Motor Trend MPG Regression Model Analysis

Trent Parkinson

January 15, 2018

Overview

This project explores the `mtcars` data set and explores how miles per gallon (MPG) is affected by different variables, specifically the affect automatic and manual transmissions have on MPG. The following will be answered,

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

Setting up environment

Necessary libraries for loading, plotting, and model selection. Reading the `mtcars` dataset and making a copy in a `data.table`.

```
library(data.table)
library(ggplot2)
library(leaps)
library(printr)

data("mtcars")
mtcars_num <-copy(mtcars)
```

Data Structure

Viewing `mtcars` data, and viewing structure of variables.

```
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

```
as.data.frame(t(apply(mtcars,2,class)))
```

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric

Data Processing

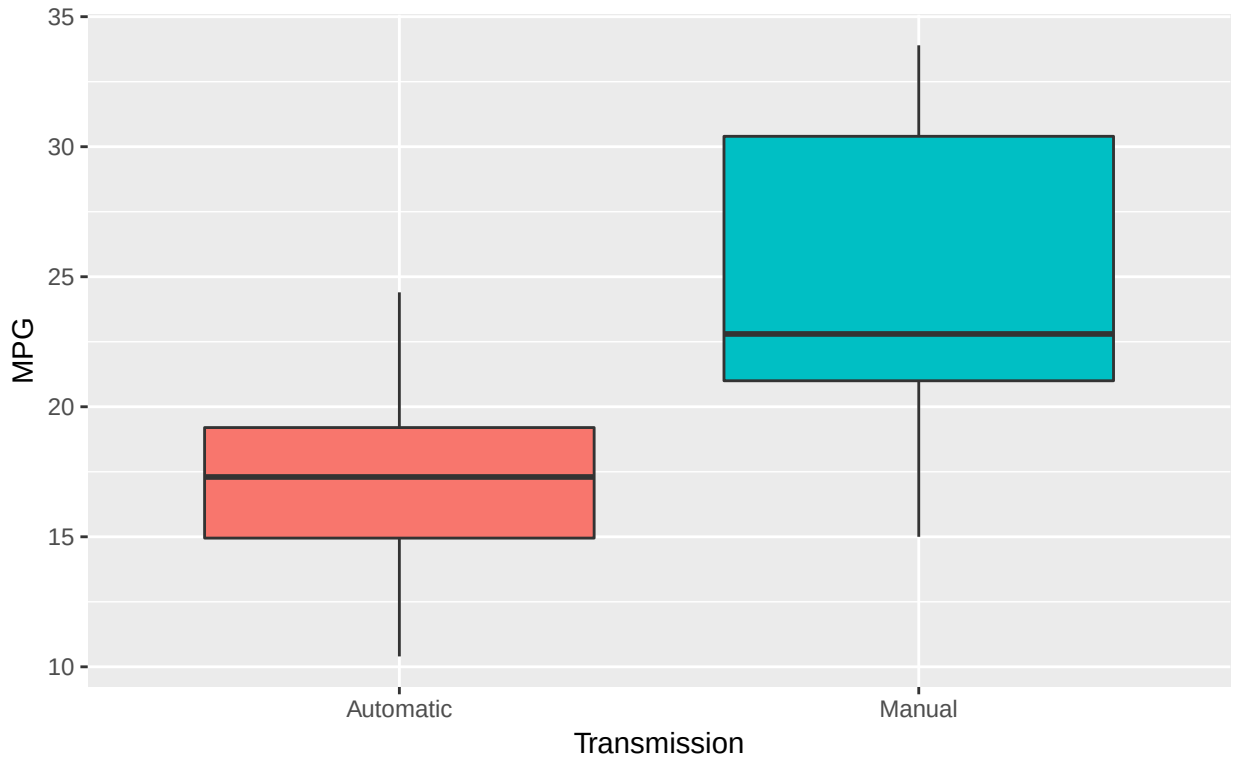
Changing categorical variables to factors. Relabeling `am` to `Automatic` and `Manual`.

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels = c("Automatic","Manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

Visualizations

Plotting the miles per gallon (MPG) for automatic and manual transmissions.

```
plot1 <- ggplot(mtcars, aes(x=am, y=mpg)) +
    geom_boxplot(aes(fill = am)) +
    xlab("Transmission") +
    ylab("MPG") +
    theme(legend.position = "none")
plot1
```



Analysis

It looks like there is a definite difference in the type of transmission for MPG. Performing a t-test will help verify if the difference in means is significant.

```
auto_vs_manu_ttest <- t.test(mpg ~ am, mtcars)
auto_vs_manu_ttest
```

```
##
##  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 Automatic    mean in group Manual
##                17.14737                24.39231
```

The t-test rejected the null-hypothesis that the difference in means is equal to zero, with a p-value of .0014. Therefore there is a difference in transmission type, with manual transmissions having a higher MPG.

Linear Regression Fitting

Since the project is trying to quantify the difference in MPG for automatic and manual transmissions. The best starting place is a simple linear model with transmission type as the dependent variable.

```
basic_fit <- lm(mpg ~ am, mtcars)
summary(basic_fit)$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
amManual	7.244939	1.764422	4.106127	0.000285

```
summary(basic_fit)$r.squared
```

```
## [1] 0.3597989
```

The basic linear model with `am` as the only regressor explains 36% of the variation, not a very good model. To gain a better model it gets tricky after one variable, since regressors can correlate with not only the predictor but also other regressors adding a variable that is highly correlated could help, but could also hurt the prediction.

One method is called stepwise regression which uses AIC to choose the best model, the other method is called best subsets regression which goes through all possible models with the specified regressors and chooses the best model based on different criterion.

```
everything_fit <- lm(mpg ~ ., mtcars)
step_fit <- step(everything_fit,direction="both",trace=FALSE)

best_subset <- regsubsets(mpg ~ ., mtcars, nvmax = 25)
best_subset_summary <- summary(best_subset)
adjr2 <- which.max(best_subset_summary$adjr2)
cp <- which.min(best_subset_summary$cp)
bic <- which.min(best_subset_summary$bic)
best_set <- best_subset_summary$outmat[c(adjr2,cp),]
best_set[,1:13]
```

	cyl6	cyl8	disp	hp	drat	wt	qsec	vs1	amManual	gear4	gear5	carb2	carb3
5 (1)	*			*		*		*	*				
3 (1)						*	*		*				

```
sub3_fit <- lm(mpg ~ am + wt + qsec, mtcars)
sub5_fit <- lm(mpg ~ am + cyl + hp + wt + vs, mtcars)
```

Model Selection

Stepwise regression gave us a best model, but best subsets gave us two different models as well. Using Mallows's C_p and BIC both returned model three as the best, while model five has the best for the adjusted R^2. The code below grabs the adjusted R^2 and also the p-value for the transmission type in the regression coefficients. Since the goal of the project is to quantify MPG, the best model would have confidence in this coefficient as well as explain the variance well.

```
models <- c("mpg ~ am + wt + qsec", "mpg ~ am + wt + cyl + hp", "mpg ~ am + wt + cyl + hp + vs")
adj_r_squared <- round(c(summary(sub3_fit)$adj.r.squared,
                         summary(step_fit)$adj.r.squared,
                         summary(sub5_fit)$adj.r.squared),4)
amManual_Pvalues <- round(c(summary(sub3_fit)$coefficients["amManual",4],
                            summary(step_fit)$coefficients["amManual",4],
                            summary(sub5_fit)$coefficients["amManual",4]),4)
results <- as.data.frame(cbind(models,adj_r_squared,amManual_Pvalues))
results
```

models	adj_r_squared	amManual_Pvalues
mpg ~ am + wt + qsec	0.8336	0.0467
mpg ~ am + wt + cyl + hp	0.8401	0.2065
mpg ~ am + wt + cyl + hp + vs	0.8418	0.1032

Checking Model

The only model with a p-value for transmission type below 5% is `mpg ~ am + wt + qsec`, it doesn't have the highest adjusted R^2 but its not much lower than the other two models.

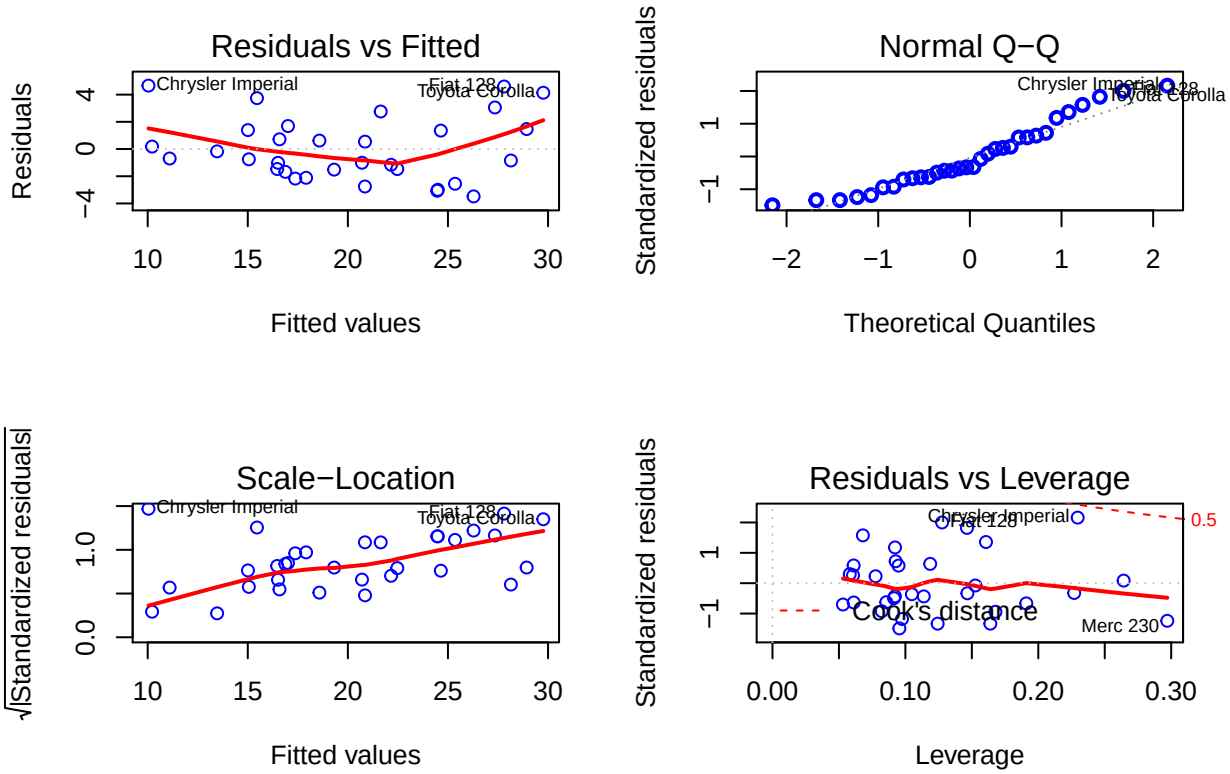
```
summary(sub3_fit)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, 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
## amManual      2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## 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
```

Everything so far looks solid, but lets make sure this model fits our data well by printing the diagnostic plots.

```
par(mfrow = c(2,2))
plot(sub3_fit, col = "blue", lwd = 2)
```



- Residuals vs Fitted: The points are randomly scattered, but may have a slight non-linear relationship.
- Normal Q-Q: The points pass normality, they deviate slightly from the diagonal, but they follow the diagonal fairly close.
- Scale-Location: The upward slope line is worrisome, the residues spread slightly wider.
- Residuals vs Leverage: No high leverage points.

Conclusions

The best transmission type for MPG would have to be the manual transmission. Its confirmed by the t-test, as well as our final linear model. By having a manual transmission instead of an automatic the MPG will increase by 2.94 as can be seen in the best model's `amManual` coefficient.

The model fit well with a $p < 0.05$ and and $R^2 = 0.85$, but the diagnostic plots did warn us that something may be missing in our model. I believe the true cause for these trends are do to the small sample size with little overlap on the parameters `wt` and `qsec`.