

STAT387 HOMEWORK#1

INTRODUCTION TO STATISTICAL LEARNING

Saeah Go

Due January 14, 12:00PM

Conceptual

1. 2(a, b, and c)

2. 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 .

(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.

This scenario is a **regression** problem, since the response (the CEO salary) is quantitative(continuous). And we are most interested in **inference** because we would like to figure out the relations how predictors affect CEO salary, not predicting something. n , the number of observations is 500, and p , the number of predictors is 3 (profit, number of employees, and 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.

In this scenario, we want to make sure a new launching product will be a success or failure. Since the response is a binary value (success or failure), it is categorical, thus this is a **classification** problem. Since we want to predict a result about products, we are most interested in **prediction**. In this case, n is 20 and p is 13 (price charged for the product, marketing budget, competition price, and ten other variables).

(c) We are interesting in predicting the % change in the US dollar 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 dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

The scenario (c) is a **regression** problem, since the response (the % change in the US dollar in relation to the weekly changes in the world stock

markets) is quantitative(continuous) values. And in this scenario, we are interested in predicting the response, thus this is a **prediction** problem. n in this case is 52, and p is 3 (the % change in the US market, the % change in the British market, the % change in the German market).

2. 4(a and b)

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.

1. Stock price: Classify whether a stock will go up or down in price the next day given a set of financial data & news from the preceding week. The aim is to make a prediction. The response would be stock market result, (go up or down), and the predictors can be the dividend-price ratio, the earnings growth rate, and the price-earnings ratio growth rate, etc.

2. House value: Consider a real estate setting where one would like to relate values of homes to variables such as quality of schools, crime rates, closeness to a park. Say we are interested in predicting the value of a house given its characteristics. Since we want to predict a house value, the aim is to make a prediction. The response is values of homes, and the predictors are quality of schools, crime rates, and closeness to a park.

3. Weather: Say we want to see if Portland will be rain or not tomorrow. We are interested in the thing that it will rain or not, so our goal is to make a prediction. The response will be the result (rain or not), and the predictors can be the atmospheric temperature (if temperature is below 32, it's possible to have snow rather than rain), wind speed, wind direction, amount of clouds (no clouds, then probably no rain!).

(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. Advertising data: Consists of sales of particular product in 200 different markets, together with advertising budgets for the product in TV, radio, and newspaper. The goal is to figure out how to allocate the advertising budgets in the three media. The response would be the sales, and the predictors will be budgets for TV, radio, and newspaper. Since Y (sales) is continuous values, regression might be useful for this problem. The goal of this application is inference since we want to see the relationship between Y and X .

2. House value: Consider a real estate setting where one would like to relate values of homes to variables such as quality of schools, crime rates, closeness to a park. Let's say we are interested in how the predictors affect the price of the house. We want to understand the relationship between the Y (house value) and the X (the elements we assumed these might affect house values), thus inference is the goal. The response is values of homes, and the predictors are quality of schools, crime rates, closeness to parks.

3. Salary: Let's say I am thinking of changing jobs (companies). And say I got a job offer and want to see if the offered salary is higher, lower, or about to average. This case our goal is inference because we want to look how the predictors affect the salary. The response is salary, and the predictors are years of experience, education level, industry, location (big city, suburban or rural area).

Applied

3. 9(a, b, c, d, e, f)

This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

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

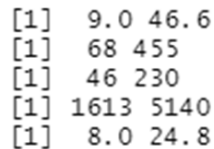
Quantitative Variables are the variables whose values result from counting or measuring something. (Continuous) Qualitative Variables are the variables that are not measurement variables. (Categorical) Their values do not result from measuring or counting.

Thus, the quantitative predictors are: mpg, displacement, horsepower, weight, and acceleration.

And, the qualitative predictors are: cylinders, year, origin, and name.

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

I used R to answer this question.



```
[1] 9.0 46.6
[1] 68 455
[1] 46 230
[1] 1613 5140
[1] 8.0 24.8
```

The range of mpg is $46.6 - 9.0 = 37.6$

The range of displacement is: $455 - 68 = 387$

The range of horsepower is: $230 - 46 = 184$

The range of weight is: $5140 - 1613 = 3527$

The range of acceleration is: $24.8 - 8.0 = 16.8$

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

I also used R to answer this question. I rounded to the nearest 0.01.

```

[1] 23.44592
[1] 194.412
[1] 104.4694
[1] 2977.584
[1] 15.54133
[1] 7.805007
[1] 104.644
[1] 38.49116
[1] 849.4026
[1] 2.758864
      mpg displacement horsepower weight acceleration
23.44592 194.41199 104.46939 2977.58418 15.54133
[1] 7.805007 104.644004 38.491160 849.402560 2.758864

```

The mean of mpg is: 23.45

The mean of displacement is: 194.41

The mean of horsepower is: 104.47

The mean of weight is: 2977.58

The mean of acceleration is: 15.54

The standard deviation of mpg is: 7.81

The standard deviation of displacement is: 104.64

The standard deviation of horsepower is: 38.49

The standard deviation of weight is: 849.40

The standard deviation of acceleration is: 2.76

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

I used R for this question again.

```

'data.frame': 392 obs. of 9 variables:
 $ mpg      : num 18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders: num 8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num 307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower: num 130 165 150 150 140 198 220 215 225 190 ...
 $ weight    : num 3504 3693 3436 3433 3449 ...
 $ acceleration: num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year      : num 70 70 70 70 70 70 70 70 70 70 ...
 $ origin    : num 1 1 1 1 1 1 1 1 1 1 ...
 $ name      : Factor w/ 304 levels "amc ambassador brougham"...: 49 36 231 14 161 141 54 223 241 2 ...
      mpg cylinders displacement horsepower weight acceleration year origin
[1] 24.368454 5.381703 187.753943 100.955836 2939.643533 15.718297 77.132492 1.599369
[1] 7.8808983 1.6581348 99.9394881 35.8955668 812.6496293 2.6938126 3.1100263 0.8193079
      mpg cylinders displacement horsepower weight acceleration year origin
[1,] 11.0 3 68 46 1649 8.5 70 1
[2,] 46.6 8 455 230 4997 24.8 82 3

```

The range of mpg is $46.6 - 11.0 = 35.6$

The range of cylinders is: $8 - 3 = 5$

The range of displacement is: $455 - 68 = 387$

The range of horsepower is: $230 - 46 = 184$

The range of weight is: $4997 - 1649 = 3348$

The range of acceleration is: $24.8 - 8.5 = 16.3$

The range of year is: $82 - 70 = 12$

The range of origin is: $3 - 1 = 2$

The mean of mpg is: 24.37

The mean of cylinders is: 5.38

The mean of displacement is: 187.75

The mean of horsepower is: 100.96

The mean of weight is: 2939.64

The mean of acceleration is: 15.72

The mean of year is: 77.13

The mean of origin is: 1.60

The standard deviation of mpg is: 7.88

The standard deviation of cylinders is: 1.66

The standard deviation of displacement is: 99.93

The standard deviation of horsepower is: 35.90

The standard deviation of weight is: 812.65

The standard deviation of acceleration is: 2.69

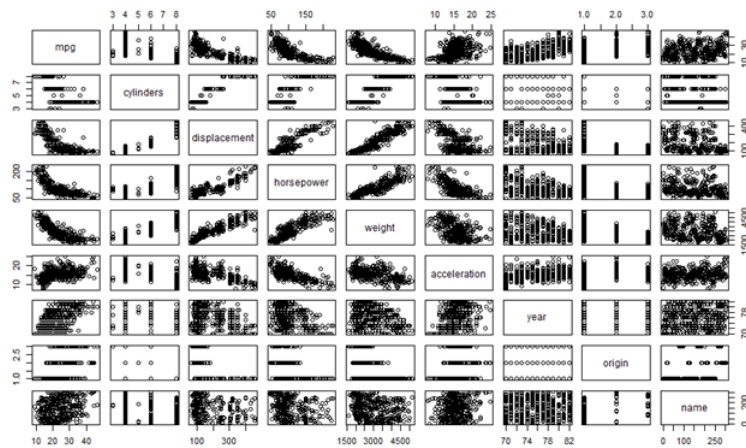
The standard deviation of year is: 3.11

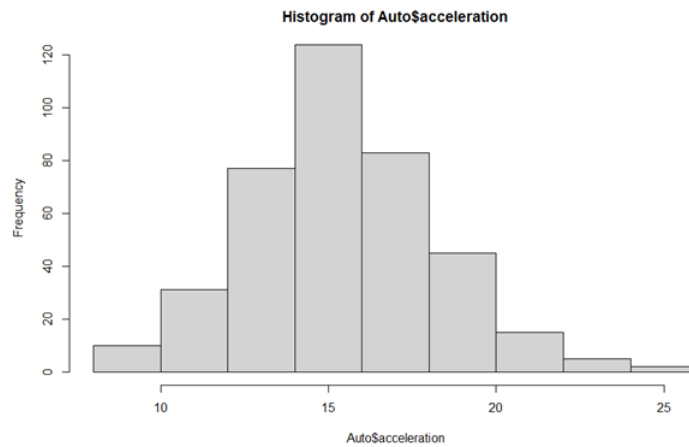
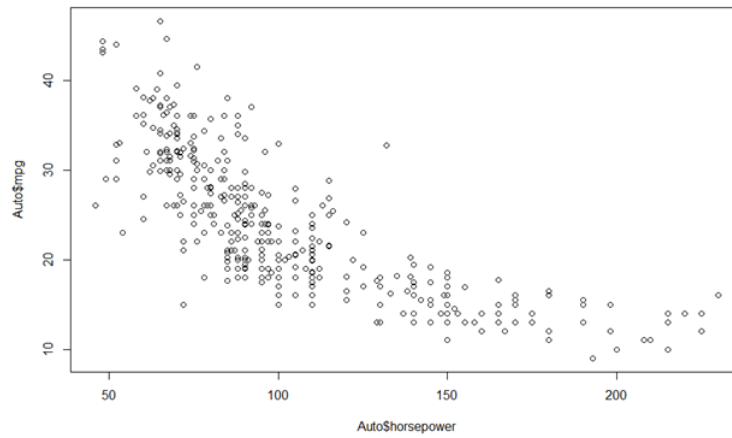
The standard deviation of origin is: 0.82

(e) Using the full data set, investigate the predictors graphically, using scatter plots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

I used R to create some plots.

Plots





My findings

- * The histogram for 'acceleration' resembles a normal distribution.
- * From the pairs(Auto), I could find that 'displacement' and 'weight' have a strong linear relationship.
- * 'mpg' has a non-linear relationship with 'displacement', 'horsepower', and 'weight'.

(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.

Yes. My plots suggest that year, horsepower, weight, and displacement might be useful in predicting mpg. I noticed these variables have positive or negative relationships to the mpg outcome. For example, I could notice that year and mpg have a positive relationship (As the year increases, the mpg also tends to increase). Also I could find that horsepower and mpg appear to have a

negative relationship (As the horsepower increases, the mpg tends to decrease). Similarly, weight and mpg, have a negative relationship too. (As the weight of car increases, the mpg tends to decrease). Displacement and mpg also have a negative relationship. (As the displacement increases, the mpg tends to decrease).

R Code

```
1 # Applied
2 ## 3. 9 (a, b, c, d, e, f)
3
4 #####
5 ## Problem (a)
6 #####
7 data(Auto, package = "ISLR") # load the data
8 fix(Auto) # since I already have loaded the ISLR package with the "
           # library" command, I don't need to use "read.table" command to
           # load the "Auto" data. It is already loaded in R. I can view the
           # file using the command "fix(Auto)".
9
10
11 #####
12 ## Problem (b)
13 #####
14 range(Auto$mpg)
15 range(Auto$displacement)
16 range(Auto$horsepower)
17 range(Auto$weight)
18 range(Auto$acceleration)
19
20
21 #####
22 ## Problem (c)
23 #####
24 # Method 1 (simply use mean() and sd() functions)
25 mean(Auto$mpg)
26 mean(Auto$displacement)
27 mean(Auto$horsepower)
28 mean(Auto$weight)
29 mean(Auto$acceleration)
30
31 sd(Auto$mpg)
32 sd(Auto$displacement)
33 sd(Auto$horsepower)
34 sd(Auto$weight)
35 sd(Auto$acceleration)
36
37 # Method 2 (use colMeans() and colSds())
38 colMeans(Auto[,c(1, 3:6)])
39 colSds(as.matrix(Auto[,c(1, 3:6)])) # need to make the data to
           # matrix form
40
41 #####
42 ## Problem (d)
```

```

43 #####
44 # drop rows using slice() function in the dplyr package
45 new_Auto <- Auto %>% slice(-c(10:84)) # remove from 10th to 85th
    observations
46
47 # check the structure of the Auto dataset
48 str(Auto) # since name's structure is factor (not numeric values),
    I cannot get mean and standard deviation with the name column.
49
50 # find means and standard deviations of only numeric columns (
    except name)
51 colMeans(new_Auto[,c(1:8)])
52 colSds(as.matrix(new_Auto[,c(1:8)]))
53 sapply(new_Auto[, c(1:8)], range)
54
55
56 #####
57 ### Problem (e)
58 #####
59 pairs(Auto) # method 1
60 pairs(~ mpg + cylinders + displacement + horsepower + weight +
    acceleration + year + origin + name, Auto) # method 2
61
62 plot(Auto$horsepower, Auto$mpg)
63 hist(Auto$acceleration)

```