cars-regression-assignment.R

William Alexander

May 16, 2018

Executive Summary

In this report we will use the mtcars dataset from the 1974 Motor Trend US magazine to answer the following questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

MPG is miles per gallon (MPG). How different is the MPG between automatic and manual transmissions?

Using hypothesis testing and simple linear regression, we determine that there is a signficant difference between the mean MPG for automatic and manual transmission cars. It is found that the manual transmission cars has 7.245 more MPGs on average. However, in order to adjust for other confounding variables such as the weight and horsepower of the car, we ran a multivariate regression to get a better estimate the impact of transmission type on MPG. After validating the model using ANOVA, the results from the multivariate regression reveal that, on average, manual transmission cars get 2.084 miles per gallon more than automatic transmission cars.

Data Processing

Reading the mtcars data

```
data(mtcars)
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 : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ 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  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Here we that am is numeric. In order to have better interpretation we change it to factor variable with two levels

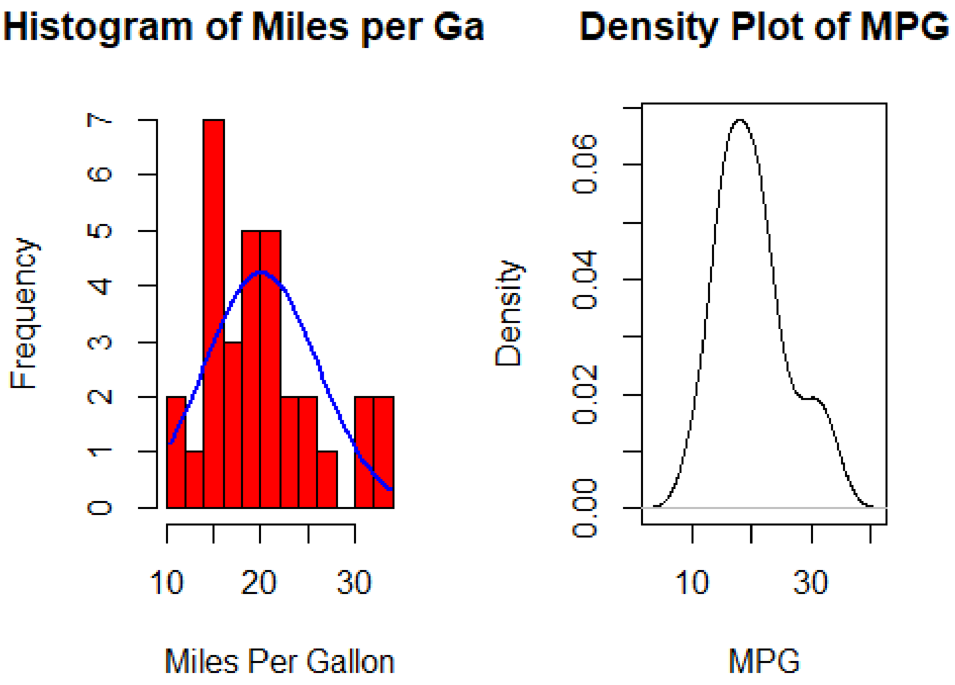
```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
```

Exploratory Data Analysis

Since we are going to run a linear regression, let us plot mpg to see its distribution.

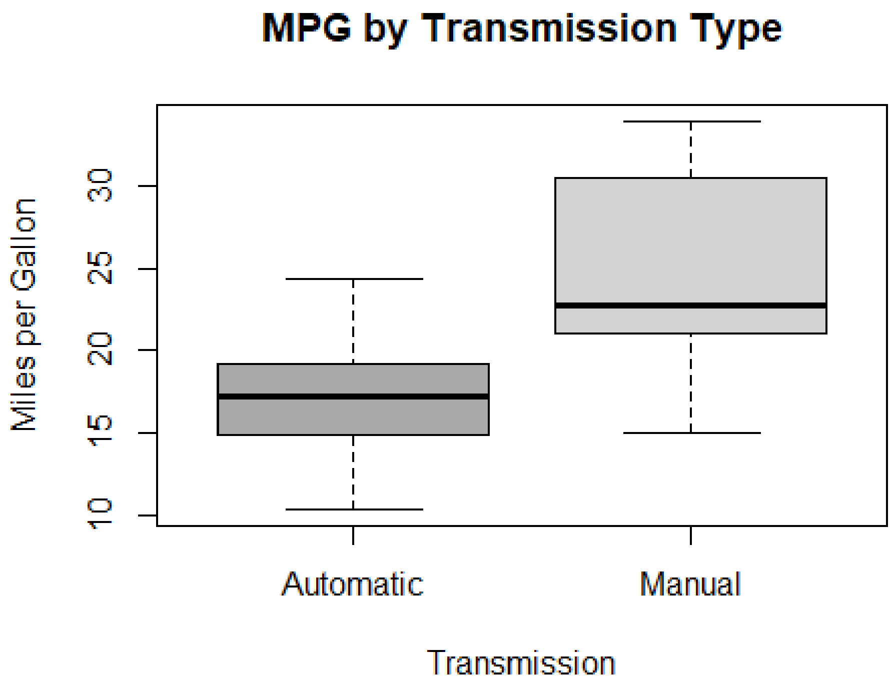
```
par(mfrow = c(1, 2))
# Histogram with Normal Curve
x <- mtcars$mpg
h<-hist(x, breaks=10, col="red", xlab="Miles Per Gallon",
   main="Histogram of Miles per Gallon")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)

# Kernel Density Plot
d <- density(mtcars$mpg)
plot(d, xlab = "MPG", main ="Density Plot of MPG")
```



The graph shows that the mpg variable is almost a normal distribution and there are no outliers or skewing. Now let us see if MPG varies for automatic vs manual transmission cars.

```r
boxplot(mpg~am, data = mtcars,
        col = c("dark grey", "light grey"),
        xlab = "Transmission",
        ylab = "Miles per Gallon",
        main = "MPG by Transmission Type")
```



We can easily see there is a difference in mpg for manual vs automatic and manual seems to have higher MPG compared to automatic.

Hypothesis Testing

```r
aggregate(mpg~am, data = mtcars, mean)

##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

The mean MPG of manual transmission cars is 7.245 MPGs higher than that of automatic transmission cars.

Building our Model

Correlation

We check which predictors should go into our model

```
data(mtcars)
sort(cor(mtcars)[1,])
```

```
##         wt        cyl       disp         hp       carb       qsec 
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840 
##       gear         am         vs       drat        mpg 
##  0.4802848  0.5998324  0.6640389  0.6811719  1.0000000 
```

In addition to am (which by default must be included in our regression model), we see that wt, cyl, disp, and hp are highly correlated with our dependent variable mpg. As such, they may be good candidates to include in our model. However, if we look at the correlation matrix, we also see that cyl and disp are highly correlated with each other. Since predictors should not exhibit collinearity, we should not have cyl and disp in in our model.

Regression Model

Simple Linear Regression

```
fit <- lm(mpg~am, data = mtcars)
summary(fit)
```

```
## 
## Call:
## lm(formula = mpg ~ am, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.3923 -3.0923 -0.2974  3.2439  9.5077 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385 
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

We do not gain much more information from our hypothesis test using this model. Interpreting the coefficient and intercepts, we say that, on average, automatic cars have

17.147 MPG and manual transmission cars have 7.245 MPGs more. In addition, we see that the R^2 value is 0.3598. This means that our model only explains 35.98% of the variance.

Multivariate Linear Regression

Next, we fit a multivariate linear regression for mpg on am, wt, and hp. Since we have two models of the same data, we run an ANOVA to compare the two models and see if they are significantly different.

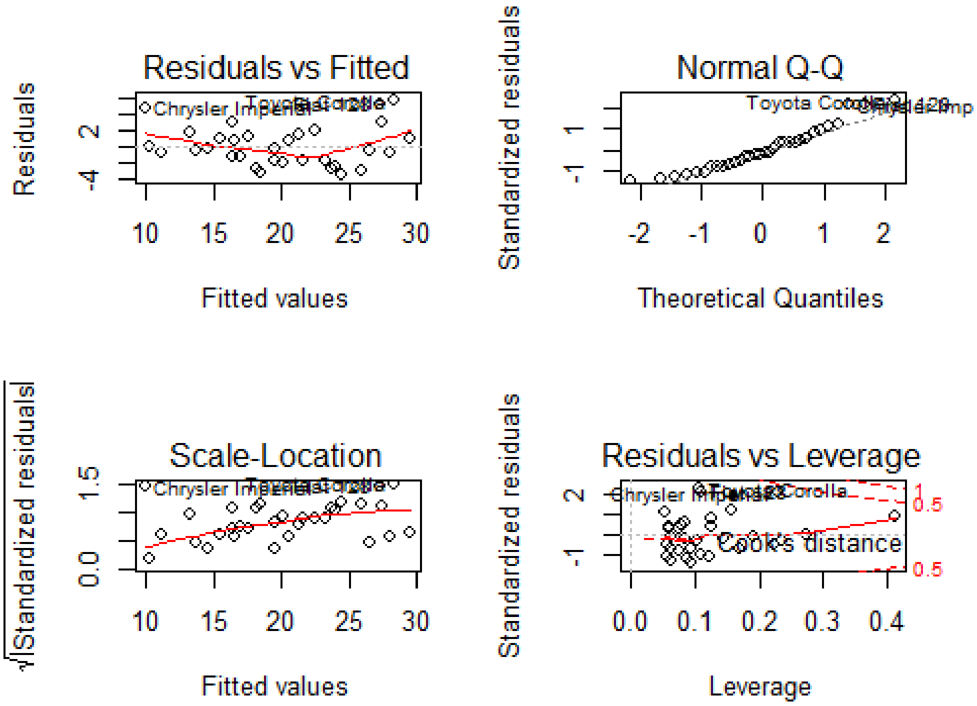
```r
bestfit <- lm(mpg~am + wt + hp, data = mtcars)
anova(fit, bestfit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 3.745e-09, we reject the null hypothesis and claim that our multivariate model is significantly different from our simple model.

Before we report the details of our model, it is important to check the residuals for any signs of non-normality and examine the residuals vs. fitted values plot to spot for any signs of heteroskedasticity.

```r
par(mfrow = c(2,2))
plot(bestfit)
```



Our residuals are normally distributed and homoskedastic. We can now report the estimates from our final model.

```
summary(bestfit)

##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am           2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

Final Findings

This model explains over 83.99% of the variance. Moreover, we see that wt and hp did indeed confound the relationship between am and mpg (mostly wt). Now when we read the coefficient for am, we say that, on average, manual transmission cars have 2.084 MPGs more than automatic transmission cars.