Impact of Transmission Type on Mileage: Regression Analysis Using mtcars Dataset

Ajay Ghanti

12/2/2016

Overview

This report is an analysis conducted for *Motor Trend*, a magazine about the automobile industry. We are considering a dataset of a collection of cars, and are interested in exploring the relationship between a set of variables and miles per gallon (MPG). We are particularly interested in answering the following two questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions?

Data Processing

The dataset we are using, **mtcars** comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). Below is a description of the variables in the dataset.

Var	Meaning
mpg	Miles per (US) Gallon
cyl	Number of Cylinders
disp	Displacement (cu.in.)
hp	Gross Horsepower
drat	Rear Axle Ratio
wt	Weight (1000 lbs)
qsec	1/4 Mile Time (Acceleration)
vs	V Engine/Straight Engine
am	Transmission (Automatic/Manual)
gear	Number of Forward Gears
carb	Number of Carburetors

```
data(mtcars)
#look ahead; compute correlation matrix before we convert data to factors
cor <- round(cor(mtcars), 2)

#convert transmission type to factor
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Auto", "Manual")
#convert engine mount type to factor (V vs. Straight Engine)
mtcars$vs <- as.factor(mtcars$vs)
levels(mtcars$vs) <- c("V", "S")
```

Exploratory Data Analysis

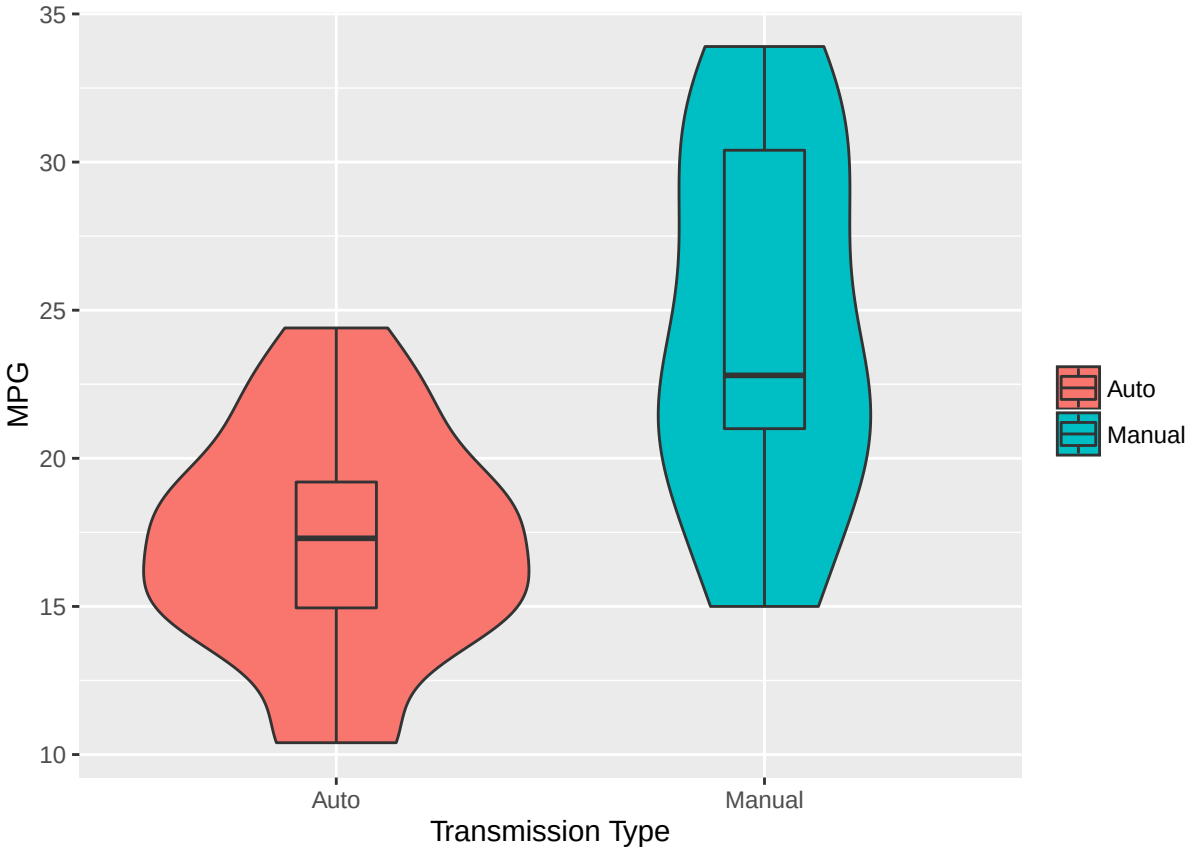
Let us start with some exploratory data analysis, to observe patterns between variables in the dataset. Here is a sample first few rows of the data.

```r
head(mtcars, 10)
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs     am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  V Manual    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  V Manual    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  S Manual    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  S   Auto    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  V   Auto    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  S   Auto    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  V   Auto    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  S   Auto    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  S   Auto    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  S   Auto    4    4
```

The outcome is mileage (mpg), and we want to study the effect of transmission type on the mileage; so transmission type (am) is the predictor. Let's plot the transmission type against the mileage.

```r
ggplot(mtcars, aes(am, mpg, fill = am)) +
    geom_violin() +
    geom_boxplot(width=.25) +
    xlab("Transmission Type") +
    ylab("MPG") +
    guides(fill=guide_legend(title=""))
```



From the plot, it can be seen that manual transmission cars have a higher average mileage than automatics. But based on this alone, we cannot conclude as there are other variables in the dataset, which might impact the mileage. We will need to perform further testing, starting with a hypothesis test.

Let us state our null hypothesis (H_0) and alternate hypotheses (H_a)

$\boldsymbol{H_0: X_{Auto} = X_{Manual}}$
There is NO difference in mileage (mpg) for each transmission type (automatic vs. manual)

$\boldsymbol{H_a: X_{Auto} \sim= X_{Manual}}$
There IS a difference in mileage (mpg) for each transmission type (automatic vs. manual)

```
# Conduct a two-sided t-test on the data
result <- t.test(mpg ~ am, data=mtcars, var.equal=FALSE, paired=FALSE)
result
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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 in group Auto mean in group Manual
##             17.14737             24.39231
```

From the t-test, we obtain a p-value of 0.001374, which is statistically significant (< 0.05); so we reject the null hypothesis, i.e., there is indeed a difference in mileage due to the transmission type. Further it is also seen that the average mileage for automatic cars (17.1474) is lesser than the mileage for manual transmission cars (24.3923).

Despite this test, which ascertains the influence of transmission type on mileage, the test actually assumes that all other variables are held constant, which is not the case; hence, we will need to proceed with regression analysis.

Regression Analysis - Model Selection for Best Fit

Simple Linear Regression

Let us first fit a linear model with only one predictor - the transmission type (am).

```
fitOne <- lm(mpg ~ am, mtcars)
summary(fitOne)
```

```
##
## 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 ***
## amManual       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
```

As seen from the above model, the intercept is 17.1474, and the coefficient for amManual is 7.2449, which means the mean mpg for manual cars is 7.2449 miles more than that for automatic cars. Also, the p-value is 2.85×10^{-4} which is statistically significant (< 0.05). However, the R^2 is 0.3598, i.e, this model only explains 35.98% of the variance.

In order to quantify the difference in mpg between automatic and manual cars, we will need to obtain the 'best fit' model with the right predictors.

Multivariable Linear Regression

The first thought would be to fit a model by adding all variables as predictors. But, we cannot add all of them to our regression model, as it might introduce the problem of collinearity and overfitting. So, lets look at the correlation matrix to figure out what predictors can go into our model.

```
#get the lower triangle of the correlation matrix
cord <- as.dist(cor)
#filter out values that are not significant
cord[which(abs(cord) < 0.70)] = NA
cord
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear
## cyl  -0.85
## disp -0.85  0.90
## hp   -0.78  0.83  0.79
## drat    NA -0.70 -0.71    NA
## wt   -0.87  0.78  0.89    NA -0.71
## qsec    NA    NA    NA -0.71    NA    NA
## vs      NA -0.81 -0.71 -0.72    NA    NA  0.74
## am      NA    NA    NA    NA  0.71    NA    NA    NA
## gear    NA    NA    NA    NA  0.70    NA    NA    NA  0.79
## carb    NA    NA    NA  0.75    NA    NA    NA    NA    NA    NA
```

By looking at the correlation table, here are some relationships we observe.

- Mileage (mpg) is influenced by weight (wt), # of cylinders (cyl), engine displacement (disp), & power (hp)
- But # of cylinders (cyl) and engine displacement (disp) are very strongly correlated
- Weight (wt) is also strongly correlated with # of cylinders (cyl) & engine displacement (disp)
- Along with the above, engine type (vs) & # of gears (gear) would bear impact on power (hp), which would in turn strongly influence acceleration (qsec - 1/4 mile time)

Based on these observations, it is evident that some of these variables do indeed bear influence on mileage, as well as each other in various ways.

We can try fitting a few models with different combinations of the above variables as predictors, and then comparing those to find the best model fit. However, R provides a simpler way - the step() function which chooses a formula-based model by AIC in a stepwise algorithm, including the predictors that best explain the regression.

```
#start with a model with all variables, and select the best one in a stepwise algorithm
fitBest <- step(lm(mpg ~ ., mtcars), trace=0)
summary(fitBest)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
anova(fitOne, fitBest)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Inference

The 'best fit' model includes the weight (wt), acceleration (qsec - 1/4 mile time), and the transmission type (am). Weight and mpg have a negative relation, whereas acceleration and transmission type change positively.

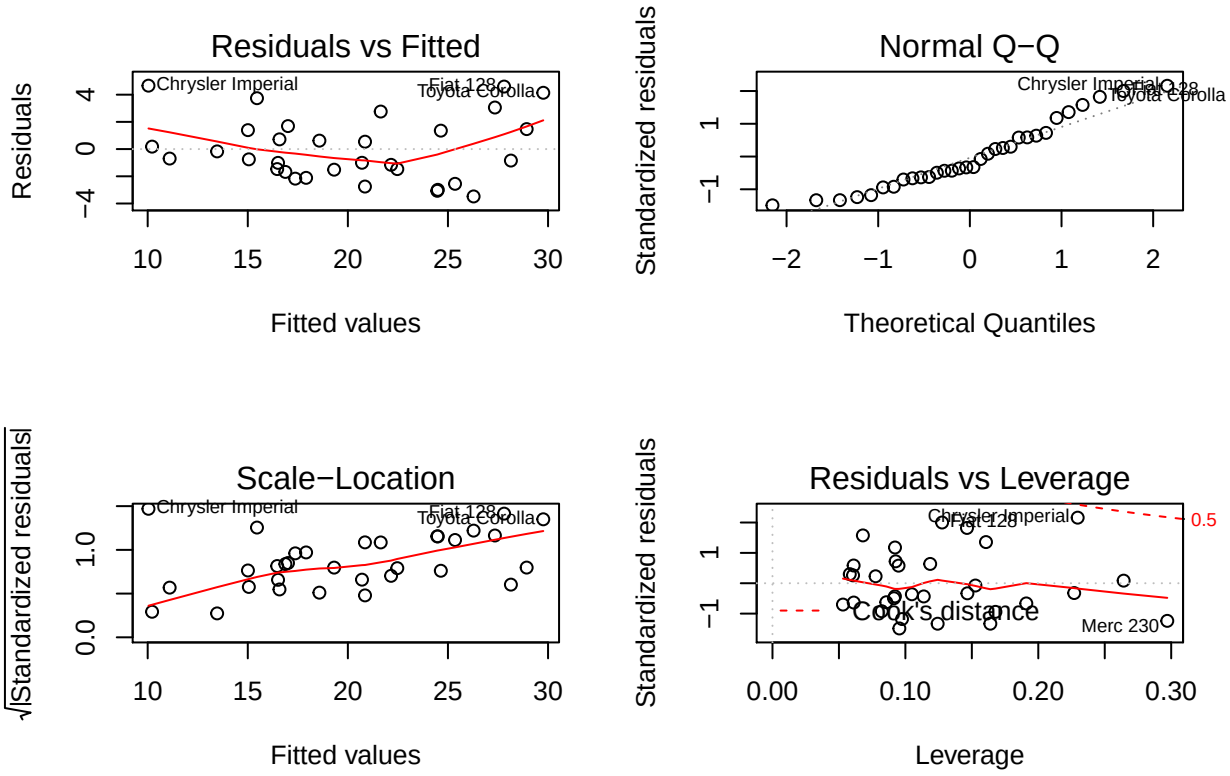
The coefficient for amManual is 2.9358, which means the mean mpg for manual cars is about 2.94 miles more than that for automatic cars. Also, the p-value is 0.046716 which is statistically significant (< 0.05). The R^2 is 0.8497, i.e, this model satisfactorily explains 84.97% of the variance.

The p-value in the ANOVA table also proves our claim that our multivariable regression model is significantly different from the simple one.

Residuals and Diagnostics

In order to validate our model better, let us plot the residual and diagnostic plots and analyse them.

```
#set a 2x2 panel
par(mfrow=c(2,2))
plot(fitBest)
```



- **Residuals vs. Fitted**: The residuals seem to be fairly equally spread out around the horizontal line, indicating that there are no non-linear patterns between predictors and outcome; there are a few outliers, though;
- **Normal Q-Q**: The residuals follow along a straight line, showing that they are normally distributed;
- **Scale-Location**: This spread-location plot shows equally randomly spread points, indicating that the variability of residuals is homoscedastic;
- **Residuals vs Leverage**: There aren't any influential points outside of the Cook's distance, so the outliers we suspected earlier don't seem to be influential cases;

Summary

From our 'best fit' model, we can conclude that the multivariable regression model accounts for most of the variance (84.97%) of the mileage (mpg), after adjusting for weight (wt) and acceleration (qsec).

We also arrive at the following answers to our earlier questions:

1. Is an automatic or manual transmission better for MPG?
   - **Manual transmission gives better mileage than an automatic** (after adjusting for weight and acceleration)
2. Quantify the MPG difference between automatic and manual transmissions?
   - On an average, manual transmission cars have a mileage of **2.94 miles more** than automatic cars