Chapter 2 HW

Fabiani Rafael

Conceptual Questions

Exercise 2: Explain whether each scenario is a classification or regression prob-lem, and indicate whether we are most interested in inference or pre-diction. Finally, provide n and p.

(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.

(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 prod- uct 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 the British market, and the % change in the German market.

Exercise 4: You will now think of some real-life applications for statistical learn-ing.

(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.

Applied Questions

Exercise 8: This exercise relates to the College data set, which can be found in the file College.csv on the book website. It contains a number of variables for 777 different universities and colleges in the US. The variables are • Private : Public/private indicator • Apps : Number of applications received • Accept : Number of applicants accepted • Enroll : Number of new students enrolled • Top10perc : New students from top 10 % of high school class • Top25perc : New students from top 25 % of high school class2.4 Exercises 55 • F.Undergrad : Number of full-time undergraduates • P.Undergrad : Number of part-time undergraduates • Outstate : Out-of-state tuition • Room.Board : Room and board costs • Books : Estimated book costs • Personal : Estimated personal spending • PhD : Percent of faculty with Ph.D.'s • Terminal : Percent of faculty with terminal degree • S.F.Ratio : Student/faculty ratio • perc.alumni : Percent of alumni who donate • Expend : Instructional expenditure per student • Grad.Rate : Graduation rate Before reading the data into R, it can be viewed in Excel or a text editor. (a) Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data. (b) Look at the data using the View() function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

rownames(college) <- college[, 1] View(college)

You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try > college <- college[, -1] > View(college) Now you should see that the first data column is Private. Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row. (c) i. Use the summary() function to produce a numerical summary of the variables in the data set.56 2. Statistical Learning ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10]. iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private. iv. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %. > Elite <- rep("No", nrow(college)) > Elite[college$Top10perc > 50] <- "Yes" > Elite <- as.factor(Elite) > college <- data.frame(college , Elite) Use the summary() function to see how many elite univer- sities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite. v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative vari- ables. You may find the command par(mfrow = c(2, 2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways. vi. Continue exploring the data, and provide a brief summary of what you discover.

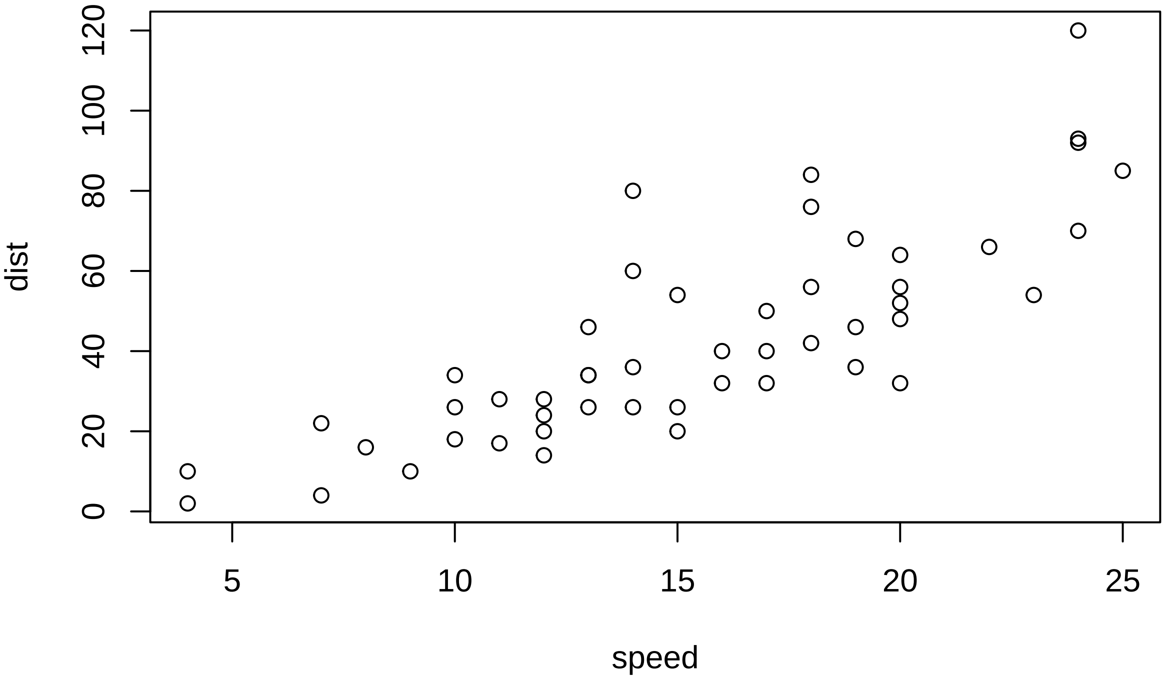
Exercise 10 This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the ISLR2 library. > library(ISLR2) 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?
(b) Make some 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) Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.
(e) How many of the census tracts in this data set bound the Charles river?
(f) What is the median pupil-teacher ratio among the towns in this data set?
(g) Which census tract of Boston has lowest median value of owner- occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.
(h) In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

```r
plot(cars)
```


dist
0
20
40
60
80
100
120
5
10
15
20
25
speed