Math 189: Midterm Project 1

Xiangyu Wei

January 24, 2021

Introduction

The Motor Trend Car Road Tests dataset contains information from the 1974 Motor Trend US magazine, of 32 observations on 11 variables. In this study, I aim to explore the relationship between weight (wt) and miles per gallon (mpg) by examining the scatterplot of the relationship, the sample means of each variable, the sample correlation/covariance between two variables, and a linear regression model. Moreover, I also want to see if this relationship is dependent on the number of cylinders (cyl). I will use the same methods mentioned before on each level of cyl, and compare the differences between the output statistic of each level to see if the relationship between wt and mpg has a dependence on cyl.

Tasks & Analysis

The following are the packages that are used in the study.

```
library(tidyverse)
```

Dataset

The Motor Trend Car Road Tests dataset was 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). It was also used in Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411. I extracted the dataset from https://github.com/tuckermcelroy/ma189/blob/main/Data/mtcars.csv at 2021-01-14 20:18:53 PST (when my groupmate and I did homework 2).

Load Dataset

```
mtcars <- read.csv('Data/mtcars.csv')
head(mtcars)
```

```
##               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
```

As we can see from the printed dataframe, the 2nd (mpg), 3rd (cyl), and 7th (wt) columns are what we need for this study.

Check Dataset

```
table(mtcars$cyl)
```

```
##
##  4  6  8
## 11  7 14
```

Upon investigation, it seems that cyl only has three variables, making it more like a categorical variable instead of a continuous numeric variable. Thus, I decided to covnert it into a factor using `as.factor()`.

Data Cleaning

```
mtcars <- mtcars %>%
  select(c(2,3,7)) %>%
  mutate(cyl.fact = as.factor(cyl))
```

I select 2nd, 3rd, and 7th columns because they are what's needed for this study. I add a column called `cyl.fact`, which converts the number of cylinder (cyl) into a factor because of my investigation above.

Method

To investigate the relationship between weight (wt) and miles per gallon (mpg), I will look at scatterplot of the relationship, the sample means of each variable, the sample correlation/covariance between two variables, and a linear regression model. To examine whether the wt and mpg relationship is dependent on the number of cylinders (cyl), I will use the same methods mentioned above on the relationship of wt and mpg within each level of cyl, and compare those statistics to see if there is a difference.

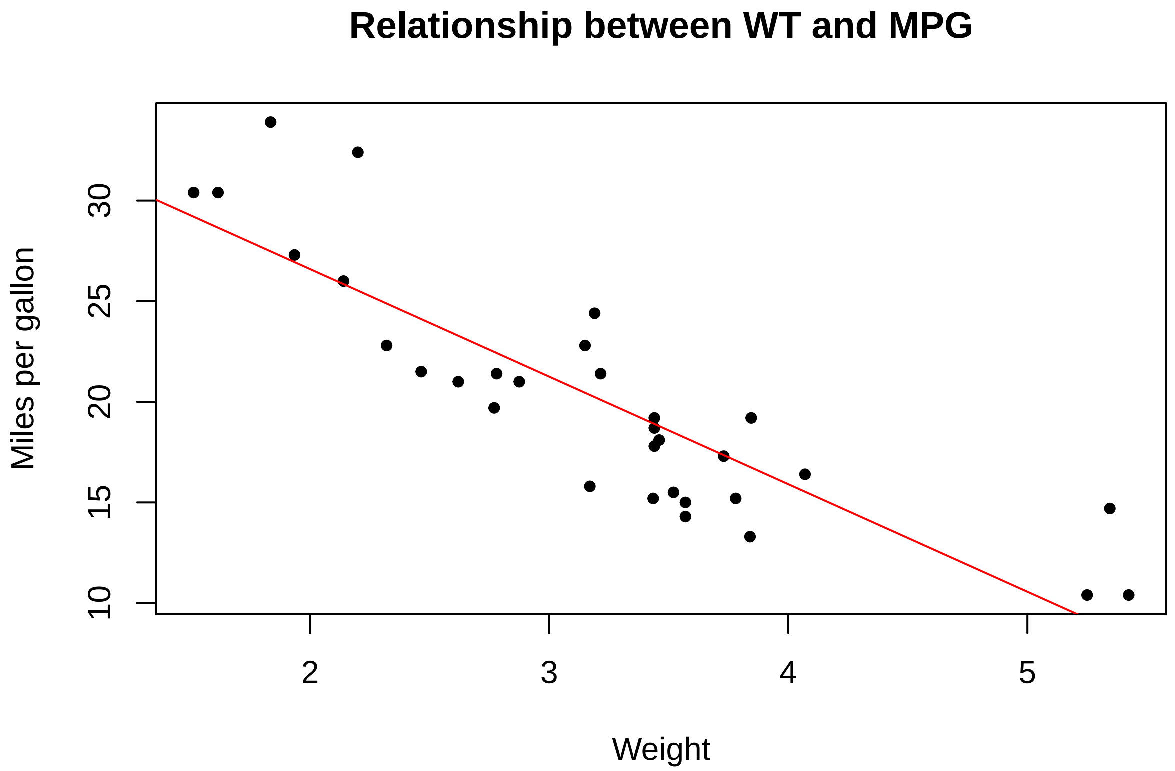
Analysis

What is the relationship between wt (Weight) and mpg (Miles per gallon)?

2-D Scatterplot To first investigate the relationship between wt and mpg, I use `plot()` to plot out the scatterplot with wt as the predictor and mpg as the response variable to examine the relationship intuitively. A linear model is also built using `lm()` to plot out the regression line in red using `abline()`.

```
# scatterplot, relationship between wt and mpg
plot(x = mtcars$wt, y = mtcars$mpg,
     xlab = "Weight", ylab = "Miles per gallon",
     main = "Relationship between WT and MPG", pch = 20)

# regression line
wt.mpg.mdl <- lm(data = mtcars, mpg ~ wt)
abline(wt.mpg.mdl, col = "red")
```



From the graph we can see that there seems to be a negative relationship between wt and mpg, which means as wt increases, mpg decreases. To further confirm this relationship, I calculate the sample mean, sample correlation, and sample covariance to check with the scatterplot, and to further quantify this relationship.

Sample Mean Sample means are calculated using `mean()` for wt and mpg to verify the variables shown on the graph. I am using sample mean instead of other measures because mean is an **unbiased** estimator.

```
c(mean(mtcars$wt), mean(mtcars$mpg))
```

```
## [1]  3.21725 20.09062
```

The mean for wt is 3.217, while the mean for mpg is 20.091. Both calculated means correspond to what are shown in the scatterplot above.

Sample Correlation Then, the sample correlation is calculated using `cor()` for wt and mpg. Sample correlation is used because it is a standard measure for checking association between two variables. However, sample correlation is only **asymptotic unbiased** instead of completely unbiased.

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

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

The sample correlation coefficient between wt and mpg is -0.868, which means that there is a **strong negative linear** relationship between wt and mpg. This conclusion corresponds with what we observed in the scatterplot above. However, since sample correlation is only *unbiased asymptotically*, I want to use another measure that is completely unbiased to measure the relationship.

Sample Covariance Thus, the sample covariance is calculated using `cov()` for wt and mpg. Sample covariance is used because it is also a measure for association between two variables, and it is **unbiased**.

```
cov(mtcars$wt, mtcars$mpg)
```

```
## [1] -5.116685
```

The sample covariance is -5.117, which is negative and means that there is a **negative association** between wt and mpg. This conclusion corresponds to our conclusion using sample correlation, and with our scatterplot above. But how exactly are wt and mpg negatively correlated? What is the intercept β_0 and the slope β_1 if the linear relationship is in the form $mpg = \beta_0 + \beta_1 \times wt$?

Linear Model In order to investigate the specific relationship in the form $mpg = \beta_0 + \beta_1 \times wt$, a linear model is created between wt as the predictor and mpg as the response variable. Since the linear model is already created using `lm()` when trying to plot the regression line above, there I will just use `summary()` to check the results.

```
summary(wt.mpg.mdl)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

From the above summary, we know that the intercept $\beta_0 = 37.285$ and the slope $\beta_1 = -5.345$, which means the specific linear relationship between wt and mpg is as follows: $mpg = 37.285 - 5.345 \times wt$.

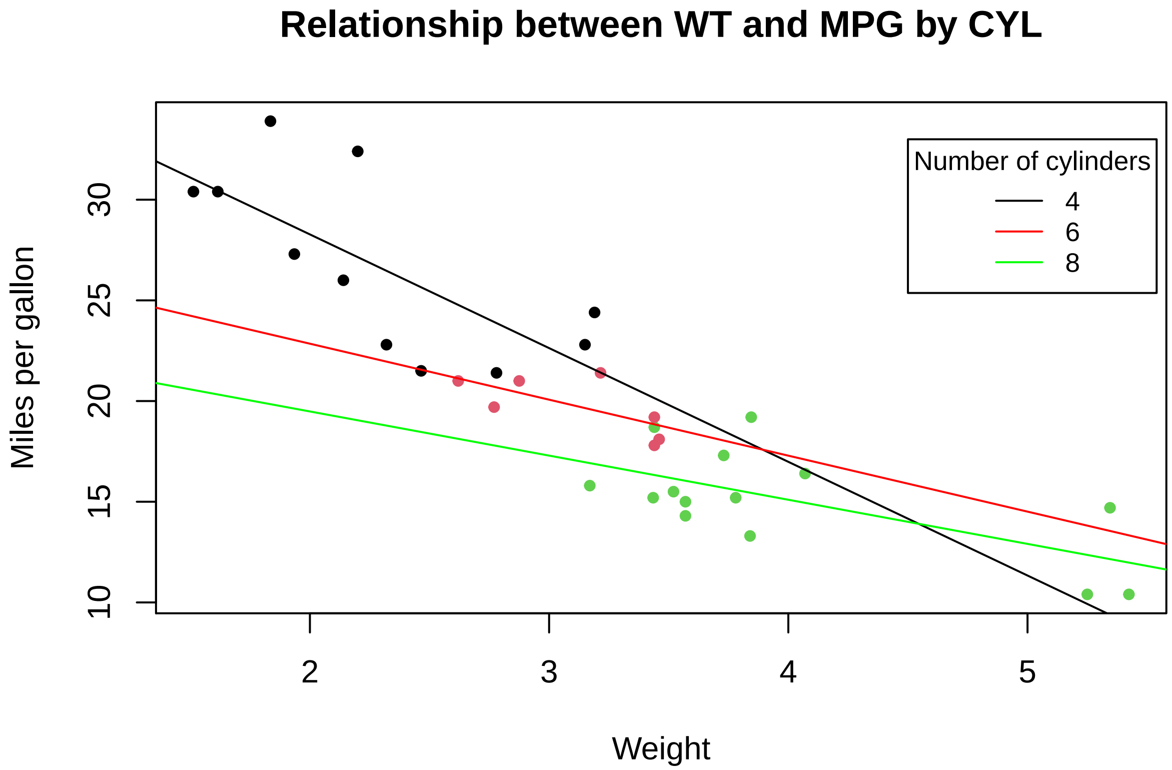
Does the relationship between wt (Weight) and mpg (Miles per gallon) depend on the number of cylinders (cyl)?

2-D Scatterplot (by color) To first investigate whether the negative relationship between wt and mpg is dependent on cyl, I use `plot()` for the same scatterplot above but separate different levels of cyl to examine the relationship individually. Three different linear models are also built using `lm()` to plot out the regression lines for the three different levels of cyl using `abline()`. The legend for the meaning of each color is also created using `legend()`.

```
# scatterplot, relationship between wt and mpg by cyl (color)
plot(x = mtcars$wt, y = mtcars$mpg,
     xlab = "Weight", ylab = "Miles per gallon",
     main = "Relationship between WT and MPG by CYL",
     col = mtcars$cyl.fact, pch = 20)

# regression lines, by cyl
wt.mpg.cyl4.mdl <- lm(data = mtcars %>% filter(cyl == 4), mpg ~ wt)
wt.mpg.cyl6.mdl <- lm(data = mtcars %>% filter(cyl == 6), mpg ~ wt)
wt.mpg.cyl8.mdl <- lm(data = mtcars %>% filter(cyl == 8), mpg ~ wt)
abline(wt.mpg.cyl4.mdl, col = "black")
abline(wt.mpg.cyl6.mdl, col = "red")
abline(wt.mpg.cyl8.mdl, col = "green")

# add legend
legend(4.5, 33, legend=c("4", "6", "8"),
       col=c("black", "red", "green"), lty = 1,
       cex=0.8, title="Number of cylinders")
```



From the graph we can see that all three linear relationships for wt and mpg seem to be negative. However, the slope and intercept for each negative linear relationship are different. To further quantify the differences, I calculate the sample mean, sample correlation, and sample covariance for each level of cyl.

Sample Mean Sample mean is **unbiased**, and is thus calcluated for wt and mpg for each level of cyl to check the differences in association.

```
mtcars %>%
  group_by(cyl) %>%
  summarize(wt.mean = mean(wt),
            mpg.mean = mean(mpg),
            .groups = 'drop')
```

```
## # A tibble: 3 x 3
##     cyl wt.mean mpg.mean
##   <int>   <dbl>    <dbl>
## 1     4    2.29     26.7
## 2     6    3.12     19.7
## 3     8    4.00     15.1
```

When cyl is 4, the mean for wt is 2.286, while the mean for mpg is 26.664. When cyl is 6, the mean for wt is 3.117, while the mean for mpg is 19.743. When cyl is 8, the mean for wt is 3.999, while the mean for mpg is 15.100. Therefore, the mean of wt and the mean of mpg seem to be different for all levels of cyl. And as cyl increases, the mean of wt increases but the mean of mpg decreases.

Sample Correlation Sample correlation is only **asymptotic unbiased** as mentioned above, and is calculated for wt and mpg for each level of cyl to check the differences in association.

```
mtcars %>%
  group_by(cyl) %>%
  summarize(correlation = cor(wt, mpg),
            .groups = 'drop')
```

```
## # A tibble: 3 x 2
##     cyl correlation
##   <int>       <dbl>
## 1     4      -0.713
## 2     6      -0.682
## 3     8      -0.650
```

When cyl is 4, the sub-sample correlation is -0.713. When cyl is 6, the sub-sample correlation is -0.682. When cyl is 8, the sub-sample correlation is -0.650. Therefore, the correlations between wt and mpg are all negative, which means that wt and mpg have a **negative linear** relationship for all levels of cyl. The correlations are also all different for all levels of cyl. And as cyl increases, the correlation decreases.

Sample Covariance Sample covariance is **unbiased** as mentioned above, and is calculated for wt and mpg for each level of cyl to check the differences in association

```
mtcars %>%
  group_by(cyl) %>%
  summarize(correlation = cov(wt, mpg),
            .groups = 'drop')
```

```
## # A tibble: 3 x 2
##     cyl correlation
##   <int>       <dbl>
## 1     4      -1.83
## 2     6      -0.353
## 3     8      -1.26
```

When cyl is 4, the sub-sample covariance is -1.832. When cyl is 6, the sub-sample covariance is -0.353. When cyl is 8, the sub-sample covariance is -1.264. Therefore, the covariances between wt and mpg are all negative, which means that wt and mpg have a **negative** relationship for all levels of cyl. The covariances are also all different for all levels of cyl.

Linear Model (separate) In order to investigate the specific relationship in the form $mpg = \beta_0 + \beta_1 \times wt$ for each level of cyl, three linear models are created between wt as the predictor and mpg as the response variable. Since the linear model is already created using `lm()` when trying to plot the regression lines above, there I will just use `summary()` to check all the three results.

```
summary(wt.mpg.cyl4.mdl)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars %>% filter(cyl == 4))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1513 -1.9795 -0.6272  1.9299  5.2523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.571      4.347   9.104 7.77e-06 ***
## wt            -5.647      1.850  -3.052   0.0137 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.332 on 9 degrees of freedom
## Multiple R-squared:  0.5086, Adjusted R-squared:  0.454
## F-statistic: 9.316 on 1 and 9 DF,  p-value: 0.01374
```

```
summary(wt.mpg.cyl6.mdl)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars %>% filter(cyl == 6))
##
## Residuals:
##       1       2       3       4       5       6       7
## -0.1250  0.5840  1.9292 -0.6897  0.3547 -1.0453 -1.0080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   28.409      4.184   6.789  0.00105 **
## wt            -2.780      1.335  -2.083  0.09176 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.165 on 5 degrees of freedom
## Multiple R-squared:  0.4645, Adjusted R-squared:  0.3574
## F-statistic: 4.337 on 1 and 5 DF,  p-value: 0.09176
```

```
summary(wt.mpg.cyl8.mdl)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars %>% filter(cyl == 8))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1491 -1.4664 -0.8458  1.5711  3.7619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.8680     3.0055   7.942 4.05e-06 ***
## wt           -2.1924     0.7392  -2.966   0.0118 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.024 on 12 degrees of freedom
## Multiple R-squared:  0.423,  Adjusted R-squared:  0.3749
## F-statistic: 8.796 on 1 and 12 DF,  p-value: 0.01179
```

When cyl is 4, the intercept $\beta_0 = 39.571$ and the slope $\beta_1 = -5.647$, which means the specific linear relationship between wt and mpg is as follows: $mpg = 39.571 - 5.647 \times wt$. In this model, wt **significantly predicts** mpg since $p = 0.0137 < \alpha = 0.05$.

When cyl is 6, the intercept $\beta_0 = 28.409$ and the slope $\beta_1 = -2.780$, which means the specific linear relationship between wt and mpg is as follows: $mpg = 28.409 - 2.780 \times wt$. In this model, wt **does not significantly predict** mpg since $p = 0.0918 > \alpha = 0.05$.

When cyl is 8, the intercept $\beta_0 = 23.868$ and the slope $\beta_1 = -2.192$, which means the specific linear relationship between wt and mpg is as follows: $mpg = 23.868 - 2.192 \times wt$. In this model, wt **significantly predicts** mpg since $p = 0.0118 < \alpha = 0.05$.

From the above summary, we know that the intercepts and slopes for the linear relationship between wt and mpg are all different for all levels of cyl And only when cyl is 4 and 8 is wt significantly predicting mpg. When cyl is 6, wt does not significantly predict mpg.

Conclusion

In summary, from our examination using sample correlation, sample covariance, and a linear model, we can see that there is a clear strong negative linear relationship between wt and mpg. And the linear relationship is in the form $mpg = 37.285 - 5.345 \times wt$.

After the investigation on the effect of cyl on the relationship between wt and mpg, we see that there’s a difference in sample means of wt, sample means of mpg, sample correlations between wt and mpg, and sample covariances between wt and mpg. More specifically, for the sample means of wt and mpg, as cyl increases, the mean of wt increases but the mean of mpg decreases. For the sample correlation between wt and mpg, as cyl increases, the sample correlation decreases. And for the sample covariances between wt and mpg, the sample covariances are all different for all levels of cyl.

When we compare the linear models for each level of cyl, we see that when cyl is 4, wt significantly predicts mpg, and the relationship is in the form $mpg = 39.571 - 5.647 \times wt$. When cyl is 6, wt does not significantly predict mpg, and the relationship is in the form $mpg = 28.409 - 2.780 \times wt$. When cyl is 8, wt significantly predicts mpg, and the relationship is in the form $mpg = 23.868 - 2.192 \times wt$. Thus, the analyses above seem to suggest that even though the relationship between wt and mpg are all negative for all levels of cyl, there still seems to be an effect of cyl on this relationship since the statistic for sample means, sample correlation, sample covariance, regerssion slope/intercept, and whether wt signficantly predicts mpg (p-value) are all different from each other for all levels of cyl.

Discussion

The study shows that there is a strong negative linear relationship between wt and mpg, and that cyl seems to have an effect on this relationship based on examination of correlation/covariance and linear regression models for each level of cyl. And if we were to think about real-life events, it follows that a heavier car needs more cylinders to power, and the higher the number of cylinders, the more fuel consumption which leads to lower miles per gallon. However, since this dataset is from several decades ago, the relationship might not apply anymore given more advanced techonology on energy consumption. Further explorations using more current datasets are needed to uncover the true relationship nowadays after 2020/2021. Other variables, such as the brand of cars or the type of cars, could also play a role in wt and mpg relationship, thus needs examination using similar or more advanced statistical tools.

Through simple comparison between statistics, such as sample correlation/covariance and linear model coefficients, I am able to tell that there seems to be a difference between each relationship of wt and mpg across levels of cyl, thus suggesting that this relationship betweenw wt and mpg depends on cyl. However, this simple comparison does not tell whether the difference is signficant. Therefore, a more wholistic model is also needed to compare the differences by including both cyl and wt into predicting mpg, which would allow us to see whether the difference in each level of cyl is signficantly different from each other. But since we are not familiar with this type of tool, we will continue learning more statitical analysis methods in MATH 189, and apply more advanced tools for future examination.