MATH 4322 Homework 1 Solutions

Cathy Poliak

9/9/2021

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
Regression
Inference
$n = 500$
$p = 4$

(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
Classification
Prediction
$n = 20$
$p = 14$

(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 Regression
Prediction
$n = 52$ (weeks) $p = 4$

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

Answer This is subjective. Everyone will have different answers. Remember *classification* requires the response variable (output) to be categorical, the *regression* problem requires the response variable (output) to be quantitative, and the *cluster* problem does not have a response variable.

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?

Advantages: Can fit many different possible functional forms for f. **Disadvantage**: Requires a greater number of parameters. Also, it is harder to interpret f. **Why use a more flexible approach?**: if ware are only interested in prediction and not concerned about the interpretabiltiy of the model. **Why use less flexible approach?**: if we are mainly interested in inference (interpretation of f).

Problem 4

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

(a) Which of the predictors are quantitative, and which are qualitative?
(b) What is the *range* of each quantitative predictor? You can answer this using the `range()` function.
(c) What is the mean and standard deviation of each quantitative predictor?
(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?
(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.
(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.

Answer

(a) *mtcars* data

```
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Quantitative variables: mpg, disp, hp, drat, wt, qsec
Categorical variables: cyl,vs, am, gear, carb

(b) Range

```
mtcars.quant<- mtcars[c(1,3,4,5,6,7)]
sapply(mtcars.quant,max) - sapply(mtcars.quant,min)
```

```
##     mpg    disp      hp    drat      wt    qsec
##  23.500 400.900 283.000   2.170   3.911   8.400
```

(c) Mean and standard deviation

```
sapply(mtcars.quant,mean)
```

```
##        mpg       disp         hp       drat         wt       qsec
##  20.090625 230.721875 146.687500   3.596563   3.217250  17.848750
```

```
sapply(mtcars.quant,sd)
```

```
##         mpg        disp          hp        drat          wt        qsec
##   6.0269481 123.9386938  68.5628685   0.5346787   0.9784574   1.7869432
```

variable	Mean	Standard Deviation
mpg	20.09	6.027
disp	230.72	123.94
hp	146.69	68.56
drat	3.60	0.53
wt	3.22	0.98
qsec	17.85	1.79

(d) Without the 10th through 32 observations

```
mtcars.quant2 <- mtcars.quant[1:9,]
sapply(mtcars.quant2,max) - sapply(mtcars.quant2,min) #range
```

```
##    mpg   disp     hp   drat     wt   qsec
##  10.10 252.00 183.00   1.16   1.25   7.06
```

```
sapply(mtcars.quant2,mean) #mean
```

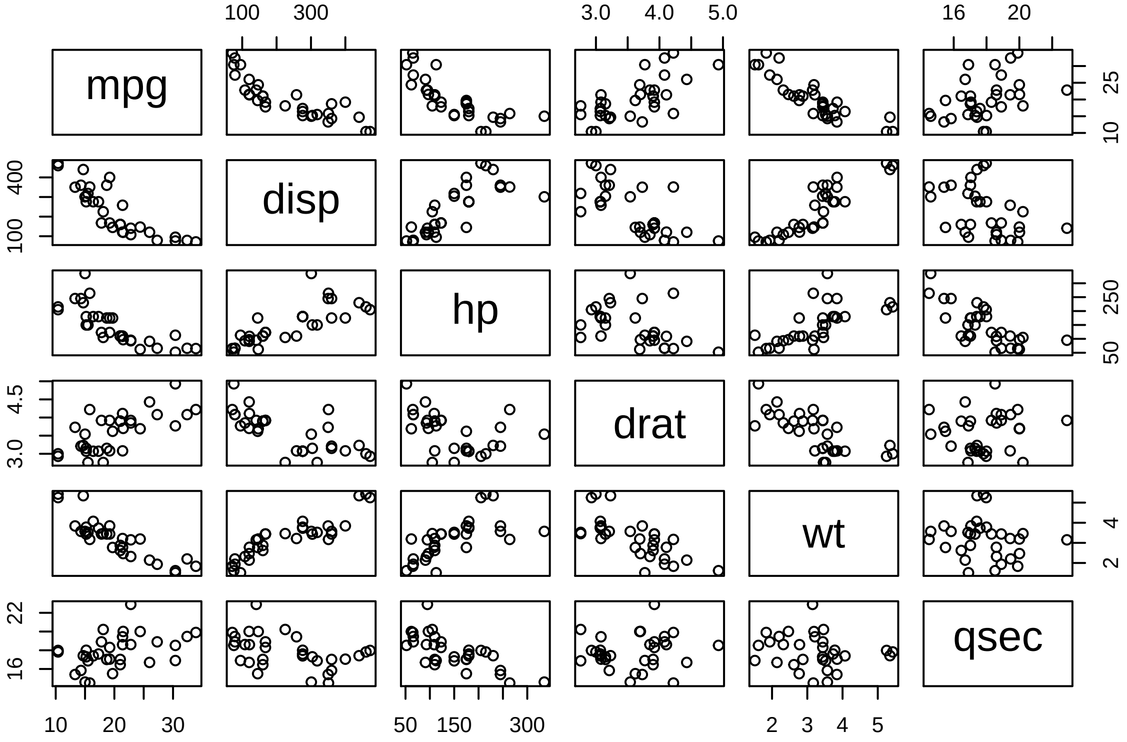
```
##        mpg       disp         hp       drat         wt       qsec
##  20.500000 213.166667 122.777778   3.495556   3.093333  18.612222
```

```
sapply(mtcars.quant2,sd) #standard deviation
```

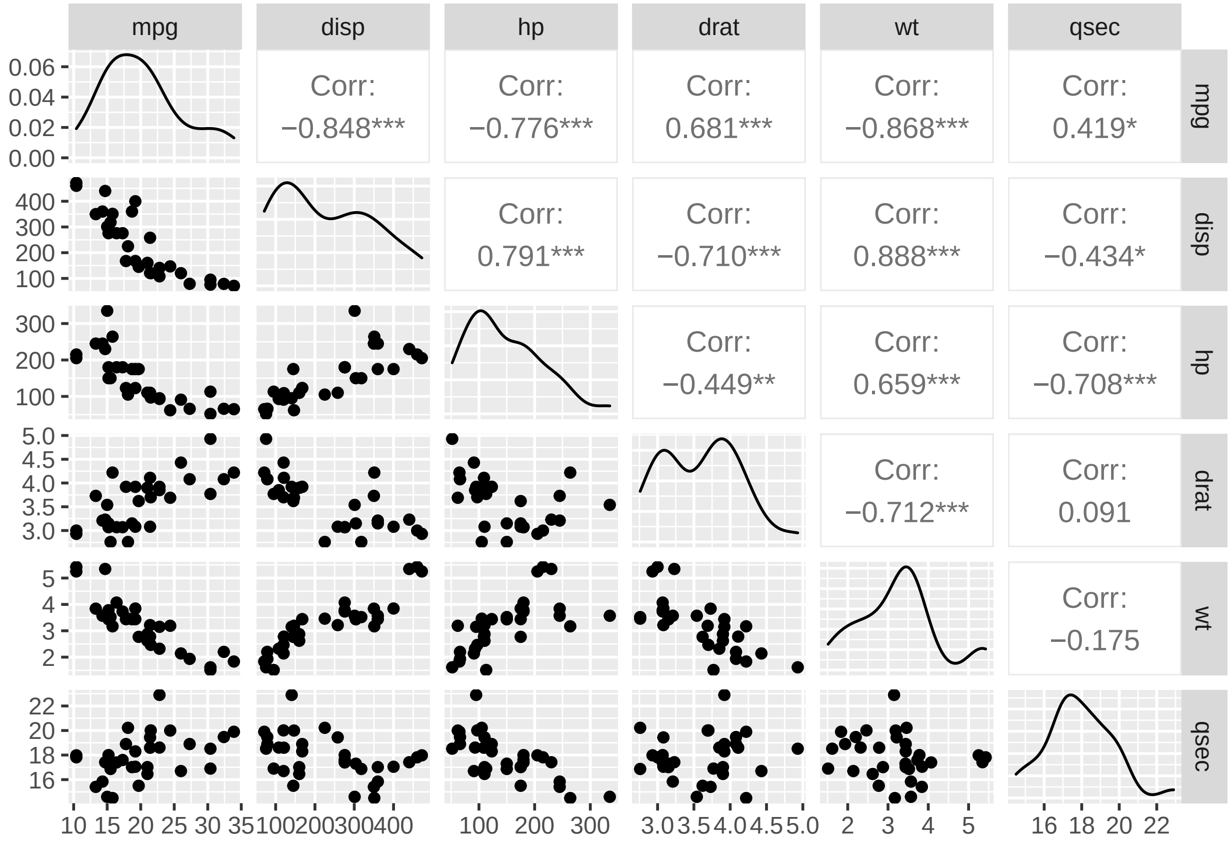
```
##        mpg       disp         hp       drat         wt       qsec
##  3.0524580 94.6297258 54.5705456  0.4451716  0.4151732  2.2629835
```

(e) Graphical functions

```
#You only need to do one of these plots.
pairs(mtcars.quant)
#install.packages("ggplot2")
#install.packages("GGally")
library(ggplot2)
```



```
library(GGally)
ggpairs(mtcars.quant) #This includes the correlations
```



Observations:
mpg appears to be associated with the all of the variables except *qsec*.
It appears that the only variable not associated with the other variables is *qsec*.

(f) From the plot it appears that we might be able to use all execpt *qsec*.

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:

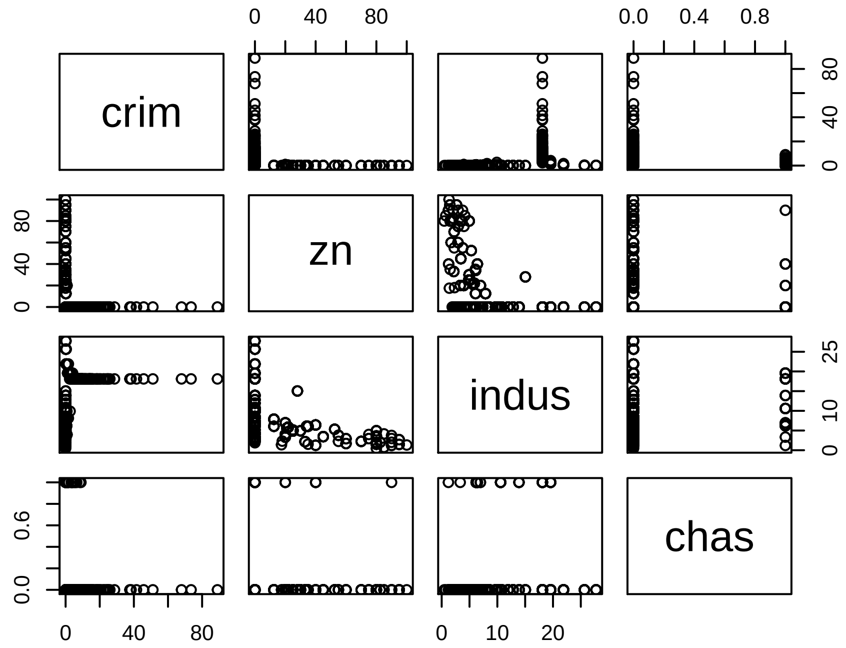
```
#?Boston
```

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

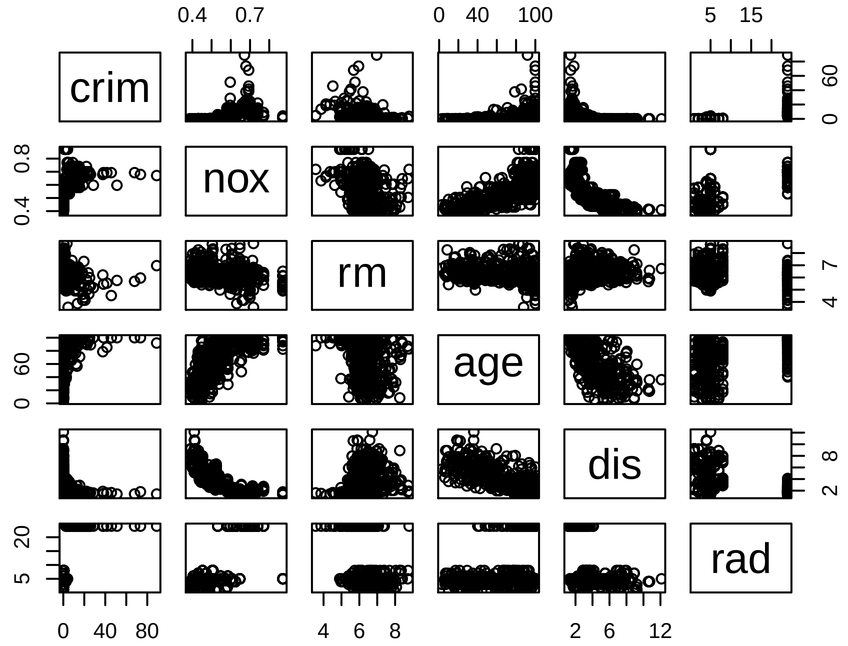
Answer
There are 506 rows - number of observations, n.
There are 14 columns - number of variables, p.

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

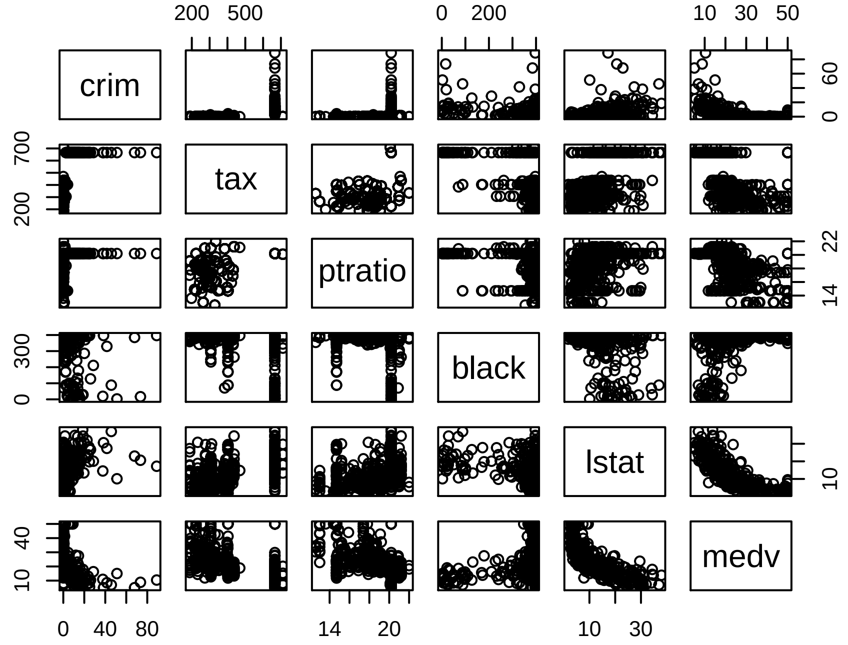
```
pairs(Boston[,1:4])
```



```
pairs(Boston[,c(1,5:9)])
```



```
pairs(Boston[,c(1,10:14)])
```



There appears to be some relationship between *lstat* and *medv*, *nox* and *age*, *nox* and *dis*, *crim* and *nox*, *crim* and *age*, *crim* and *dis*, *nox* and *rm*, *indus* and *crim*, *crim* and *chas*, and *zn* and *indus*. Although the plots are very hard to look at.

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

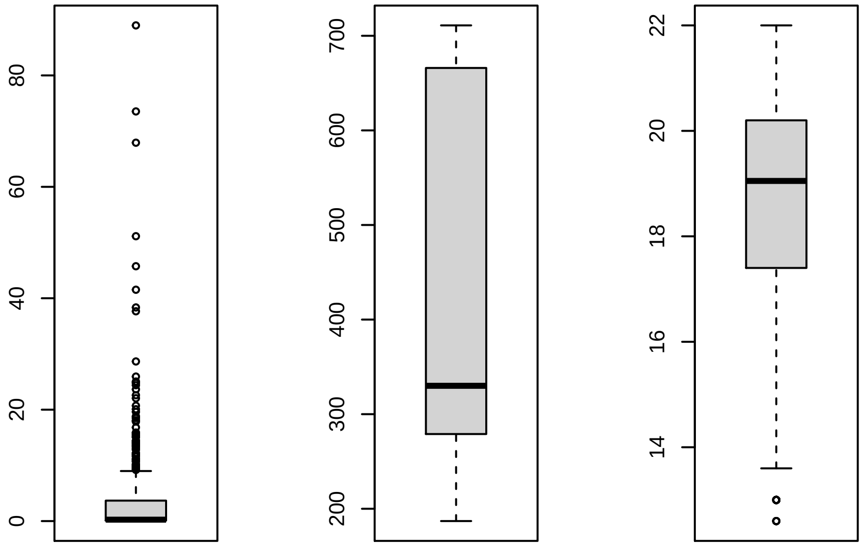
```
sort(cor(Boston)[1,])
```

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

The strongest correlation with crime appears to be *rad*. All of the other variables do not appear to have an association with crime.

(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.

```
par(mfrow=c(1,3))
boxplot(Boston$crim)
boxplot(Boston$tax)
boxplot(Boston$ptratio)
```



```
max(Boston$crim) - min(Boston$crim)
```

```
## [1] 88.96988
```

```r
max(Boston$tax) - min(Boston$tax)
```

```
## [1] 524
```

```r
max(Boston$ptratio) - min(Boston$ptratio)
```

```
## [1] 9.4
```

Looking at the box plots it appears that some suburbs have unusually high crime. The tax and parent to teacher ratio do not appear to not have high values. However, the parent to teacher ratio is extremely lower in some of the suburbs.

The following gives you which row (neighborhood) has the highest crime and the lowest parent teacher ratio.

```r
which.max(Boston$crim)
```

```
## [1] 381
```

```r
which.min(Boston$ptratio)
```

```
## [1] 197
```

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

```r
sum(Boston$chas)
```

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

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

```r
median(Boston$ptratio[Boston$chas == 1])
```

```
## [1] 17.6
```

(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.

```r
m = which.min(Boston$medv)
m
```

```
## [1] 399
```

```r
rbind(Boston[m,],sapply(Boston,median))
```

```
##         crim zn indus chas   nox     rm   age     dis rad tax ptratio  black
## 399 38.35180  0 18.10    0 0.693 5.4530 100.0 1.48960  24 666   20.20 396.90
## 2    0.25651  0  9.69    0 0.538 6.2085  77.5 3.20745   5 330   19.05 391.44
##     lstat medv
## 399 30.59  5.0
## 2   11.36 21.2
```

Here are some notable observations comparing this neighborhood to the median.
Crime rate is high.
Proportion of non-retail business acre per town is high.
All of the houses are built before 1940.
Index of accessibility to radial highways is high.
Tax is higher.
Percent of lower status of the population is high.
Median value of a home is low.

(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.

```
length(Boston$rm[Boston$rm > 7]) #More than 7 rooms
```

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

```
length(Boston$rm[Boston$rm > 8]) #More than 8 rooms
```

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

```
rbind(sapply(Boston[Boston$rm > 8,],median),sapply(Boston,median))
```

```
##          crim zn indus chas   nox     rm  age     dis rad tax ptratio  black
## [1,] 0.52014  0  6.20    0 0.507 8.2970 78.3 2.89440   7 307   17.40 386.86
## [2,] 0.25651  0  9.69    0 0.538 6.2085 77.5 3.20745   5 330   19.05 391.44
##      lstat medv
## [1,]  4.14 48.3
## [2,] 11.36 21.2
```

First row is the median values with homes with more than 8 rooms. Second row is median values of all the neighborhood. Some observations.

Crime rate is higher than the median.
Percent of lower status of the population is low.
Median value of homes in that neighborhood is high.