Midterm Project 1

Jerry Chan

Introduction

In this report, I explore the Motor Trend Car Road Tests dataset and analyze the relationship between weight (wt), miles per gallon (mpg), and the number of cylinders (cyl). In the first section, I exam the distribution and relationship of wt and mpg. Next, I check if the distribution of wt and mpg depends on cyl. Last, I run a permutation test to see if the relationship of wt and mpg depends on cyl.

Data Description

The motor trend car road tests dataset [1] contrains data extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).[2]

Metadata [3]

1. **model**: the car model
2. **mpg**: Miles/(US) gallon
3. **cyl**: Number of cylinders
4. **disp**: Displacement (cu.in.)
5. **hp**: Gross horsepower
6. **drat**: Rear axle ratio
7. **wt**: Weight (lb/1000)
8. **qsec**: 1/4 mile time
9. **vs**: V/S
10. **am**: Transmission (0 = automatic, 1 = manual)
11. **gear**: Number of forward gears
12. **carb**: Number of carburetors

Load Data

```
df = read.csv('mtcars.csv', header = TRUE)
head(df)
```

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

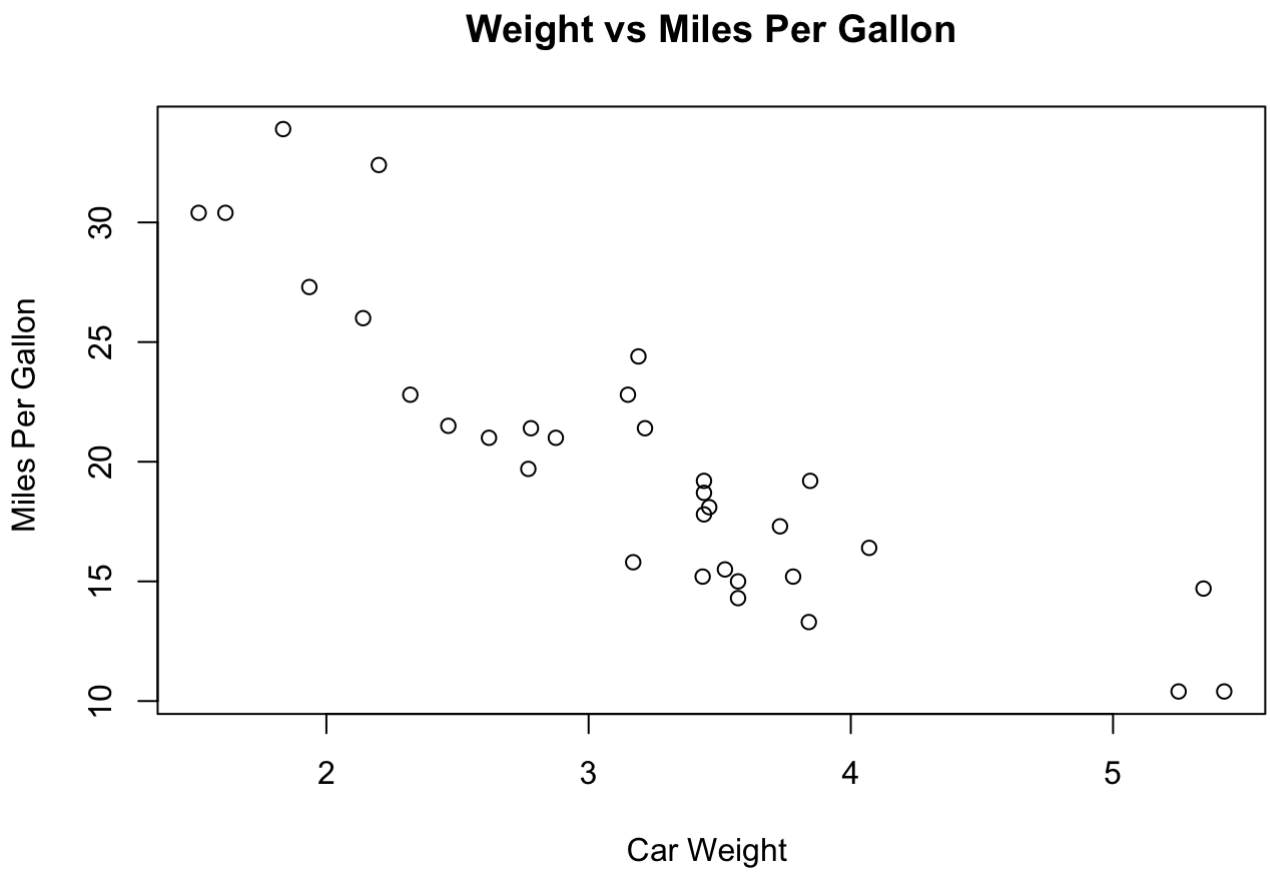
6 rows | 1-10 of 13 columns

Body

1. Weight vs Miles Per Gallon

First, I'm going to plot a scatter plot of Weight vs Miles Per Gallon to observe distribution.

```
plot(df$wt, df$mpg, main="Weight vs Miles Per Gallon",
   xlab="Car Weight ", ylab="Miles Per Gallon ")
```



From the plot we cn see that as the Car Weight increases, the Miles Per Gallon decreases and there doesn't seems to have any outliers.

Next, I use the code below to get the correlation of the two variable.

```
cor(df$wt, df$mpg)
```

```
## [1] -0.8676594
```

The function below fits a linear regression model to our data:

```
lm(df$mpg ~ df$wt)
```

```
##
## Call:
## lm(formula = df$mpg ~ df$wt)
##
## Coefficients:
## (Intercept)        df$wt
##      37.285       -5.344
```

2. Weight vs Miles Per Gallon vs Number of Cylinders

```
library(plotly)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

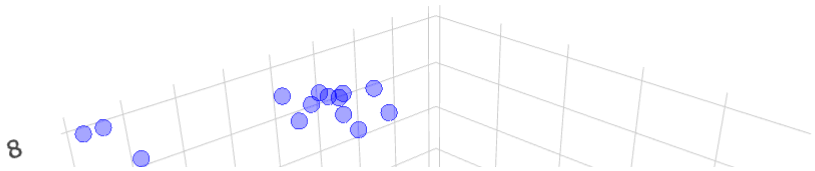
```
##
## Attaching package: 'plotly'
```

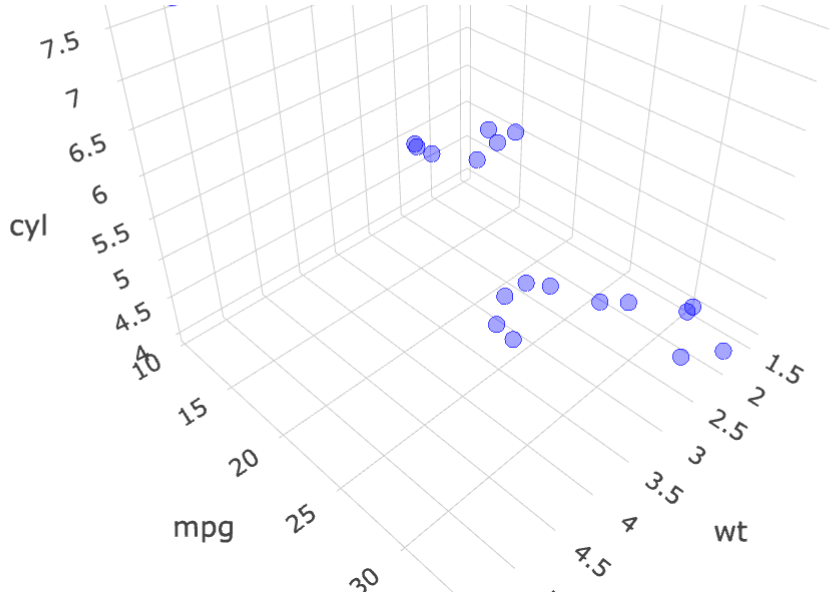
```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
fig <- plot_ly(df, x = ~wt, y = ~mpg, z = ~cyl, color = I("blue"), alpha = 0.5, size
 = 1)
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title = 'wt'),
                                   yaxis = list(title = 'mpg'),
                                   zaxis = list(title = 'cyl')))
fig
```

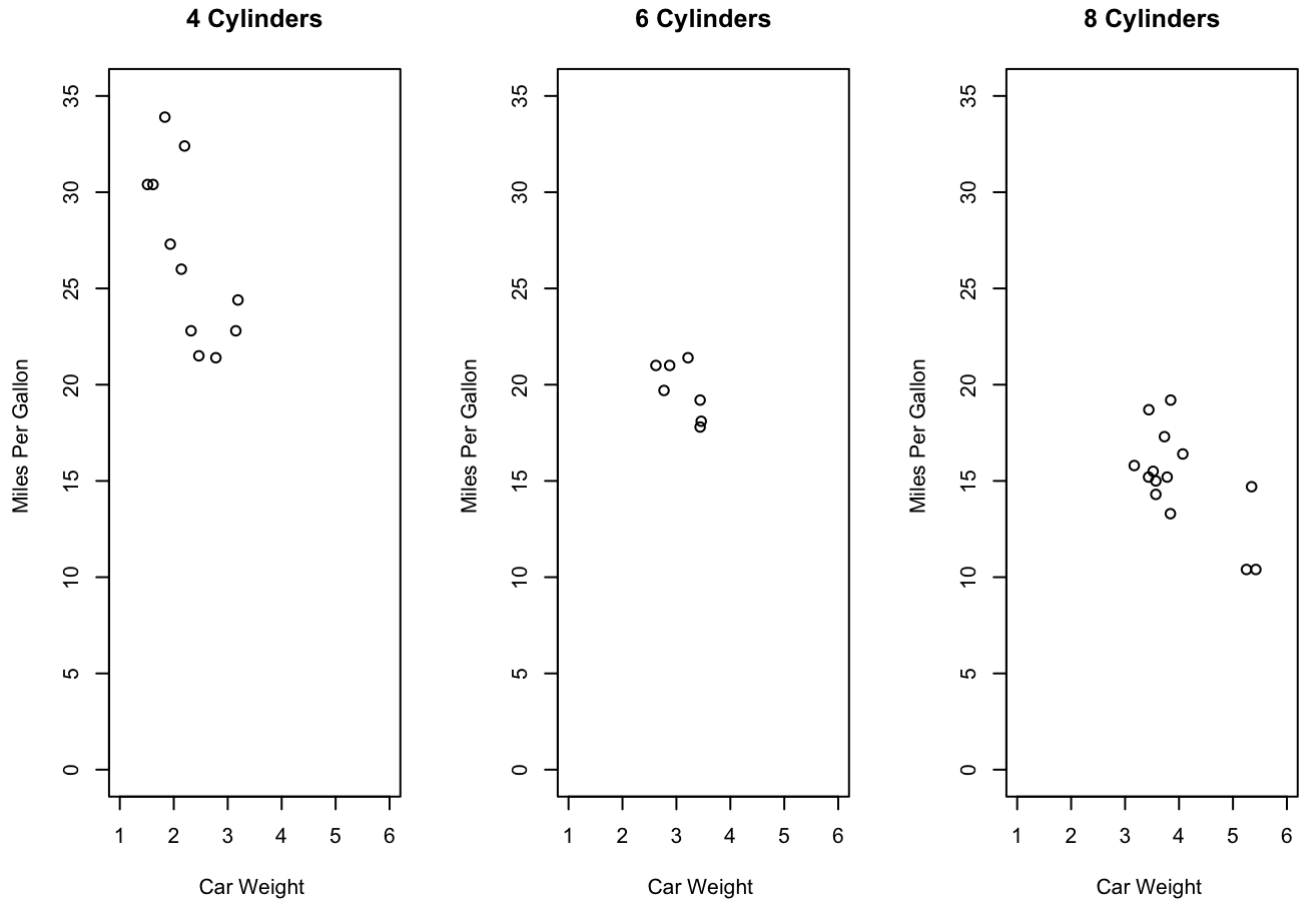




From the graph, we can see that there are only three possible values for Number of Cylinders, which are 4, 6, and 8. There are a clearly different distribution of Weight and Mules Per Gallon for different value of Number of Cylinders. We plot the wt vs mpg plot subject to each cyl and compute the grouped mean and variance of Weight and Miles Per Gallon.

Plots:

```
par(mfrow=c(1,3))
d1 = df[df$cyl == 4, ]
d2 = df[df$cyl == 6, ]
d3 = df[df$cyl == 8, ]
plot(d1$wt, d1$mpg, main="4 Cylinders",
   xlab="Car Weight ", ylab="Miles Per Gallon ",xlim=c(1, 6), ylim=c(0,35))
plot(d2$wt, d2$mpg, main="6 Cylinders",
   xlab="Car Weight ", ylab="Miles Per Gallon ",xlim=c(1, 6), ylim=c(0,35))
plot(d3$wt, d3$mpg, main="8 Cylinders",
   xlab="Car Weight ", ylab="Miles Per Gallon ",xlim=c(1, 6), ylim=c(0,35))
```



Grouped Mean:

```
aggregate(df[, c(2,7)], list(df$cyl), mean)
```

Group.1 <int>	mpg <dbl>	wt <dbl>
4	26.66364	2.285727
6	19.74286	3.117143
8	15.10000	3.999214

3 rows

Grouped Variance:

```
aggregate(df[, c(2,7)], list(df$cyl), var)
```

Group.1 <int>	mpg <dbl>	wt <dbl>
4	20.338545	0.3244028
6	2.112857	0.1269821
8	6.553846	0.5766956

3 rows

It's very clear that mpg and wt's distributions depends on cyl. The more Cylinders a car has, the less its mpg is and the heavier it weights. However this doesn't mean the relationship between mpg and wt depends on cyl. I will try to figure that out in the next section.

3. Does the relationship of Weight and Miles Per Gallon depends on the Number of Cylinders?

I choose to use a permutation test to test the dependency.

p value threshold: 0.05 Null Hypothesis: The relationship of wt and mpg doesn't depends on cyl. Therefore, the regression slope wt vs mpg of the grouped data should be about the same as the ungrouped data. Alternative Hypothesis: The relationship of wt and mpg does depends on cyl.

Under null hypothesis, the slope of wt and mpg should be -5.344 as the ungrouped data. So, I choose to use the mean absolute distance between (slope when cyl = 4, slope when cyl = 6, slope when cyl = 8) and (-5.344, -5.344, -5.344) as the test statistic.

Compute the observed statistic.

```
get_slopes <- function(df){
  d1 <- df[df$cyl == 4, ]
  d2 <- df[df$cyl == 6, ]
  d3 <- df[df$cyl == 8, ]
  s1 <- lm(d1$mpg~d1$wt)$coefficients[2]
  s2 <- lm(d2$mpg~d2$wt)$coefficients[2]
  s3 <- lm(d3$mpg~d3$wt)$coefficients[2]
  c(s1,s2,s3)
}
null_dist <- c(-5.344, -5.344, -5.344)
obs_dist <- get_slopes(df)
obs_stats  <- mean(abs( obs_dist- null_dist))
sprintf("observed slopes: %f, %f, %f", obs_dist[1],obs_dist[2],obs_dist[3])
```

```
## [1] "observed slopes: -5.647025, -2.780106, -2.192438"
```

```
sprintf("observed statistic = %f",obs_stats)
```

```
## [1] "observed statistic = 2.006160"
```

Next, we shuffle the label (cyl) and calculate the simulated statistic. This process is repeated for 10000 time.

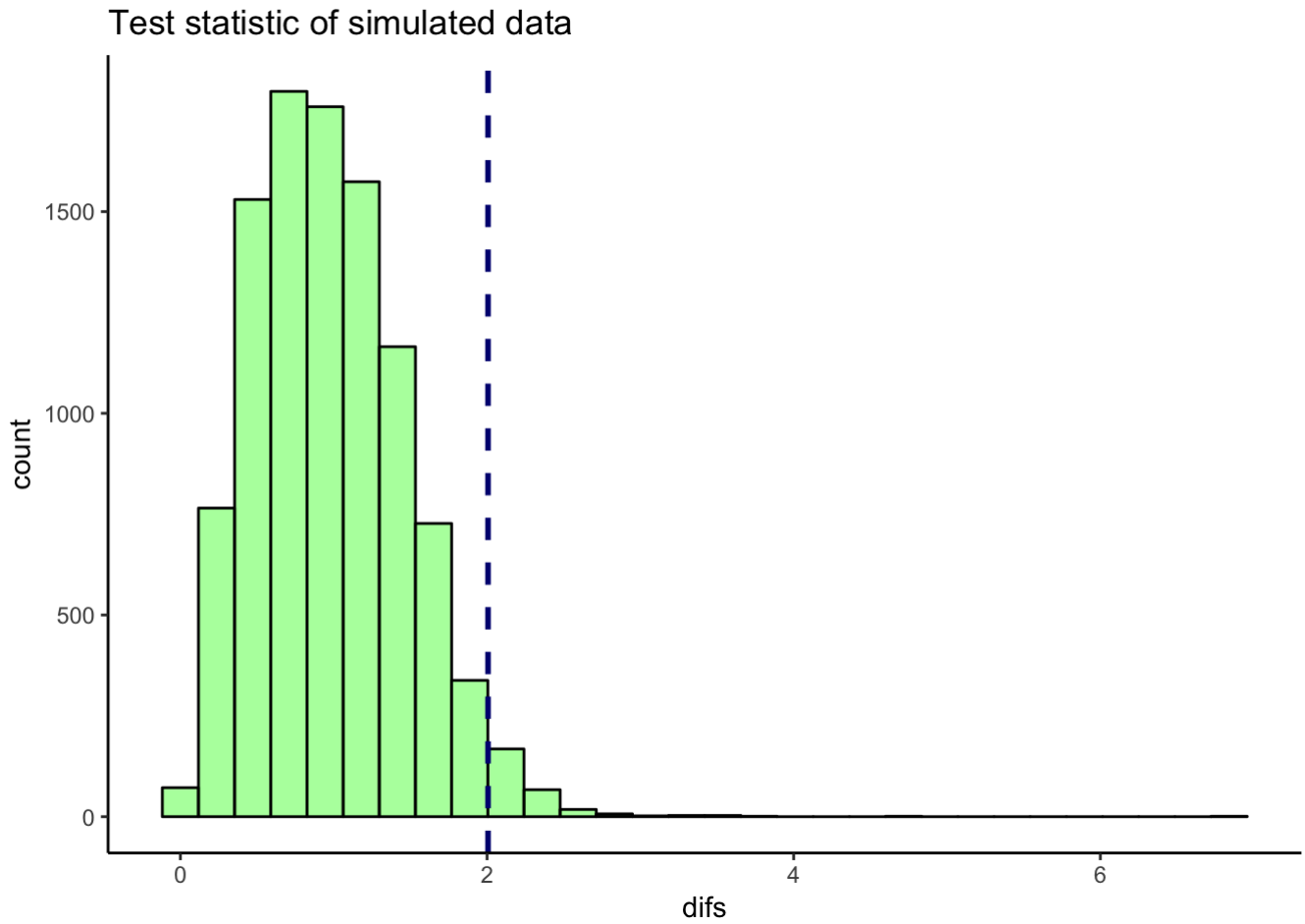
```
N = 10000
results<-vector('list',N)
for(i in 1:N){
  simulation <- transform( df, cyl = sample(cyl) )
  sim_dist <- get_slopes(simulation)
  results[[i]]<-  mean(abs( sim_dist- null_dist))
}
```

plot the results:

```
sim <- data.frame(difs = unlist(results))

ggplot(sim, aes(x=difs)) +
  geom_histogram(color="black", fill="green", alpha=.4) +
  geom_vline(color="navy",lwd=1,lty=2,xintercept = obs_stats) +
  theme_classic()+
  ggtitle("Test statistic of simulated data")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Calculate p value:

```
mean(results >= obs_stats)
```

```
## [1] 0.0269
```

Conclusion

1. In the first part, I use the linear regression model to estimate the relationship of wt and mpg. Wt and mpg have a high correlation of -0.86. The slope of wt vs mpg is -5.344 and the intercept is 37.285.
2. In the second part, we can see that the distribution of wt and the distribution of mpg depends on cyl. As cyl increases, mpg decreases and wt increases. This conclusion is based on grouped mean and variance and observing the distributions in scatterplots.

3. In the last part, I do a permutation test to see if the relationship of wt and mpg depends on cyl. The resulting p value is lower then the threshold I set. Therefore I reject the null hypothesis, which means the relationship depends on cyl.
4. The estimators I used in this project are sampled Mean and Variance and the slope of a linear regression model.

4.1 Sample mean is an unbiased estimator as $E[\bar{X}] = \mu$ where μ is the population mean. Samlpe variance is also an unbiased estimator. I'll not include the proof since I think this is not the focus of this course.

4.2 If we assume the true distribution of wt vs mpg follows a linear model, the slope of linear regression model is a unbiased estimator. [4]

---

1. Available from Math 189 Course GitHub↩
2. T. McElroy, "Ma189Homework2"↩
3. Christian, Motor Trend Car Road Analysis, Dec-2014. [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/51431_3323677bd16347fd983ba69d2aac5d64.html (https://rstudio-pubs-static.s3.amazonaws.com/51431_3323677bd16347fd983ba69d2aac5d64.html). [Accessed: 13-Jan-2021].↩
4. https://scholar.princeton.edu/sites/default/files/bstewart/files/lecture5handout.pdf (https://scholar.princeton.edu/sites/default/files/bstewart/files/lecture5handout.pdf)↩