library(ggplot2)

## Part A
df <- Boston
head(df)
dim(df)

## Part B
pairs(dplyr::select(df, tax, nox, medv))

## Part C
summary(lm(crim ~ zn, data=df))
summary(lm(crim ~ indus, data=df))
summary(lm(crim ~ chas, data=df))
summary(lm(crim ~ nox, data=df))
summary(lm(crim ~ rm, data=df))
summary(lm(crim ~ age, data=df))
summary(lm(crim ~ dis, data=df))
summary(lm(crim ~ rad, data=df))
summary(lm(crim ~ tax, data=df))
summary(lm(crim ~ ptratio, data=df))
summary(lm(crim ~ black, data=df))
summary(lm(crim ~ lstat, data=df))
summary(lm(crim ~ medv, data=df))

## Part D
summary(df$crim)
boxplot(df$crim, xlab="per capita crime rate")
summary(df$tax)
boxplot(df$tax, xlab="full-value property-tax rate per $10,000")
summary(df$ptratio)
boxplot(df$ptratio, xlab="pupil-teacher ratio by town")

## Part E
table(df$chas, exclude=NULL)

## part F
summary(df$ptratio)

## part G
arrange(df, medv)

## Part H
arrange(df, desc(rm))
df_7 <- filter(df, rm>7)
df_8 <- filter(df, rm>8)
summary(df_8)
summary(df)
```