Building Regression Models from Motor Trend Data

Executive Summary

In this project, we analyze the `mtcars` data extracted from the 1974 Motor Trend US magazine, and study the relationship between transmission methods (**automatic** vs. **manual**) and miles per gallon (**MPG**). The data of 32 automobiles (1973–74 models) include their fuel consumption and other aspects of automobile design and their performance (MPG). We perform exploratory data analyses and build several regression models to study the impact of automatic and manual transmissions on MPG. Several linear regression models are built and the one with the best adjusted R-squared value is selected. The model reveals that cars have higher MPG values when they are either lighter in weight with a manual transmission or heavier in weight with an automatic transmission.

Exploratory Data Analysis

We first load the data set `mtcars` and convert several variables from `numeric` class to `factor` class.

```
library(ggplot2)
data(mtcars)
head(mtcars) # results hidden due to space constraint
dim(mtcars) # results hidden due to space constraint
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)
attach(mtcars)
```

Our basic exploratory data analyses are given in the **Appendix** section. In summary, it can be seen from the box plot that automatic transmission method yields lower values of MPG. Additionally, higher correlations between variables such as *wt*, *disp*, *cyl* and *hp* can be seen from the pair graphs.

Inference

We perform sample T-test to validate the null hypothesis if the MPG of the automatic and manual transmissions are from the same population.

```
sampleTest <- t.test(mpg ~ am)
sampleTest$p.value  # results hidden due to space constraint
sampleTest$estimate # results hidden due to space constraint
```

The above null hypothesis can be rejected as the p-value is 0.00137. Further, the mean of MPG of manual transmitted cars (group 1) is higher than that of automatic transmitted cars (group 0).

Regression Analysis

Basic models

We first fit the full model as follows.

```
fullModel <- lm(mpg ~ ., data=mtcars)
summary(fullModel) # results are hidden due to space constraint
```

The residual standard error of this full model is 2.833 on 15 degrees of freedom, while the adjusted R-squared value is 0.779. Even though the model can explain about 78% of the variance of the MPG variable, none of the coefficients are at 0.05 significant level.

Hence, backward selection is then used to choose statistically significant variables. This model basically follows “mpg ~ wt + qsec + am”.

```
stepModel <- step(fullModel, k=log(nrow(mtcars)))
summary(stepModel) # results are hidden due to space constraint
```

The residual standard error of the above model is 2.459 on 28 degrees of freedom, while the adjusted R-squared value is 0.8336, i.e., the model can explain about 83% of the variance of the MPG variable. Further, all of the coefficients are now at 0.05 significant level.

The scatter plot in the **Appendix** indicates that there is an interaction term between variables *wt* and *am* as automatic transmitted cars often weigh heavier than manual transmitted cars. Therefore, we fit the following model that includes the interaction term:

```
interactionModel<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(interactionModel) # results are hidden due to space constraint
```

This model is better than the previous. Its residual standard error is 2.084 on 27 degrees of freedom, while the adjusted R-squared value is 0.8804, i.e., this model can explain about 88% of the variance of the MPG variable. All of the coefficients are also at 0.05 significant level.

Now, we fit a simple model for MPG feature to be predicted from transmission variable.

```
amModel<-lm(mpg ~ am, data=mtcars)
summary(amModel) # results are hidden due to space constraint
```

The results reveal that a car with automatic transmission has 17.147 MPG, whereas that of a car with manual transmission is about 7 MPG higher. The residual standard error of this model is 4.902 on 30 degrees of freedom while the adjusted R-squared value is 0.3385, i.e., the model can explain about 34% of the variance of the MPG variable. Such a low value indicates that other variables need to be added into the model.

Final model

We choose the final model as follows.

```
anova(amModel, stepModel, fullModel, interactionModel)
confint(interactionModel) # results are hidden due to space constraint
```

We choose the model considering the interaction between variables *wt* and *am*, specifically "mpg ~ wt + qsec + am + wt:am", as it has the highest adjusted R-squared value.

```
summary(interactionModel)$coef # results are hidden due to space constraint
```

As shown by the result, cars with manual transmission add about [14.079 + (-4.141)*weight] more miles per gallon on average than cars with automatic transmission when "wt" (weight of the ca) and "qsec" (1/4 mile time) remain constant.

Residual Analysis and Diagnostics

The residual plots are shown in the **Appendix** section. The following conclusions can be made. Firstly, the *Residuals vs. Fitted* plot indicates no consistent pattern, which supports the accuracy of the independence assumption. Secondly, the *Normal Q-Q* plot shows that the residuals are normally distributed as the points lie closely to the line. Thirdly, the *Scale-Location* plot affirms the constant variance assumption since the points are randomly distributed. Finally, the *Residuals vs. Leverage* plot confirms that there is no outlier because all values fall within the 0.5 band.

In addtion, we measure the impact of an observation on the estimate of a regression coefficient:

```
sum((abs(dfbetas(interactionModel)))>1)
```

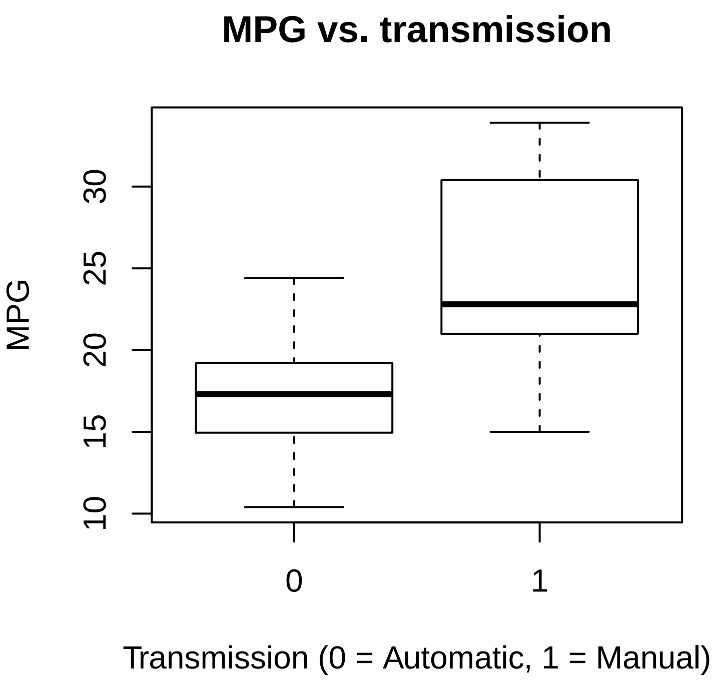
```
## [1] 0
```

In summary, all basic assumptions of linear regression are met and the above analyses answer the questions well.

Appendix

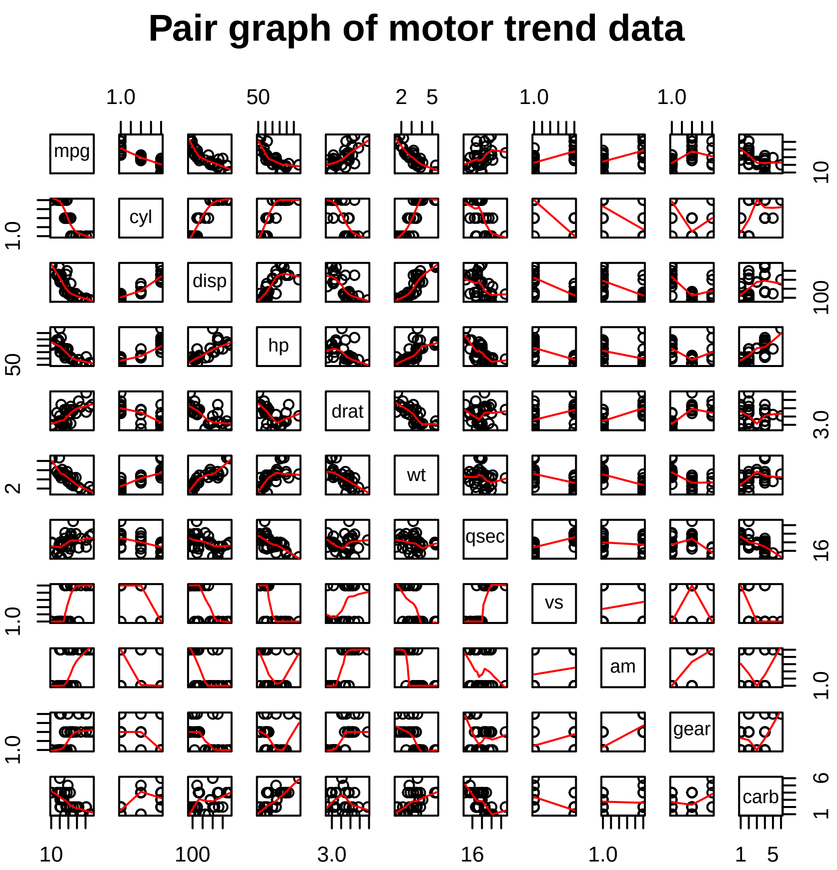
Box plot of MPG vs. transmission

```
boxplot(mpg ~ am, xlab="Transmission (0 = Automatic, 1 = Manual)", ylab="MPG",
        main="MPG vs. transmission")
```



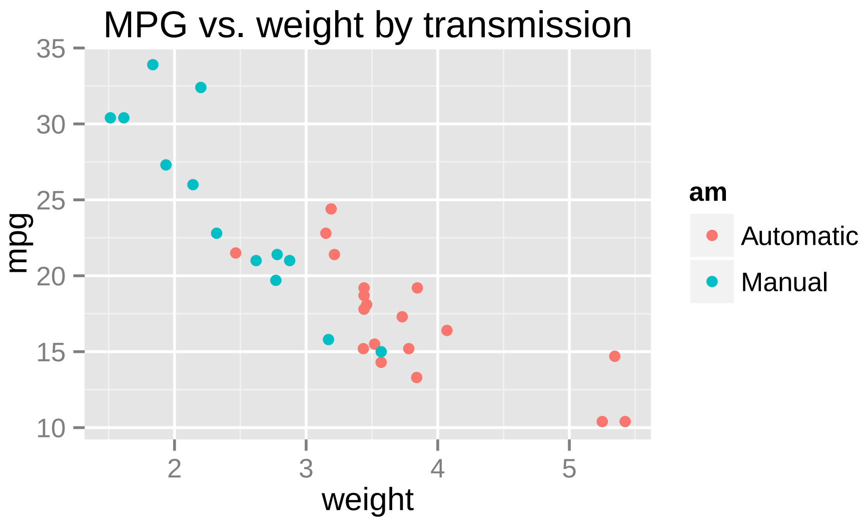
Pair graph of motor trend data

```
pairs(mtcars, panel=panel.smooth, main="Pair graph of motor trend data")
```



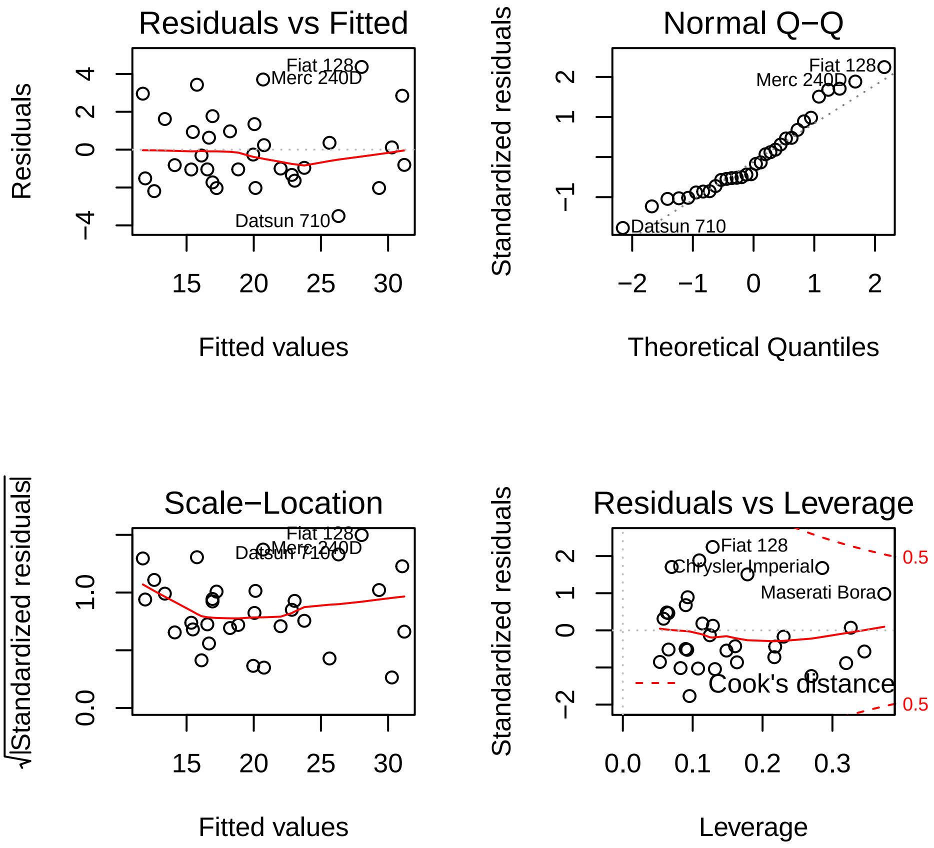
Scatter plot of MPG vs. weight by transmission

```
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + geom_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("weight") + ggtitle("MPG vs. weight by transmission")
```



Residual plots

```r
par(mfrow = c(2, 2))
plot(interactionModel)
```


Residuals vs Fitted
Residuals
−4
0
2
4
Fiat 128
Merc 240D
Datsun 710
15
20
25
30
Fitted values
Normal Q−Q
Standardized residuals
−1
1
2
Fiat 128
Merc 240D
Datsun 710
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
1.0
Fiat 128
Merc 240D
Datsun 710
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
−2
0
1
2
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5
0.0
0.1
0.2
0.3
Leverage