Motor Trend Data Analysis Report (Regression Models Project)

Kevin O'Brien

Executive Summary

In this report we explore the relationship between miles per gallon (MPG) and a set of variables from the Motor Trends car dataset. We are particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG?
- How can we quantify the difference in MPG between manual and automatic transmissions?

The data was taken from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 characteristic of automobile design and performance for 32 automobiles (1973–74 models). We use regression models and exploratory data analyses to mainly explore how *automatic* (***am = 0***) and *manual* (***am = 1***) transmissions features affect the ***mpg*** feature.

The t-test shows that the performance difference between cars with automatic and manual transmission. And it is about 7 MPG more for cars with manual transmission than those with automatic transmission. Then, we fit several linear regression models and select the one with highest Adjusted R-squared value.

So, given that weight and 1/4 mile time are held constant, manual transmitted cars are 14.079 + (-4.141)*weight more MPG (miles per gallon) on average better than automatic transmitted cars. Thus, cars that are lighter in weight with a manual transmission and cars that are heavier in weight with an automatic transmission will have higher MPG values.

Prepare the R Environment and the Dataset

In advance of the analysis, we will load some useful R packages. Also, we load the data set `mtcars` and change some variables from `numeric` class to `factor` class.

```
data(mtcars)
modeldata <- mtcars #Save Raw Dataset
```

Exploratory Data Analysis

In this section, we shall do some basic exploratory data analyses. Please refer to the **Appendix: Figures** section for the plots. According to the box plot, we see that manual transmission yields higher values of MPG in general. And as for the pair graph, we can see some higher correlations between variables like "wt", "disp", "cyl" and "hp".

Inference

At this step, we make the null hypothesis as the MPG of the automatic and manual transmissions are from the same population (assuming the MPG has a normal distribution). We use the two sample T-test to demonstrate.

```
result <- t.test(modeldata$mpg ~ modeldata$am)
```

Since the p-value is 0.0013736, we reject our null hypothesis. Therefore the automatic and manual transmissions are value drawn from two populations with different mean values. The mean mileage per gallon of manual transmitted cars is about 24.3923077, 17.1473684 more than that of automatic transmitted cars.

Regression Analysis

First, we fit the full model as the following.

```
fullModel <- lm(mpg ~ ., data=mtcars)
```

```
#Results Hidden
fullModel <- lm(mpg ~ ., data=mtcars)
summary(fullModel)
```

This model has the Residual standard error as 2.650197 on 21 degrees of freedom. And the Adjusted R-squared value is 0.8066, which means that the model can explain about 78% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significant level.

Model Selection To improve model fit, we use backward selection to select some statistically significant variables.

```
stepModel <- step(fullModel, k=log(nrow(mtcars)))
summary(stepModel) # results hidden
```

The resulting model is defined as "mpg ~ wt + qsec + am". This model has a Residual standard error as 2.459 on 28 degrees of freedom. The Adjusted R-squared value is 0.8497, which means that the model can explain about 85% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

Please refer to the **Appendix: Figures** section for the plots again. According to the scatter plot, it indicates that there appear to be an interaction term between "wt" variable and "am" variable, since automatic cars tend to weigh heavier than manual cars.

Thus, we have the following model including the interaction term:

```
amIntWtModel<-lm(mpg ~ wt + qsec + am + wt:am, data=modeldata)
summary(amIntWtModel) # results hidden
```

This model has the Residual standard error as 2.084 on 27 degrees of freedom. The Adjusted R-squared value is 0.8959, which means that the model can explain about 90% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level. This is a very satisfactory model

Next, we fit the simple model with MPG as the outcome variable and Transmission (i.e. ***am***) as the only predictor variable.

```
# results hidden
amModel<-lm(mpg ~ am, data=modeldata)
summary(amModel)
```

The output shows that on average, a car has 17.147 mpg with automatic transmission, and if it is manual transmission, 7.245 mpg is increased.

This model has the Residual standard error as 4.902 on 30 degrees of freedom. The Adjusted R-squared value is 0.3385, which means that the model can explain about 34% of the variance of the MPG variable. The low Adjusted R-squared value also indicates that we should add other variables to the model.

Finally, we select the final model.

We end up selecting the model with the highest Adjusted R-squared value, "mpg ~ wt + qsec + am + wt:am".

```
summary(amIntWtModel)$coef %>% kable(digits=4)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	23.8025	6.0768	3.9169	0.0006
wt	-7.0779	1.0946	-6.4664	0.0000
qsec	1.0170	0.2520	4.0354	0.0004
amManual	-14.0794	3.4353	-4.0985	0.0003
wt:amManual	4.1414	1.1968	3.4603	0.0018

Thus, the result shows that when "wt" (weight lb/1000) and "qsec" (1/4 mile time) remain constant, cars with manual transmission add 14.079 + (-4.141)*wt more MPG (miles per gallon) on average than cars with automatic transmission. That is, a manual transmitted car that weighs 2000 lbs have 5.797 more MPG than an automatic transmitted car that has both the same weight and 1/4 mile time.

Residual Analysis and Diagnostics

Please refer to the **Appendix: Figures** section for the plots. According to the residual plots, we can verify the following underlying assumptions:

1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.

2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line.

3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.

4. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

Therefore, the above analyses meet all basic assumptions of linear regression and well answer the questions.

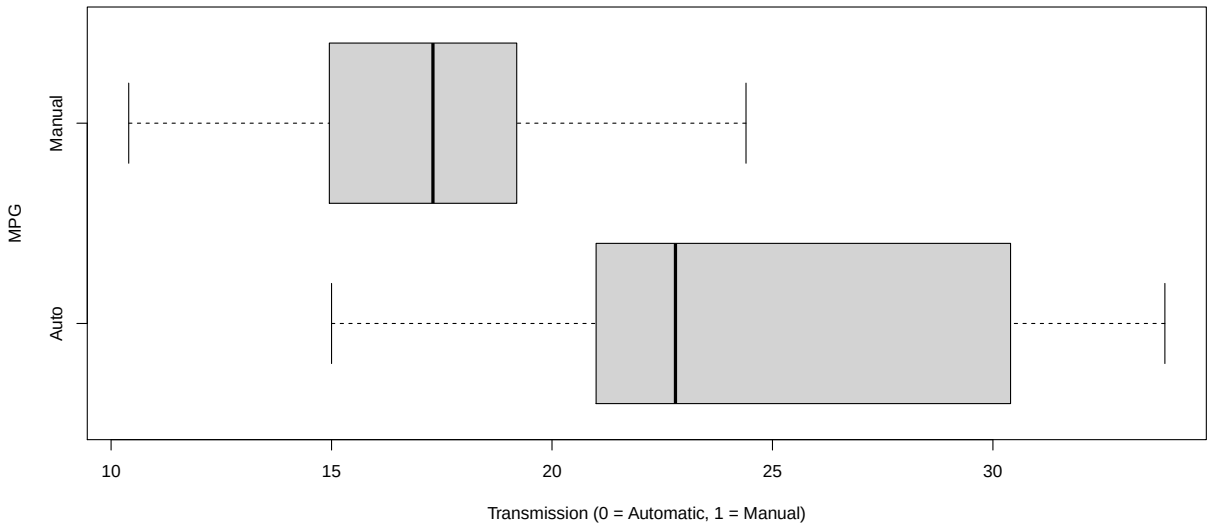
Appendix: Figures

1. Boxplot of MPG vs. Transmission

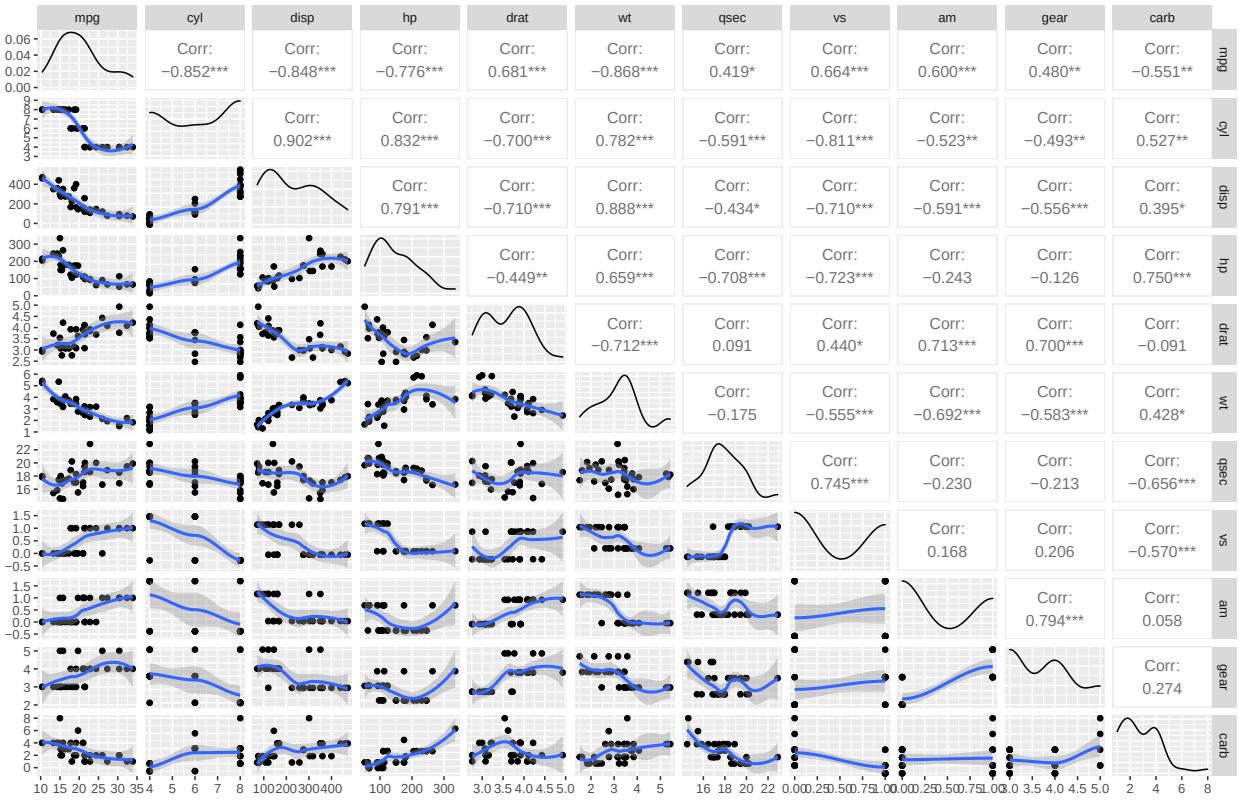
2. Pair Graph of Motor Trend Car Road Tests
3. Scatter Plot of MPG vs. Weight by Transmission

4. Residual Plots

Plot 1. Boxplot of MPG vs. Transmission

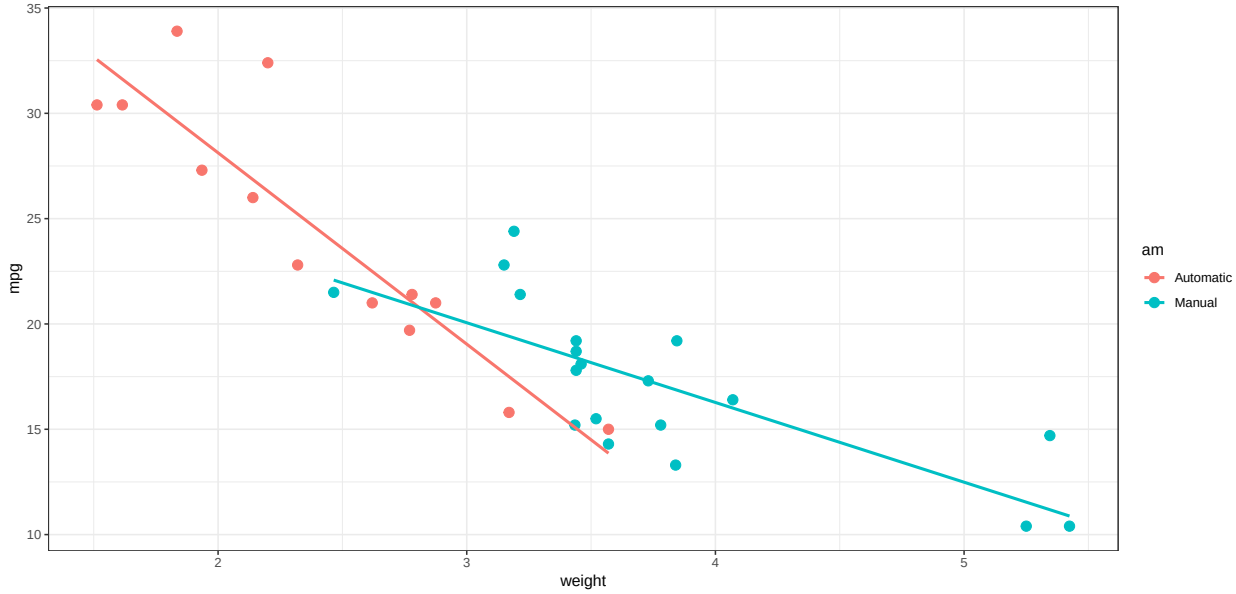


Plot 2: Pairs Plot




```
## `geom_smooth()` using formula 'y ~ x'
```

Plot 3: Scatter Plot of MPG vs. Weight by Transmission



Plot 4: Residual Plots