Title: Regression Models Course Project - Motor Trend

This report analyzes relationship between a set of variables and miles per gallaon using “mtcars” dataset, to see if a. “Is an automatic or manual transmission better for MPG” b. “Quantify the MPG difference between automatic and manual transmissions”

Exploratory Data Analysis

First, Load the dataset `mtcars` and change some variables from `numeric` class to `factor` class.

```
library(ggplot2)
data(mtcars)
mtcars[1:3, ] # Sample Data
```

```
##                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
```

```
dim(mtcars)
```

```
## [1] 32 11
```

```
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)
```

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

Next we will make 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
```

```
## [1] 0.001374
```

```
result$estimate
```

```
## mean in group 0 mean in group 1 
##           17.15           24.39
```

We got the p-value is 0.00137, we reject the null hypothesis. Applying Regression Analysis and fit the model

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

This resulted in Residual std. eror of 2.833 on 15 deg of freedon, and ajusted R squared is 0.779 - explaining 78% of variance of MPG available. But the coefficinets are significatnt at 0.05 level. Lets calculated some significant variables:

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

This model is “mpg ~ wt + qsec + am”. It has the Residual standard error as 2.459 on 28 degrees of freedom. And the Adjusted R-squared value is 0.8336, which means that the model can explain about 83% 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.

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

This model has the Residual standard error as 2.084 on 27 degrees of freedom. And the Adjusted R-squared value is 0.8804, which means that the model can explain about 89% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level. This is a pretty good one.

Next, we fit the simple model with MPG as the outcome variable and Transmission as the predictor variable.

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

It 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. And 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 need to add other variables to the model.

Finally, we select the final model.

```
anova(amModel, stepModel, fullModel, amIntWtModel)
confint(amIntWtModel) # results hidden
```

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

```
summary(amIntWtModel)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.1108925
## wt            -2.937      0.666  -4.409 0.0001489
## qsec           1.017      0.252   4.035 0.0004030
## am1           14.079      3.435   4.099 0.0003409
## wt:am1        -4.141      1.197  -3.460 0.0018086
```

Thus, the result shows that when “wt” (weight lb/1000) and “qsec” (¼ 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 ¼ 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.

As for the Dfbetas, the measure of how much an observation has effected the estimate of a regression coefficient, we get the following result:

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

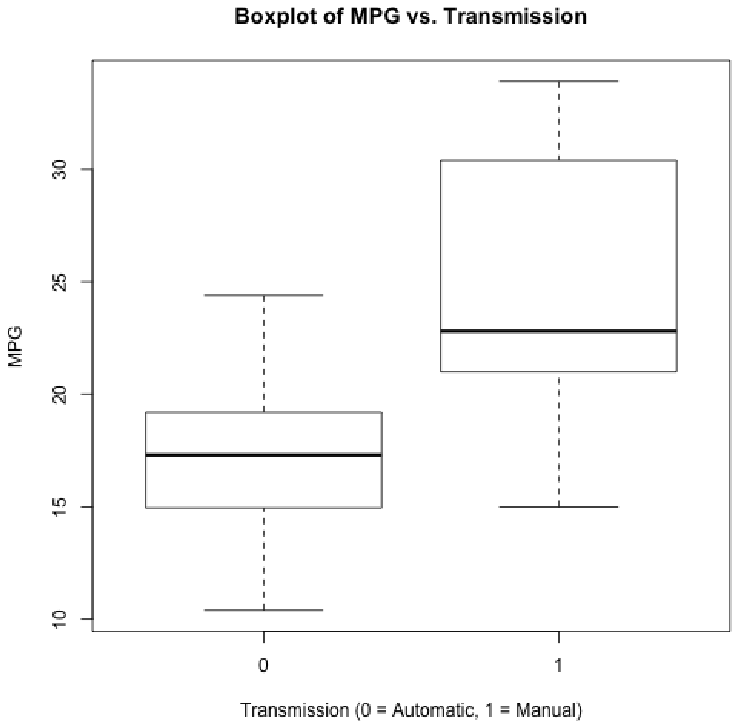
```
## [1] 0
```

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

Appendix: Figures

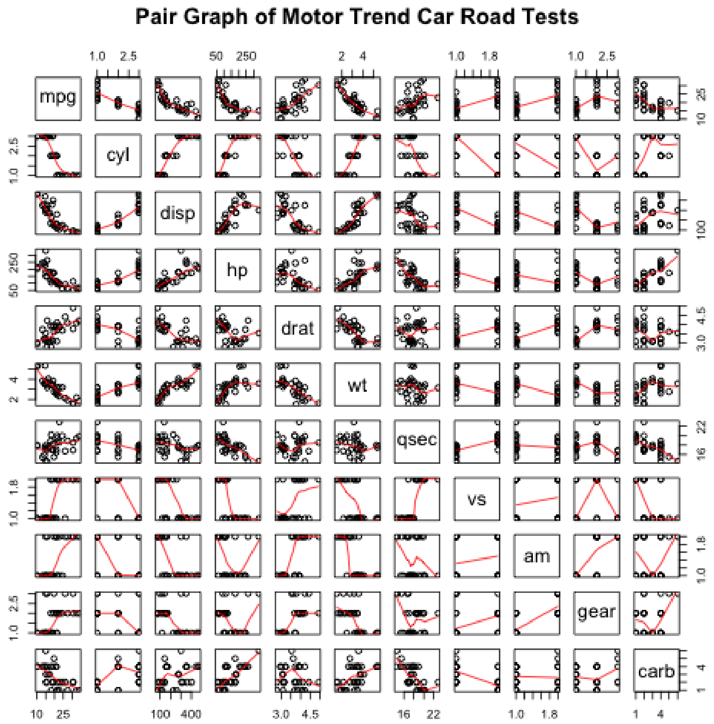
1. Boxplot of MPG vs. Transmission

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



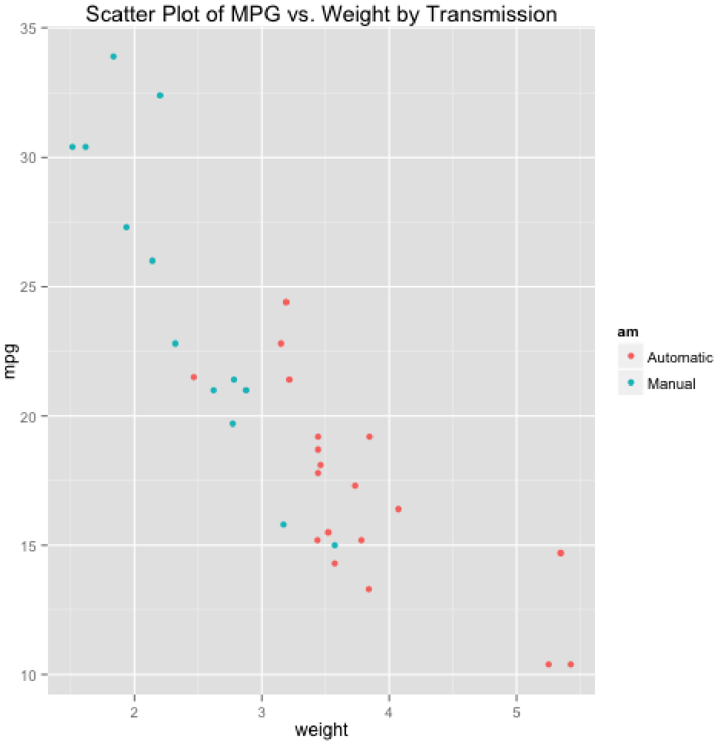
2. Pair Graph of Motor Trend Car Road Tests

```
pairs(mtcars, panel=panel.smooth, main="Pair Graph of Motor Trend Car Road Tests")
```



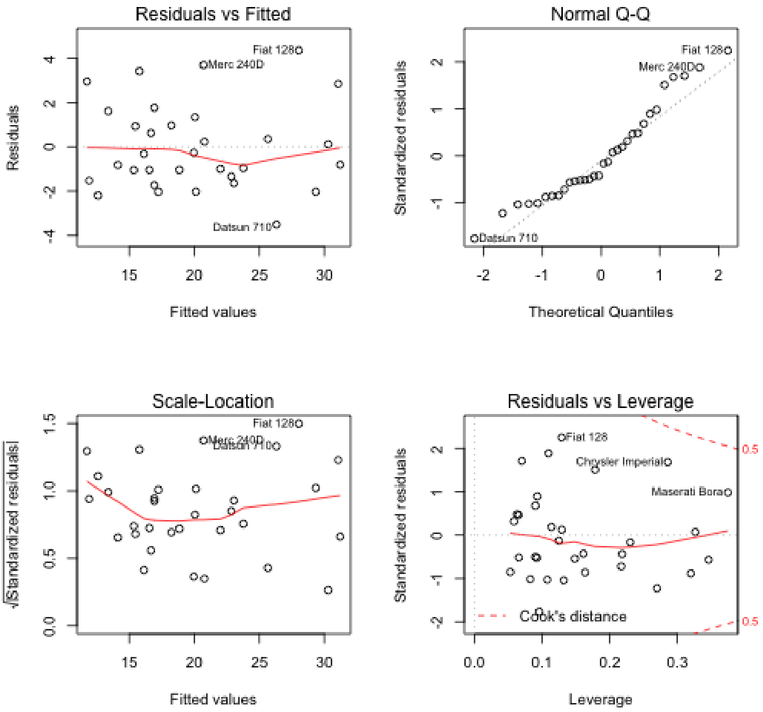
3. Scatter Plot of MPG vs. Weight by Transmission

```
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("Scatter Plot of MPG vs. Weight by Transmission")
```



4. Residual Plots

```
par(mfrow = c(2, 2))
plot(amIntWtModel)
```


Residuals vs Fitted
Residuals
Fitted values
Fiat 128
Merc 240D
Datsun 710
4
2
0
-2
-4
15
20
25
30
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Merc 240D
Datsun 710
2
1
0
-1
-2
-1
0
1
2
Scale-Location
√|Standardized residuals|
Fitted values
Fiat 128
Merc 240D
Datsun 710
1.5
1.0
0.5
0.0
15
20
25
30
Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5
2
1
0
-1
-2
0.0
0.1
0.2
0.3