Math 189: Midterm Project 1 Solution

Introduction

In this first midterm project for Math 189 the Motor Trend Car Road Tests dataset is analyzed. The goal is to answer the question: What is the relationship between weight (wt) and miles per gallon (mpg)? Does this relationship depend on the number of cylinders (cyl)? We will study these questions using techniques learned in the first 6 lectures of the course.

The dataset was provided in the course materials (https://github.com/tuckermcelroy/ma189), and was extracted from the 1974 Motor Trend US magazine. The data contains measurements on several aspects of automobile design and performance for 32 automobiles (1973–74 models).

Analysis

We begin by reading in the data, and taking a look. There are 11 variables, not including the labels of the cars.

```
cars <- read.table("mtcars.csv",sep = ",", header=TRUE)
head(cars)
```

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

We can estimate the mean and variability of the mpg and wt variables, to get an idea of these features. The sample mean vector and sample variance matrix are unbiased (so long as the data are identically distributed).

```
cars_all <- cars[,c(7,2)]
colMeans(cars_all)
```

```
##       wt      mpg
##  3.21725 20.09062
```

```
var(cars_all)
```

```
##            wt       mpg
## wt   0.957379 -5.116685
## mpg -5.116685 36.324103
```

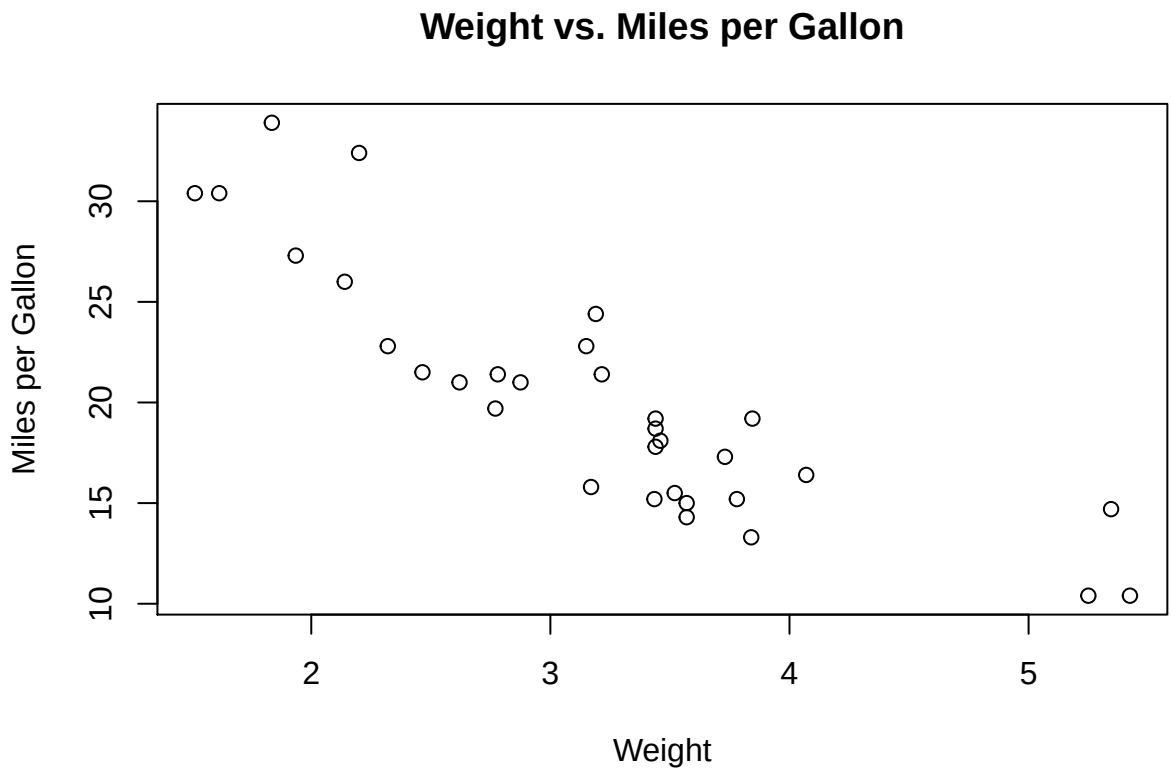
We see there is a negative association between wt and mpg, but in order to understand the strength of association we compute the sample correlation.

```
cor(cars_all)
```

```
##             wt        mpg
## wt   1.0000000 -0.8676594
## mpg -0.8676594  1.0000000
```

There seems to be a high negative association between mpg and wt, because the correlation is -0.8676594. We can also examine a scatterplot.

```
plot(x = cars$wt, y = cars$mpg,
     xlab = "Weight", ylab = "Miles per Gallon",
     main = "Weight vs. Miles per Gallon")
```



However, we have not yet considered the role of cylinders. We can update our analysis by taking these cylinders into account.

```
cars4 <- cars[cars$cyl==4,c(7,2)]
cars6 <- cars[cars$cyl==6,c(7,2)]
cars8 <- cars[cars$cyl==8,c(7,2)]
cor(cars4)
```

```
##             wt        mpg
```

```
## wt   1.0000000 -0.7131848
## mpg -0.7131848  1.0000000
```

```
cor(cars6)
```

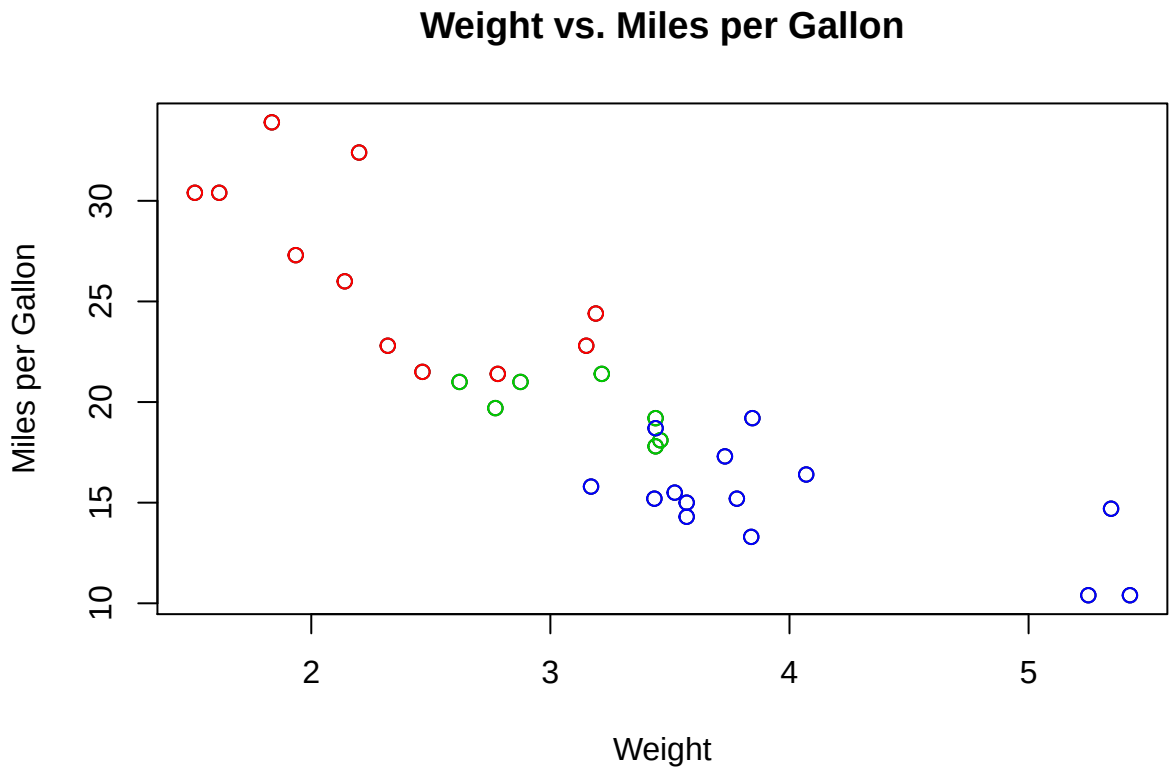
```
##             wt        mpg
## wt   1.0000000 -0.6815498
## mpg -0.6815498  1.0000000
```

```
cor(cars8)
```

```
##            wt       mpg
## wt   1.000000 -0.650358
## mpg -0.650358  1.000000
```

We see that the correlations are all negative, but differ depending on the cylinders. The updated scatterplot is below. We colored 4-cylinder red, 6-cylinder green, and 8-cylinder blue. The blue and red data are fairly well-separated, although the green is somewhat mingled with both sub-populations.

```
plot(x = cars$wt, y = cars$mpg,
     xlab = "Weight", ylab = "Miles per Gallon",
     main = "Weight vs. Miles per Gallon")
points(x = cars4[,1], y= cars4[,2],col = 2)
points(x = cars6[,1], y= cars6[,2],col = 3)
points(x = cars8[,1], y= cars8[,2],col = 4)
```



We may well expect the population mean vectors for wt and mpg to differ by cylinder. We compute the sample mean vectors for the three sub-populations, and normalize by the standard error (standard deviation divided by square root sample size) to make them easier to compare.

```
colMeans(cars4)/sqrt(diag(var(cars4))/nrow(cars4))
```

```
##       wt      mpg
## 13.31001 19.60901
```

```
colMeans(cars6)/sqrt(diag(var(cars6))/nrow(cars6))
```

```
##       wt      mpg
## 23.14379 35.93552
```

```
colMeans(cars8)/sqrt(diag(var(cars8))/nrow(cars8))
```

```
##       wt      mpg
## 19.70450 22.06952
```

These are the t-statistics for wt and mpg, for each sub-population, testing whether the mean is zero. Unsurprisingly, they are all highly significant.

Summary

There is empirical evidence that the size of cylinder (whether 4, 6, or 8) has an impact on the relationship between wt and mpg. The scatterplot exhibits three clusters with little overlap, and the sample correlations differ as well. Although we have not tested for significant differences between the estimates, the above analysis indicates that we should treat the data as consisting of three sub-populations. This is also supported by the normalized sample mean vectors for the three cylinder sub-populations.