Coursera - Regression Models - Project 1

Charles Bradbury

Executive Summary

This report will provide analysis on the `mtcars` data set and explore the relationship between the variables of that data set and miles per gallon (MPG). This report will attempt to answer two basic questions:

1. Is an automatic or manual transmission better for miles per gallon (MPG)?
2. Quantify the miles per gallon difference between automatic and manual transmissions.

The results, as our analysis below will show, is that:

1. Cars with manual transmissions provide a higher miles per gallon ratio compared to ones with automatic transmissions.
2. There is a sginifcant difference in the means and medians for MPG between vehicles with manual vs. automatic transmissions.

Load needed libraries/packages

```
library(ggplot2)
```

Data Loading & Processing

Let's load the data set `mtcars` and transform some variables to factors to make data processing more simple moving forward.

```
data(mtcars)
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

Exploratory Data Analysis

In this section we will do a more granular analysis of our data and explore various relationship between the variables of interest. You can see in *Figure 1* in the appendix that several variable including `cyl`, `disp`, `hp`, `drat`,`am`, `vs` and `wt` have recognizable affects on miles per gallon. We will use linear models in order to show these correlations below.

Being that once focus area for this report is MPG, we can show the relationship between MPG and the transmission type by using a boxplot (see *Figure 2* in the appendix). This plot shows us that there is an increase in MPG when the the car in question has a manual transmission.

Inference

At this step, we need to make some inferences based on a simple T-test.We will assume the null hypothesis as the MPG of the automatic and manual transmissions are from the same population (assuming the MPG has a normal distribution).

```
result <- t.test(mpg ~ am)
result$p.value
result$estimate
```

Since the p-value is 0.001373, we wil reject the null hypothesis. This tells us that automatic and manual transmissions are from different populations. The mean/average for auomatic transmission (group 0) is 17.15 miles per gallon and the mean/average for manual transmissions (group 1) is 24.40 miles per gallon. This is a ~7.25 miles per gallon difference in favor of the vehicles with a manual transmisson.

Regression Analysis

In this portion of the report, we will build linear regression models based on the different variables in order to determine the best model fit and show comparisons with the base model which we will obtain using "anova".

```
initialModel <- lm(mpg ~ ., data=mtcars)
bestModel <- step(initialModel)
```

The best model based on the above computations is the one that consists of variables `cyl`, `wt`, and `hp` as confounders and `am` as the independent variable.

Now, let's look at the details of the model we have chosen:

```
summary(bestModel)
```

This model is "mpg ~ cyl + hp + wt + am". It has the Residual standard error as 2.41 on 26 degrees of freedom. The Adjusted R-squared value is 0.8401, which means that the model can explain about 84% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

Now, let's fit the base model with only `am` as the predictor variable and the best model which we chose above.

```
amBaseModel <- lm(mpg ~ am, data=mtcars)
anova(amBaseModel, bestModel)
```

If we look a the above results, the p-value is higly significant which tells us we should reject the null hypothesis that the confounder variables `cyl`, `hp`, and `wt` do not contribute to the accuracy of the model.

Residual Analysis and Diagnostics

Please refer to *Figure 3* in the appendix to see a residual plot of our regression analysis. From these plots we can make the following observations:

- The randomness of the scatter plot points on teh "Residuals vs. Fitted" indicate the suspected independance.
- The "Normal Q-Q"" plot shows us that the points fall mostly on the line showing the residuals are normally distributed.

- The “Scale-Location” scatter plot shows us points scattered in a band patter, showing contant variance.
- There are distinct points of interest (our outliers) in the top right-hand corder of each of these plots.

We will investigate the top three outliers in each case of influence measurements:

```
outliers <- hatvalues(bestModel)
tail(sort(outliers),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##          0.2777872           0.2936819           0.4713671
```

```
influence <- dfbetas(bestModel)
tail(sort(influence[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

In examining the vehicles above, it is evidents that the analysis performed was accurate, as the same vehicles are shown in the resiudal plots.

Conclusion

Given the analysis above form our best fit modle, we can conclude that vehicles with manual transmissions perform better in miles per gallon compared with vehicles with automatic transmissions. Also, the miles per gallon will decrease by 2.5 for every 1,000 pounds increase in weight (wt). There is a slight, but apparent decrease in miles per gallon as horsepower (hp) increases. As for the amount of cylinders, as they increase from 4 to 6 then to 8 we see a decrease in miles per gallon by a factor of 3 and 2.2, respectiveley.

Appendix: Figures

Figure 1 - Pair Graph: Motor Trend Car Road Tests

```
pairs(mtcars, panel=panel.smooth, main="Pairs Plot for mtcars Data Set")
```

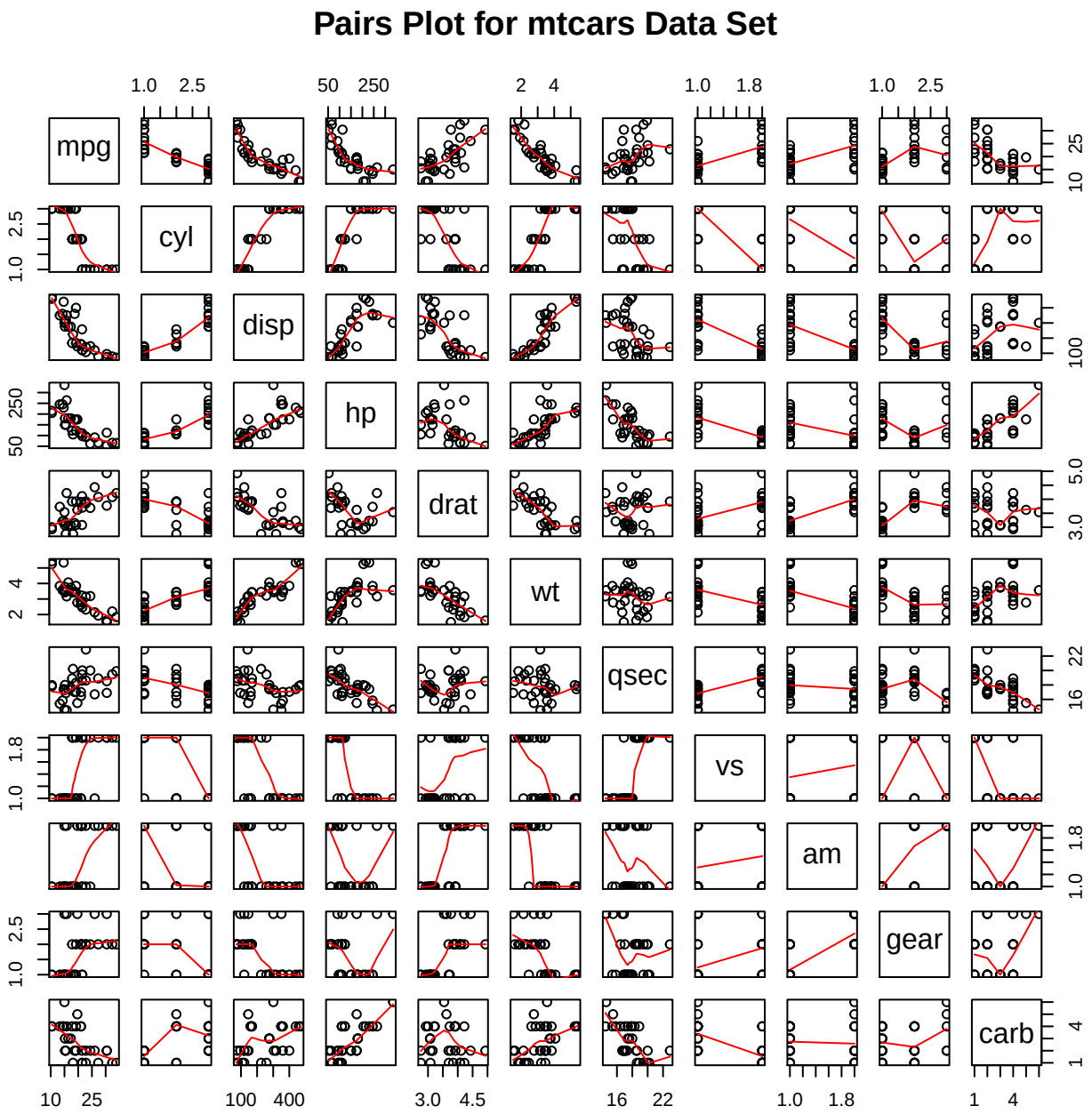
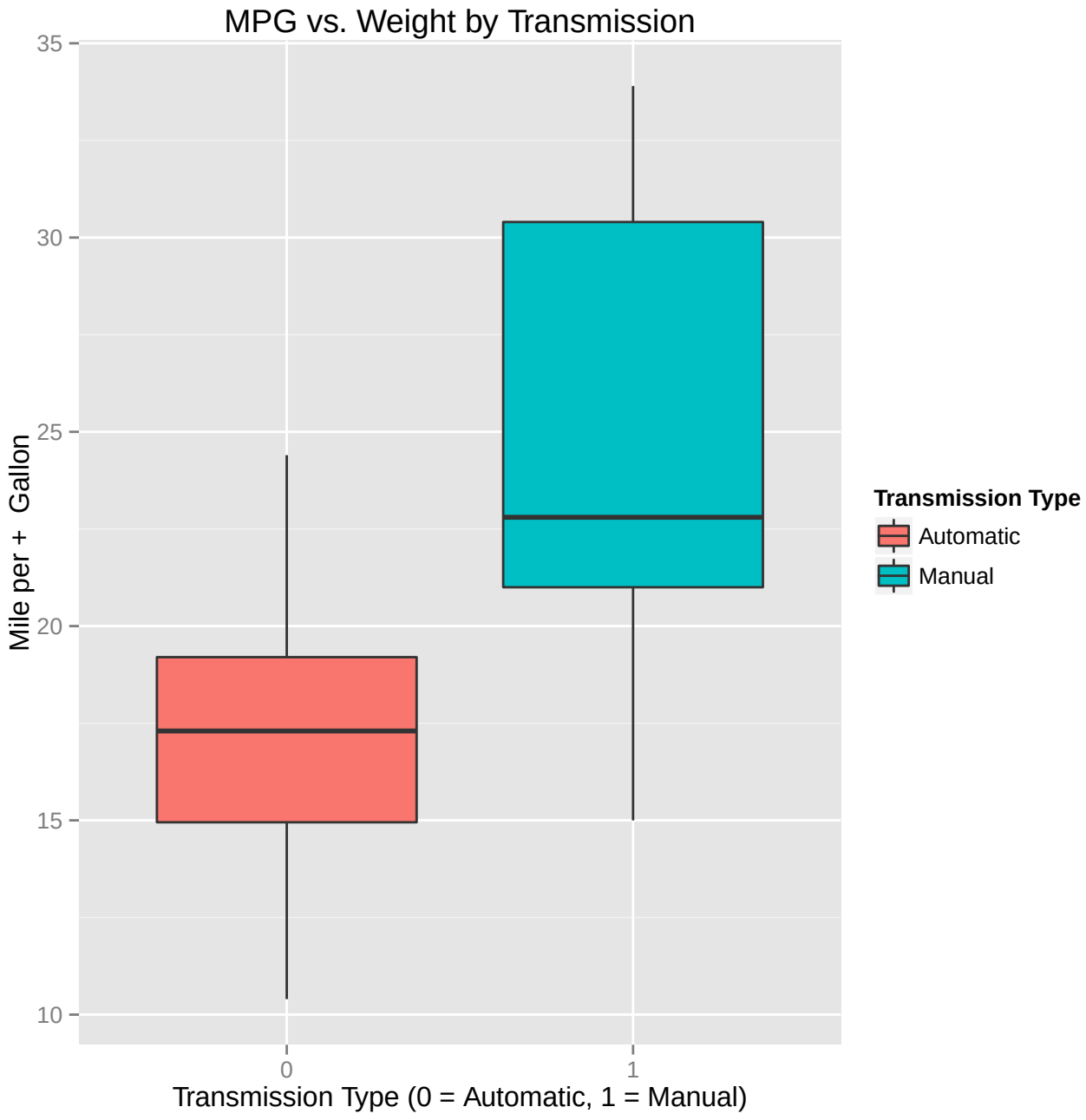


Figure 2 - Boxplot: MPG vs. Transmission

```
p <- ggplot(mtcars, aes(factor(am), mpg))
p + geom_boxplot(aes(fill = factor(am))) + xlab("Transmission Type (0 = Automatic, 1 = Manual)") +
        ylab("Mile per +  Gallon") + ggtitle("MPG vs. Weight by Transmission") +
        scale_fill_discrete(name="Transmission Type", labels=c("Automatic", "Manual"))
```



3. Residual Plots

```
residualData <- lm(mpg ~ cyl + hp + wt + am, data=mtcars)
par(mfrow = c(2, 2))
plot(residualData)
```