Regression Models Course Project

Relationship of Vehicle Transmission on MPG

Daniel Rimdzius

1/4/2020

Summary

This report is interested in exploring the relationship between a set of variables and miles per gallon (MPG). In particular I will show that a manual transmission car is better for gas mileage by roughly 1.8 mpg.

Exploratory Data Analysis

We will begin by loading the required packages and datasets, and take a look at the mtcars data.

```
require(ggplot2)
require(dplyr)
require(tidyr)
data(mtcars)
head(mtcars)
```

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

A new dataframe is created simply to use better-named variables. Discrete factors are properly changed, as well.

```
data <- data.frame(MilesPerGallon = mtcars$mpg,
             Cylinders = factor(mtcars$cyl),
             EngineShape = factor(mtcars$vs, levels = c(0,1), labels = c("V-shaped","Straight")),
             Transmission = factor(mtcars$am, levels = c(0,1), labels = c("Automatic", "Manual")),
             Gears = factor(mtcars$gear),
             Carburetors = factor(mtcars$carb),
             Weight = mtcars$wt,
             Displacement = mtcars$disp,
             Horsepower = mtcars$hp,
             RearAxleRatio = mtcars$drat,
             QuarterMile = mtcars$qsec)
```

Looking at the exploratory boxplot, PLOT1, in the Appendix, we can see there appears to be a clear difference in MPG between automatic and manual transmissions. Unadjusted for any other factors, the median MPG for manual transmissions appears to be NA versus the automatic transmission mpg median of NA.

PLOT2 in the Appendix shows the effects of each variable plotted against the Miles Per Gallon. Besides Transmission, we can see that there appear to be strong effects from Cylinders, Displacement, Horsepower and Weight.

Regression Models

To show that there is a significant difference between the automatic and manual transmissions related to mpg, we can perform a t-test:

```
AutomaticMPG <- data[data$Transmission == "Automatic",]$MilesPerGallon
ManualMPG <- data[data$Transmission == "Manual",]$MilesPerGallon
t.test(AutomaticMPG, ManualMPG)
```

```
##
##  Welch Two Sample t-test
##
## data:  AutomaticMPG and ManualMPG
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

We see there is a p-value of 0.001, and the 95% confidence interval is -3.2 to -11.3 which does not include 0. We can therefore reject the null hypothesis and conclude this is a significant variable. We can create a simple linear model for this:

```
fit1 <- lm(MilesPerGallon ~ Transmission, data)
summary(fit1)
```

```
##
## Call:
## lm(formula = MilesPerGallon ~ Transmission, data = data)
##
## 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 ***
## TransmissionManual    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
```

Again we see we can reject the null hypothesis. However, the R-squared value is only 0.3597989, which means the transmission only account for around 36% of the variance. We must add in additional variables to find a better fit.

Model Selection

As we saw above in our exploratory data analysis, and shown in PLOT2 in the Appendix, there are several other factors that appear to have a significant effect on mpg. We can setup multiple model fits and analyze the variance.

```
fit2 <- lm(MilesPerGallon ~ Transmission + Cylinders, data)
fit3 <- lm(MilesPerGallon ~ Transmission + Cylinders + Weight, data)
fit4 <- lm(MilesPerGallon ~ Transmission + Cylinders + Weight + Horsepower, data)
fit5 <- lm(MilesPerGallon ~ Transmission + Cylinders + Weight + Horsepower + Displacement, data)
anova(fit1,fit2,fit3,fit4,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: MilesPerGallon ~ Transmission
## Model 2: MilesPerGallon ~ Transmission + Cylinders
## Model 3: MilesPerGallon ~ Transmission + Cylinders + Weight
## Model 4: MilesPerGallon ~ Transmission + Cylinders + Weight + Horsepower
## Model 5: MilesPerGallon ~ Transmission + Cylinders + Weight + Horsepower +
##     Displacement
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 37.9300 2.678e-08 ***
## 3     27 182.97  1     81.53 13.5510  0.001118 **
## 4     26 151.03  1     31.94  5.3093  0.029801 *
## 5     25 150.41  1      0.62  0.1025  0.751489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see from the ANOVA table that the fit5 model has a P-value of 0.75, while fit4 and below are at 0.029 and below. Displacement is related to the number of cylinders, so we will disregard that term, and choose fit4 as our best-fit line.

```
BestFit <- fit4
summary(BestFit)
```

```
##
## Call:
## lm(formula = MilesPerGallon ~ Transmission + Cylinders + Weight +
##     Horsepower, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        33.70832    2.60489  12.940 7.73e-13 ***
## TransmissionManual  1.80921    1.39630   1.296  0.20646
## Cylinders6         -3.03134    1.40728  -2.154  0.04068 *
## Cylinders8         -2.16368    2.28425  -0.947  0.35225
## Weight             -2.49683    0.88559  -2.819  0.00908 **
## Horsepower         -0.03211    0.01369  -2.345  0.02693 *
## ---
## 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:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

```
shapiro.test(BestFit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  BestFit$residuals
## W = 0.96807, p-value = 0.4479
```

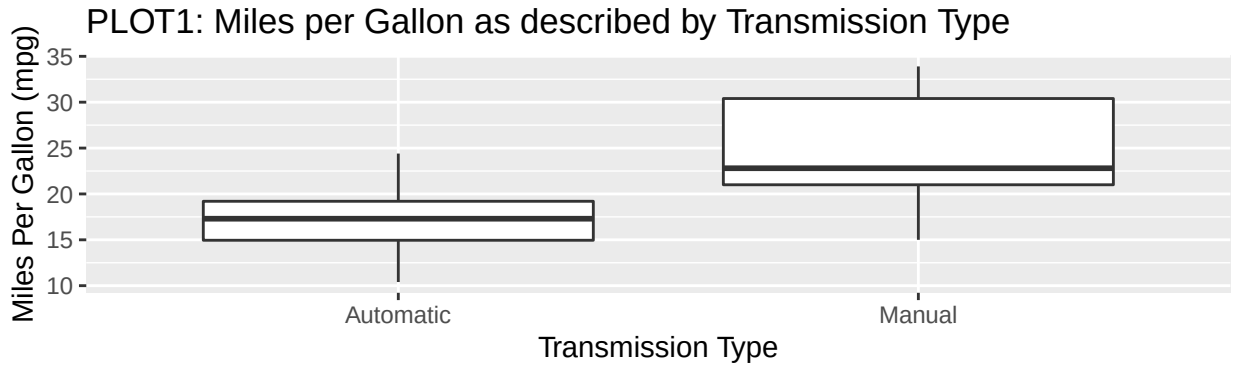
The best fit model has an R-squared value of 0.8659, accounting for about 87% of the variance and a strongly significant p-value. We see in PLOT3 in the Apendix the plot of the residuals for this model. They are normally distributed and do not show any heteroskedasticity. We also see in the shapiro test a high P-value (» 0.05) which fails to reject normality, supporting our confidence in the analysis of variance.

Conclusions

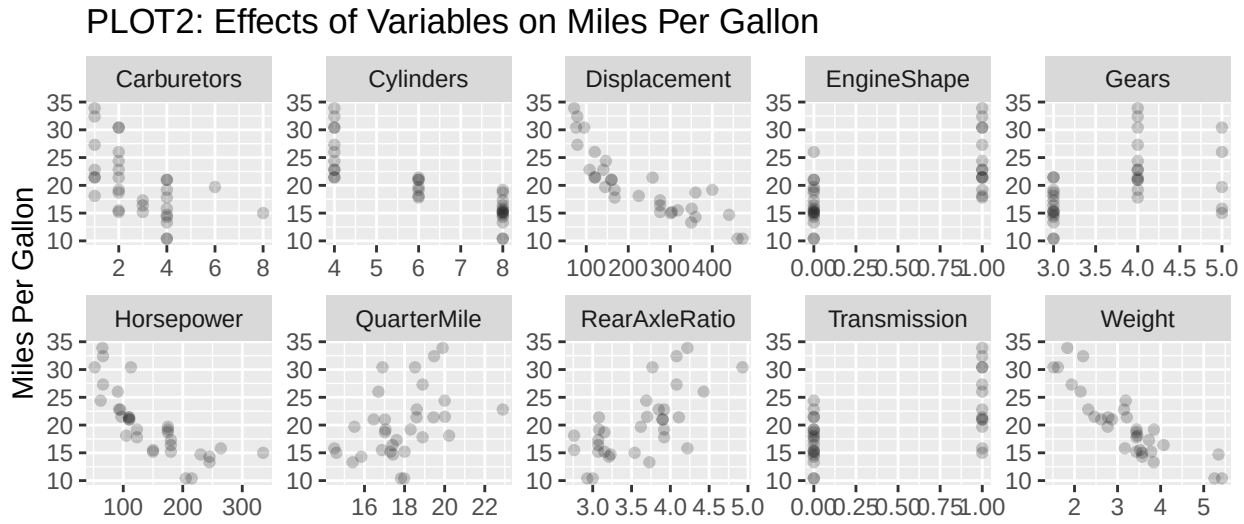
From this best fit model, we can conclude that a manual transmission car is better for MPG. The coefficient shown in the BestFit summary for the Transmission tells us the difference between manual and automatic transmissions is 1.806 mpg.

Appendix

```
ggplot(data, aes(x = Transmission, y = MilesPerGallon)) +
      geom_boxplot() +
      labs(x = "Transmission Type",
           y = "Miles Per Gallon (mpg)",
           title = "PLOT1: Miles per Gallon as described by Transmission Type")
```




```
levels(data$EngineShape) <- 0:1
levels(data$Transmission) <- 0:1
data %>% gather(-MilesPerGallon, key = "Factor", value = "Value") %>% mutate(Value = as.numeric(Value)) %>%
      ggplot(aes(x = Value, y = MilesPerGallon)) +
      geom_point(alpha = 0.2) +
      facet_wrap(~ Factor, scales= "free", ncol = 5) +
      labs(x = "", y = "Miles Per Gallon", title = "PLOT2: Effects of Variables on Miles Per Gallon")
```




```
par(mfrow = c(2,2))
plot(BestFit)
title(main = "PLOT3: Residual Plots for the Best Fit Model", sub = "", outer = TRUE, adj = 0.125, line =
```