Coursera: Regression Models - Course Project

Chun-Li Hou

07 September, 2020

Executive Summary

The 1974 Motor Trend US magazine dataset (mtcars) is used to evaluate the effect of transmission design on mpg (miles per gallon) in automobiles. Simply put we are asking the questions as following:

- **Is an automatic or manual transmission better for mpg?**
- **How is the mpg difference between automatic and manual transmissions?**

Dataset Description

The dataset consists of a dataframe with 32 observations (nrow) and 11 variables (nol).

- mpg: Miles per US gallon
- cyl: Number of cylinders
- disp: Displacement (cubic inches)
- hp: Gross horsepower
- drat: Rear axle ratio
- wt: Weight (lb / 1000)
- qsec: 1 / 4 mile time
- vs: V/S
- am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

Loading & Processing & Exploring

```
# load
data("mtcars")

# transform
mtcars$cyl = factor(mtcars$cyl)
mtcars$vs = factor(mtcars$vs)
mtcars$gear = factor(mtcars$gear)
mtcars$carb = factor(mtcars$carb)
mtcars$am = factor(mtcars$am, labels = c("Automatic", "Manual"))

# print
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

In this section, we deep dive into our data and explore various relationships between variables of interest.

Initially, we plot the relationship bewteen all the variables of the dataset (Figure.1 in the appendix). From the plot, we notice that most of the variables in the dataset seem to have correlation with mpg. So, we will use linear model to identify and qunatify that.

Since we are interested in the effects of car transmission type on mpg. we plot boxplot of the variable mpg whem am is automatic or manual (Figure.2 in the appendix). This plot clearly depicts an increase in the mpg when transmission is manual.

Regression Analysis & Inference

```
# fit.best
init.mod = lm(mpg ~ ., data = mtcars)
best.mod = step(init.mod, direction = "both", trace = FALSE)

# print
summary(best.mod)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.939 -1.256 -0.401  1.125  5.051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7083     2.6049   12.94  7.7e-13 ***
## cyl6         -3.0313     1.4073   -2.15   0.0407 *
## cyl8         -2.1637     2.2843   -0.95   0.3523
## hp           -0.0321     0.0137   -2.35   0.0269 *
## wt           -2.4968     0.8856   -2.82   0.0091 **
## amManual      1.8092     1.3963    1.30   0.2065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.84
## F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

The best model obtained from the above computations consists of the variables as cyl, wt, hp and am. From the best model, we observe that the adjusted r squared value is 0.84. Thus, we can conclude that more than **84%** of the variability is explained by the best model.

```
# fit.base
base.mod = lm(mpg ~ am, data = mtcars)

# print
summary(base.mod)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.15       1.12   15.25  1.1e-15 ***
## amManual        7.24       1.76    4.11  0.00029 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

As using only the indicated variable (am) on mpg, the adjusted r squared value is 0.34. Thus, we can conclude that more than **34%** of the variability is explained by the base model.

```
anova(base.mod, best.mod)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     26 151  4       570 24.5 1.7e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on p-value < 0.05, we reject H0 and conclude that the equations are not equivalent, which means that the variables of cyl, hp, and wt do contirbute to the accuracy of the model.

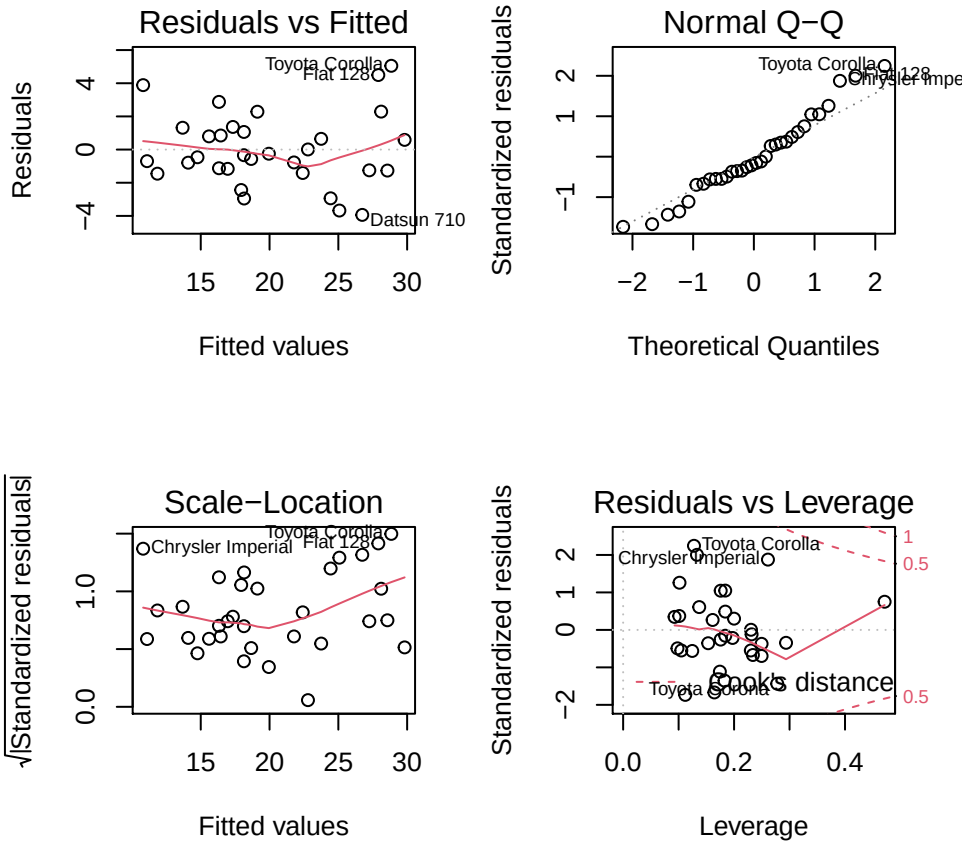
```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -4, df = 18, p-value = 0.001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
## mean in group Automatic    mean in group Manual
##                    17.1                    24.4
```

We also perform a t-test assuming that the transmission data has a normal distribution and we clearly see that the manual and automatic tramsmissions are significantly different based on p-value < 0.05.

Assumption Checking

```
par(mfrow = c(2, 2))
plot(best.mod)
```



From the above plots, we can check the following assumptions needing to be established for a regression model.

- **Residuals vs Fitted plot**: dataset seems to be ramdomly scattered on the plot and verify the independence condition.
- **Normal Q-Q plot**: the points that mostly fall on the line indicate that the residuals are normally distributed.
- **Scale-Location plot**: it consists of points scattered in a constant band pattern, which indicates homoscedastic.
- **Redisuals vs Leverage plot**: the point as an outlier shows its level in influence and leverage. We compute some regression diagonostics of our model to find out these possible outliers as below. Thus, we find out that **Toyota Corona** is the point as high influence and high leverage. This may need to take a look on the observation in detail to see whether it has a record error or not, or try to take it off from the model to see whether the conclusion change, or just report results with and without the point.

```
influence = dfbetas(best.mod)
tail(sort(influence[, 6]), 3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##             0.351             0.429             0.731
```

```
# sum((abs(dfbetas(best.mod)))>1) # default accepted influencial point
```

The influence point has extreme value of Y, so it has the power to move the line no matter about the leverage. It can be identified by the cook's distance.

```
leverage = hatvalues(best.mod)
tail(sort(leverage), 3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##               0.278               0.294               0.471
```

The leverage point has extreme value of X, so it has a greater possible ability to move the line based on the distance from the line or the overall pattern that is influence.

```
data.frame(vif(best.mod)) %>% arrange(GVIF) %>% select(GVIF) %>% t()
```

```
##        am   wt  hp  cyl
## GVIF 2.59 4.01 4.7 5.82
```

The generalized variance inflation factor (GVIF) is a measure of collinearity. The bigger number, the less independency, means higher colinearity. Thus, am is comparatively the best independent variable to mpg.

Conclusion

1. Based on the analysis result, we can conclude the following:

- mpg decreases by **2.2** unit as comparing with the car in 8 to 4 number of cylinders, and decreases by **3.0** unit as comparing with the car in 6 to 4 number of cylinders (when other variables are fixed) (Figure.3 in the appendix).
- mpg decreases negligibly by **0.03** unit as increasing 1 unit of hp (when other variables are fixed) (Figure.4 in the appendix). Or, by rescaling, mpg decreases by **3.2** unit as increasing 100 unit of hp.
- mpg decreases by **2.5** unit as increasing 1 unit of wt, which means increasing 1000 lb in weigth of car (when other variables are fixed) (Figure.5 in the appendix).
- mpg increases by **1.8** unit as comparing with the car in manual to in automatic transmission (when other variables are fixed) (Figure.2 in the appendix).

2. Above set of analysis yields the inference that **manual transmission is better than automatic transmission** with a more 1.8 miles per gallon as fixed other variables.

3. Additionally, type of transmission is the most independent variable to mpg in the model. However, it seems that wt, hp, and cyl are more statistically significant when determining mpg.

Appendix

```
g = ggpairs(mtcars,
            lower = list(continuous = wrap("smooth", method = "lm"))) +
    labs(caption = "Figure.1")
g
```

Figure.1 Overview of Dataset

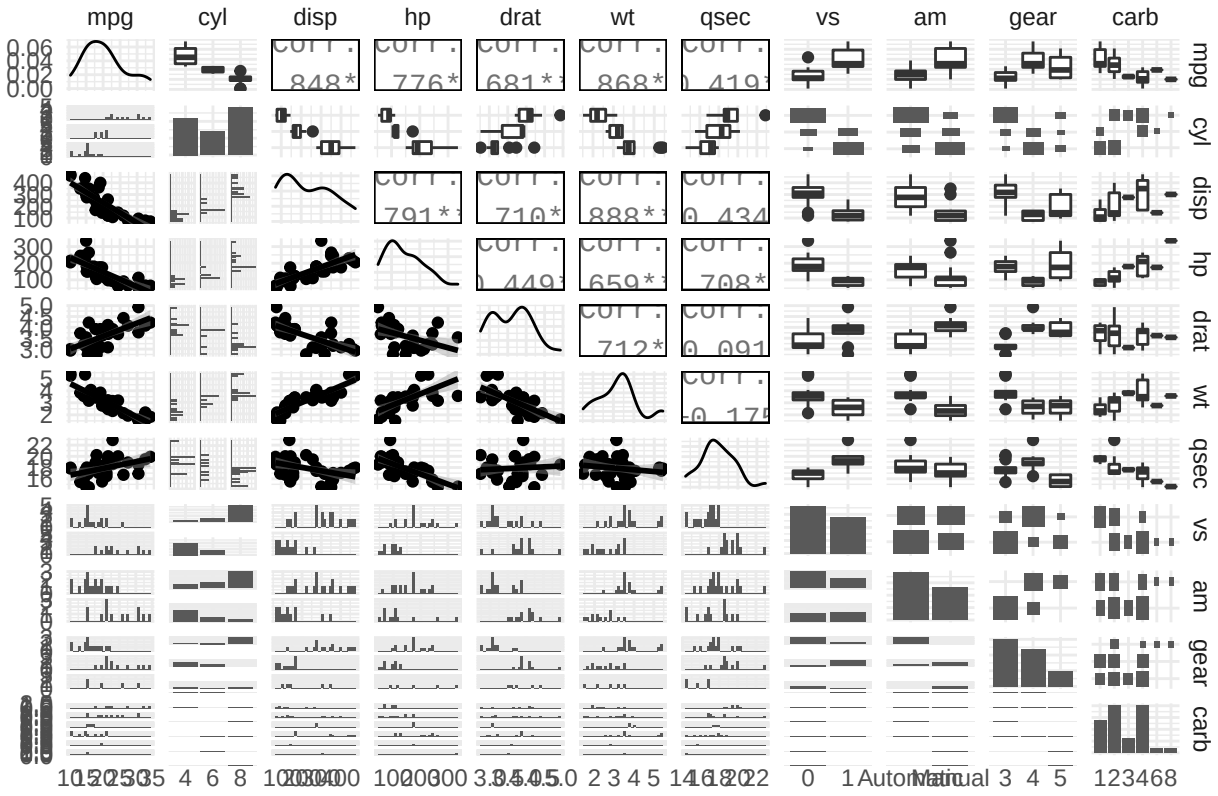


Figure.1

```
g[1, 9] +
    labs(title = "Boxplot of MPG vs Transmission",
         x = "Transmission\n(0 = Automatic, 1 = Manual)",
         y = "Miles per Gallon",
         caption = "Figure.2")
```

Figure.2 Boxplot of MPG vs Transmission

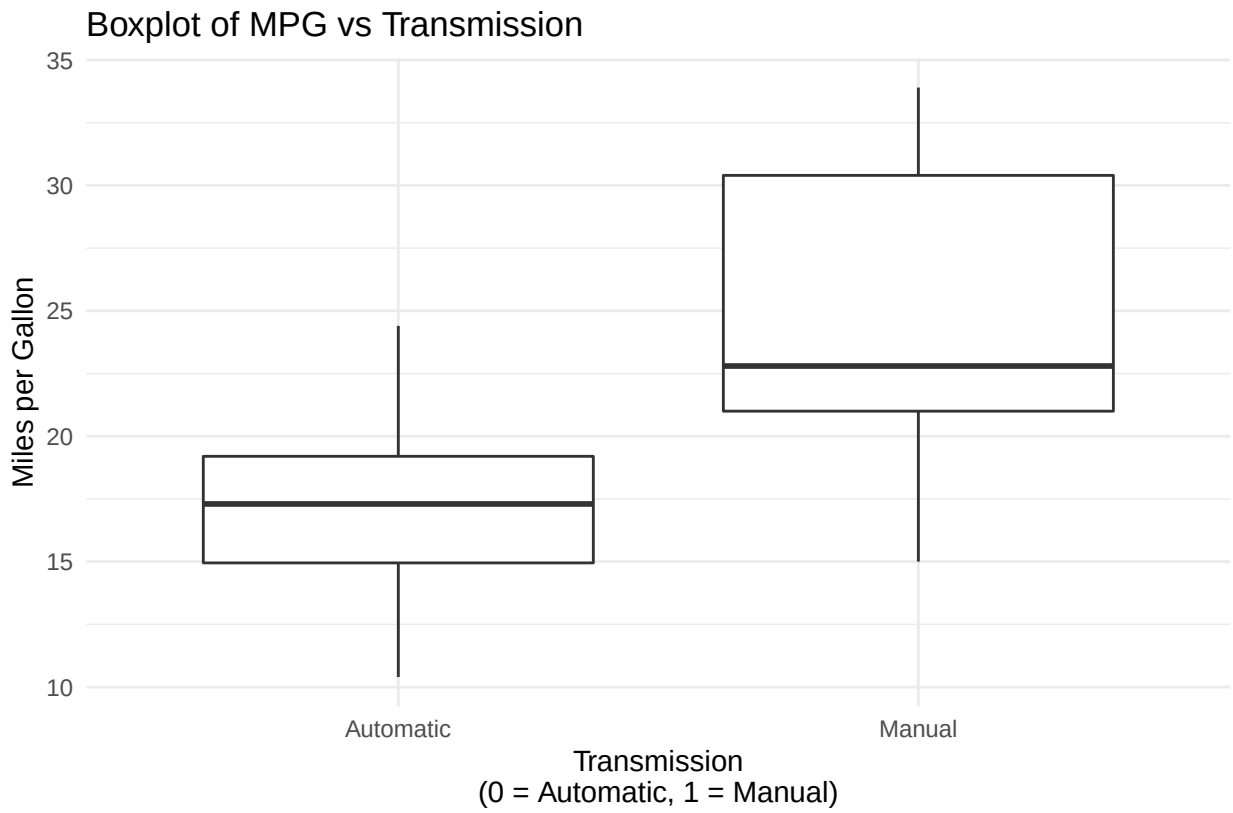


Figure.2

```
g[1, 2] +
    labs(title = "Boxplot of Mileage by Cylinder",
         x = "Number of Cylinders",
         y = "Miles per Gallon",
         caption = "Figure.3")
```

Figure.3 Boxplot of Mileage by Cylinder

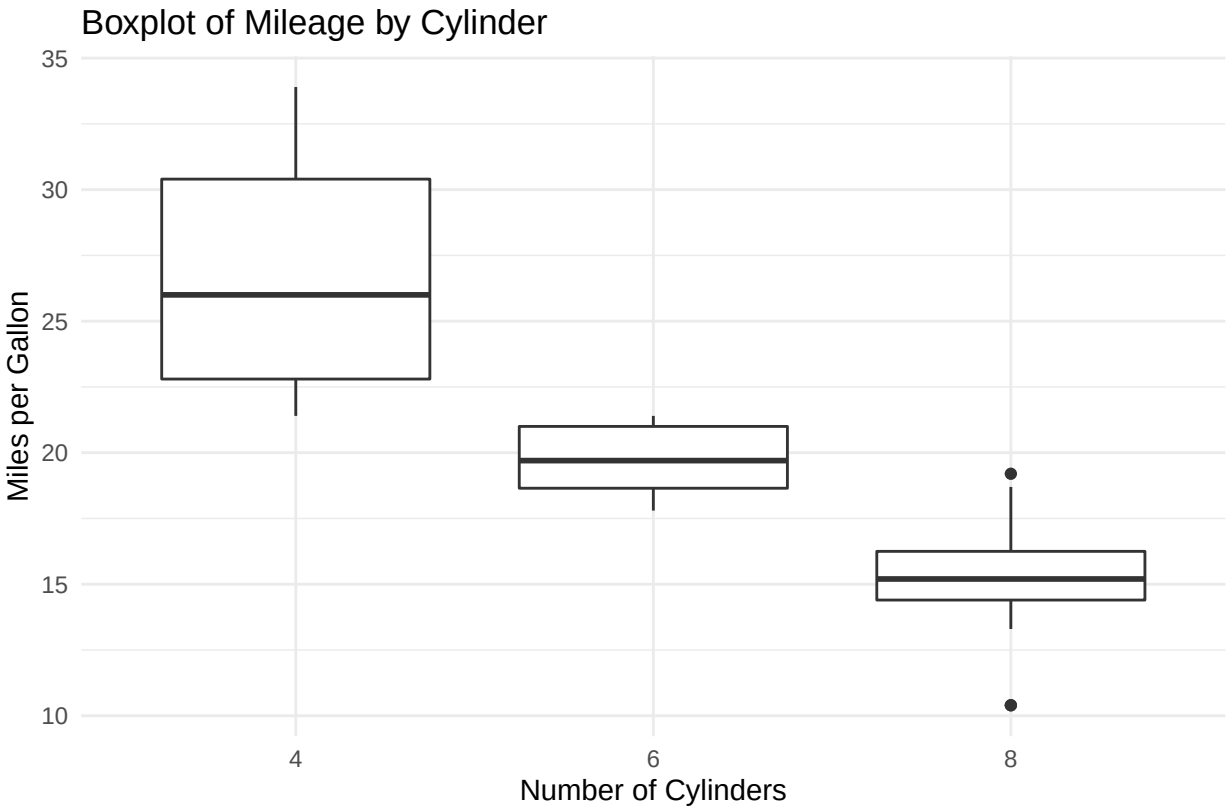


Figure.3

```
g[4, 1] +
    labs(title = "Regression Plot of Mileage by Gross Horsepower",
         y = "Gross Horsepower",
         x = "Miles per Gallon",
         caption = "Figure.4")
```

Figure.4 Regression Plot of Mileage by Gross Horsepower

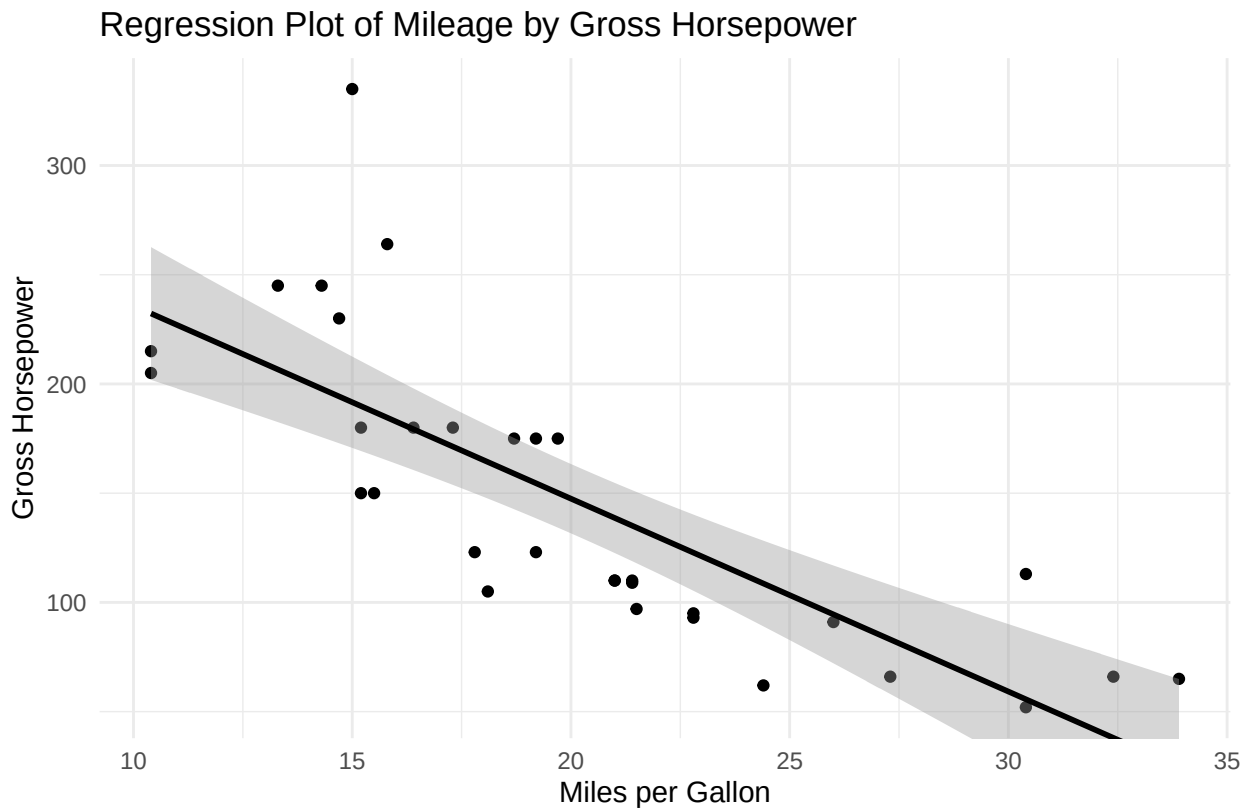


Figure.4

```
g[6, 1] +
    labs(title = "Regression Plot of Mileage by Weight",
         y = "Weight (lb / 1000)",
         x = "Miles per Gallon",
         caption = "Figure.5")
```

Figure.5 Regression Plot of Mileage by Weight

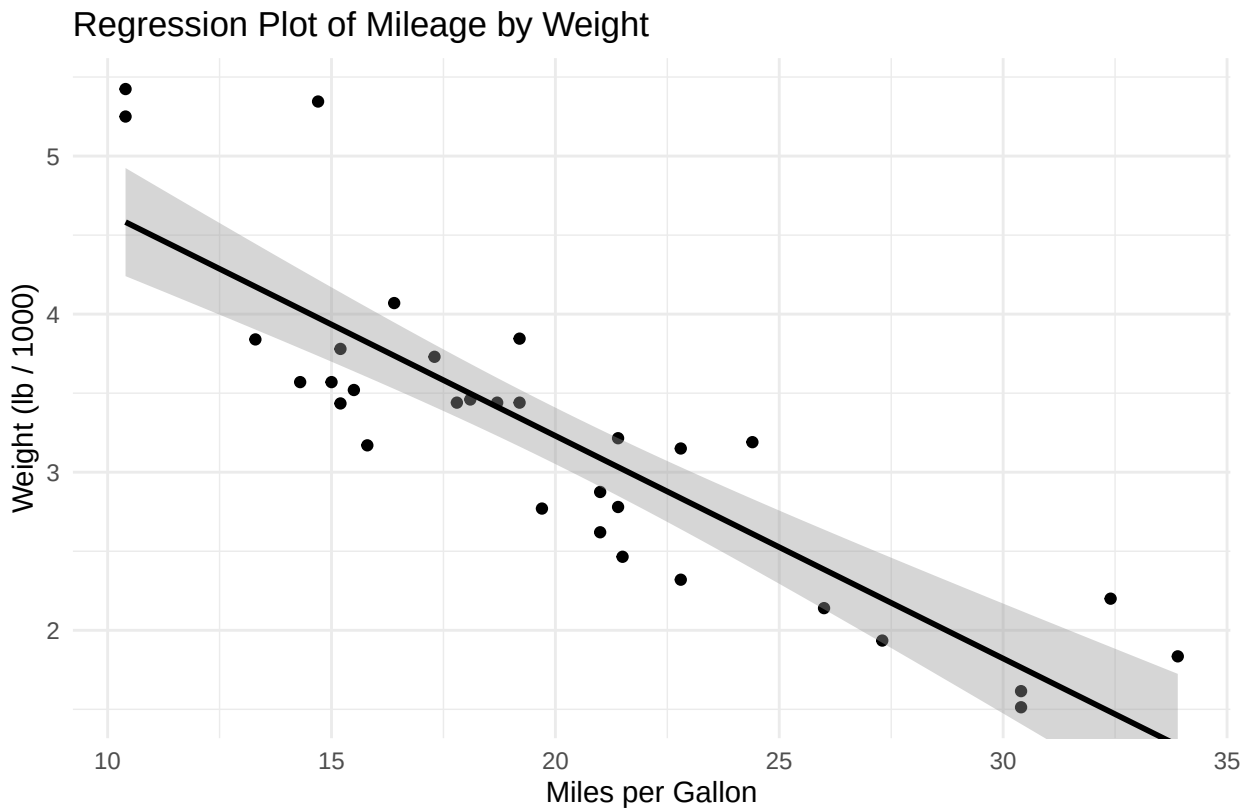


Figure.5