Regression Analysis Project

Ivan G. Palser

Saturday, June 21, 2014

Executive Summary

The purpose of this anlysis is to look at a data set of a collection of cars, and explore the relationship between a set of variables and miles per gallon (MPG) (outcome). The following two questions need to be addressed:

- Is an automatic or manual transmission better for MPG?
- Quantifying how different is the MPG between automatic and manual transmissions?

The process we will follow is hypothesis testing, simple linear regression and multivariate regression to take into account other variables above and beyond transmisson type such as vehicle weight and horsepower. I found some of the R code from various sources on the web as the mtcars data set is a popular data set for R and linear regression education.

It turns out that simple linear regression shows cars with manual transmissions to get 7.245 more miles per gallon than ones with automatic transmissions. However when performing multivariate regression to account for other factors such as eight and horsepower the manual transission cars only did 2.084 miles per gallon better.

Data Processing

Loading the mtcars data

```
data(mtcars)
```

Here we see that our predictor, am, is a numeric class, let’s convert this to a factor class and label the levels as Automatic and Manual for ease of use.

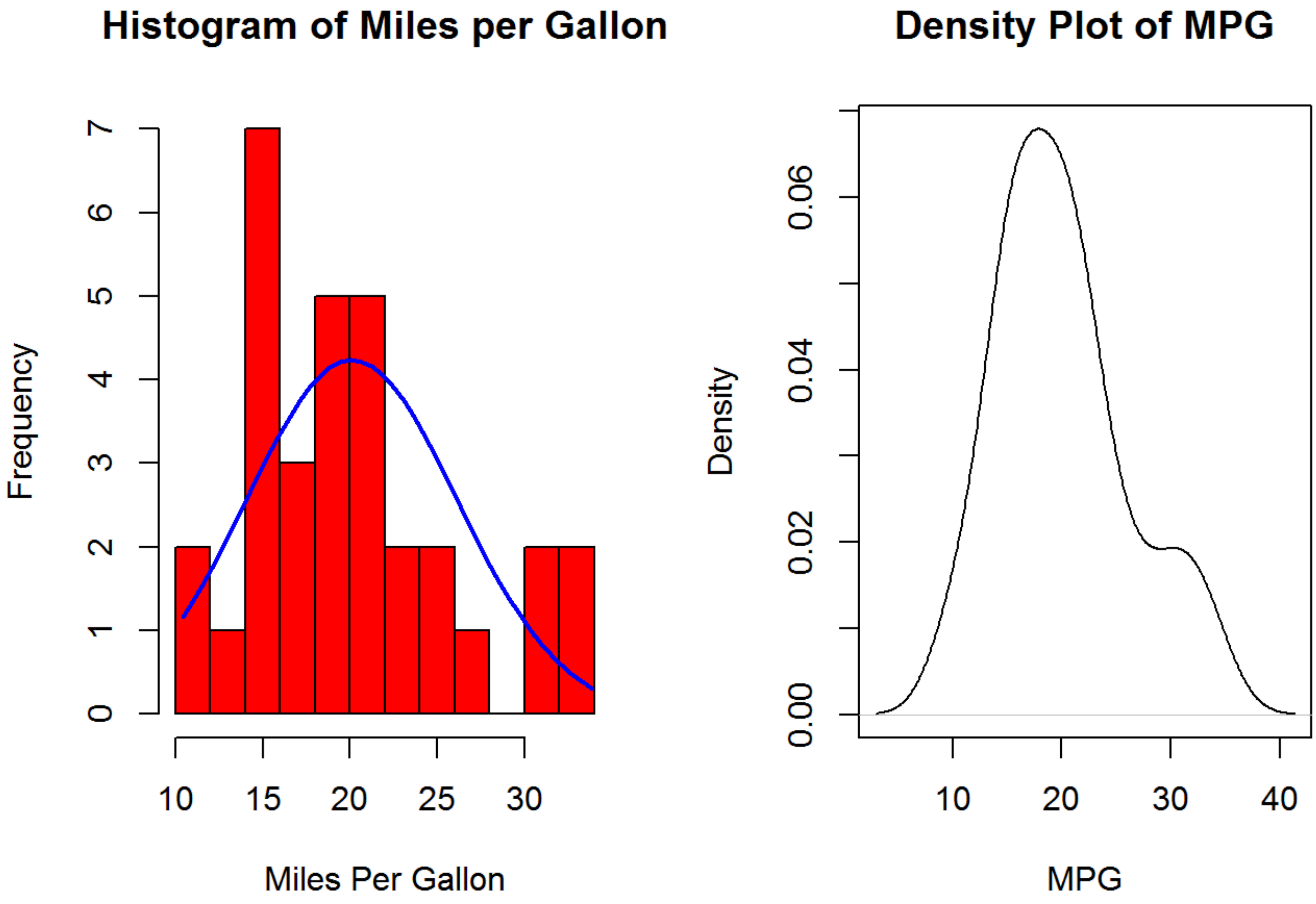
```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
```

Exploring the Data

The first step is to plot the miles per gallon and check the 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")
```

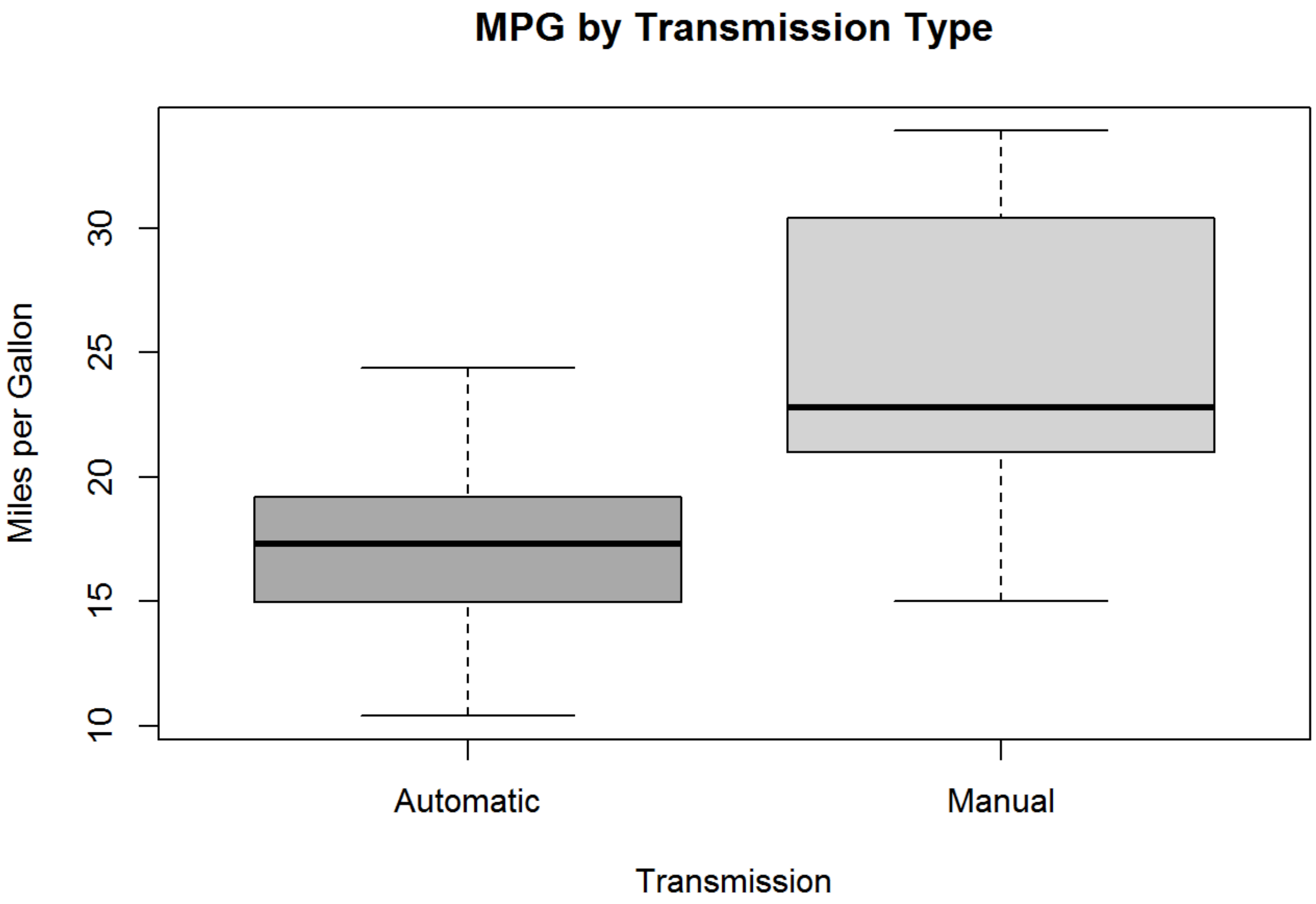


Histogram of Miles per Gallon

Density Plot of MPG

The distribution of mpg is approximately normal and there are no outliers skewing the data. Here is a box plot to see how mpg varies by automatic versus manual transmission.

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



It is clear the the MPG varies by transmission type but we need to perform further analysis.

Hypothesis Testing

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

```
##          am   mpg
## 1 Automatic 17.15
## 2    Manual 24.39
```

The mean MPG of manual transmission cars is 7.245 MPGs higher than that of automatic transmission cars. Is this a significant difference? We set our alpha-value at 0.5 and run a t-test to find out.

```
autoData <- mtcars[mtcars$am == "Automatic",]
manualData <- mtcars[mtcars$am == "Manual",]
t.test(autoData$mpg, manualData$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  autoData$mpg and manualData$mpg
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
## mean of x mean of y
##     17.15     24.39
```

With a p-value of 0.001374, we reject the null hypothesis and claim that there is a signficiant difference in the mean MPG between manual transmission cars and that of automatic transmission cars. Now we must quantify that difference.

Building the Model

Correlation

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

```
##      wt     cyl    disp      hp    carb    qsec    gear      am      vs
## -0.8677 -0.8522 -0.8476 -0.7762 -0.5509  0.4187  0.4803  0.5998  0.6640
##    drat     mpg
##  0.6812  1.0000
```

We can see that wt, cyl, disp, and hp are highly correlated with our dependent variable mpg. They are good data points to include in the model.

Including wt and hp in our regression equation makes sense - heavier cars and cars that have more horsepower should have lower MPGs.

Regression Analysis

Simple Linear Regression

To begin our model testing, we fit a simple linear regression for mpg on am.

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

```
## 
## 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 ***
## am              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
```

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.

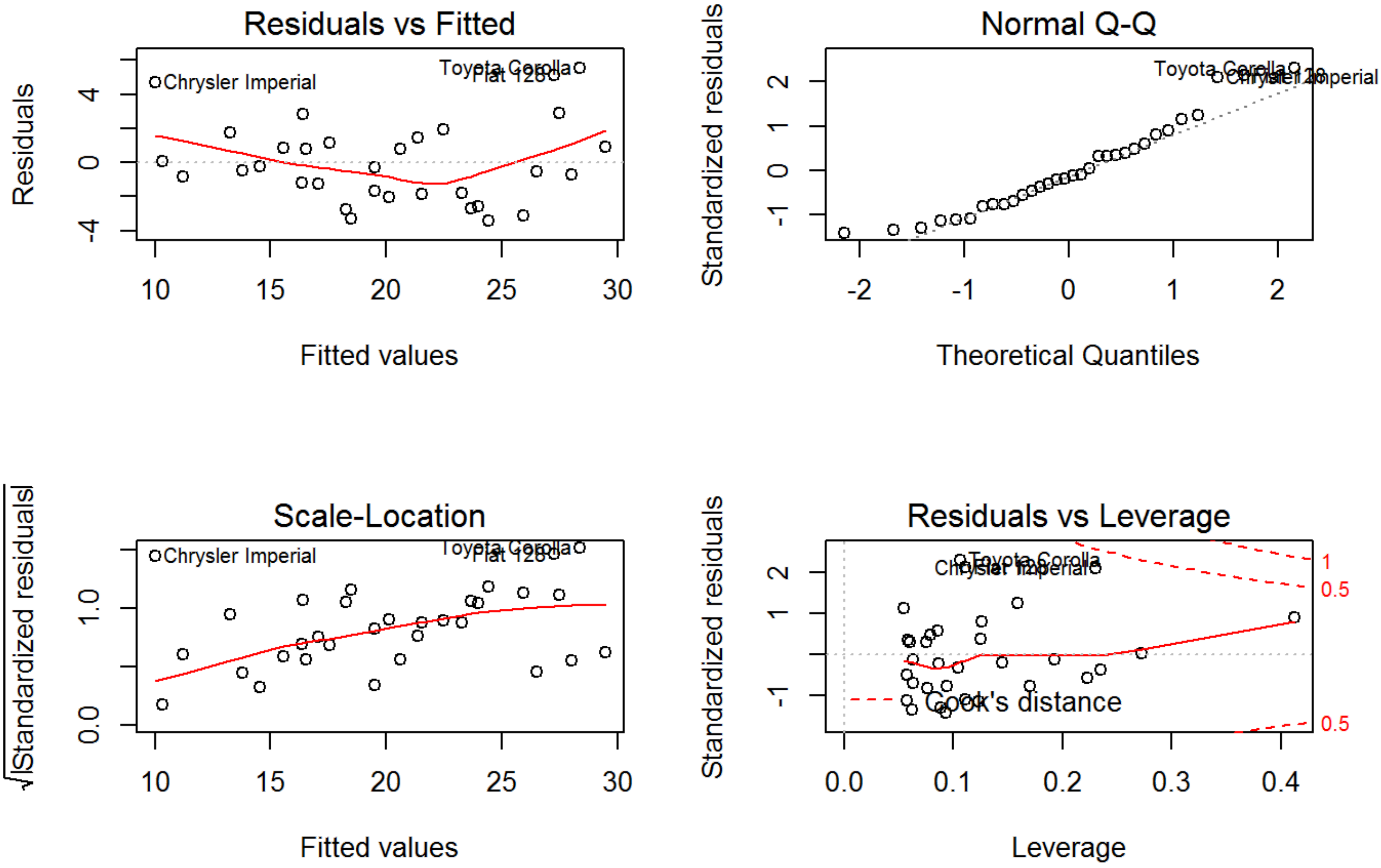
```
bestfit <- lm(mpg~am + wt + hp, data = mtcars)
anova(fit, bestfit)
```

```
## 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     30 721                            
## 2     28 180  2       541 42 3.7e-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.

```
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 
## -3.422 -1.792 -0.379  1.225  5.532 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 34.00288    2.64266   12.87  2.8e-13 ***
## am           2.08371    1.37642    1.51  0.14127    
## wt          -2.87858    0.90497   -3.18  0.00357 ** 
## hp          -0.03748    0.00961   -3.90  0.00055 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.54 on 28 degrees of freedom
## Multiple R-squared:  0.84,   Adjusted R-squared:  0.823 
## F-statistic:   49 on 3 and 28 DF,  p-value: 2.91e-11
```

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.