Regression Models Course Project

Executive Summary

Motor Trend, a magazine about the automobile industry, is interested in the following two questions:

"Is an automatic or manual transmission better for MPG" "Quantify the MPG difference between automatic and manual transmissions"

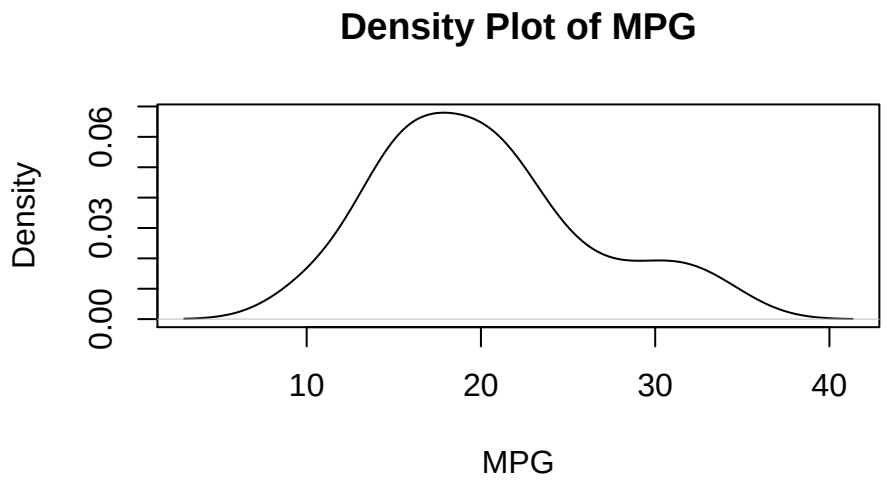
This report *used the "mtcars" data set,* checked the data for mpg from the source data set, *tested whether the transmission type causes a significant difference in mpg,* built multiple regression models, *selected one regression model with aid of ANOVA and coefficients,* run the selected models and went through the residuals, and *concluded the MPG difference between automatic and manual transmissions.

Data Processing

```
# Load the data
library(datasets);data(mtcars)
# Turn the variable for transmission type into a factor
mtcars$am <- as.factor(mtcars$am) #Appendix I for checking variables
levels(mtcars$am) <- c("Automatic", "Manual")
mtcars$cyl <- as.factor(mtcars$cyl)
```

Inference: Valid data

```
# View mpg
plot(density(mtcars$mpg), xlab = "MPG", main ="Density Plot of MPG")
```

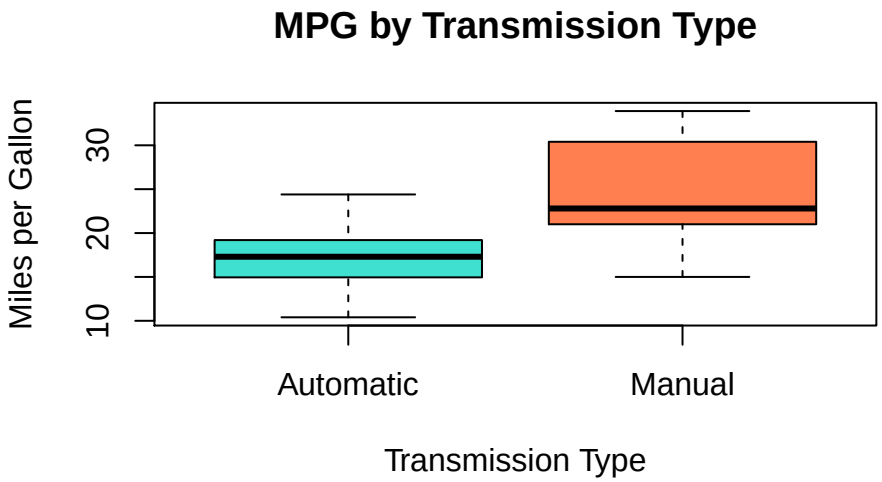


There is no obvious outliner in the plot. The data shows a similar spread in the direction of the line drawn. It is possible that mtcars$mpg is normally disributed.

Exploratory Data Analysis

Let's check mtcars$mpg against the independent variable, the transmision type.

```
# View the mpg breakdown by transmission type
boxplot(mpg ~ am, data = mtcars, col = c("turquoise", "coral"),
         xlab = "Transmission Type", ylab = "Miles per Gallon",
         main = "MPG by Transmission Type")
```



Although the MPG for automatic tranmission is evenly distributed, the box plot suggests that the manual tranmission type leverages more miles per gallon.

Inference: Hypothesis Test

From the box plot, the manual transmission type has a higher mean. Let's check whether this is a significant difference.

```
t.test(mtcars[mtcars$am == "Automatic",]$mpg, mtcars[mtcars$am == "Manual",]$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars[mtcars$am == "Automatic", ]$mpg and mtcars[mtcars$am == "Manual", ]$mpg
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
## mean of x mean of y
##     17.15     24.39
```

The p-value of 0.001374 suggests *there is significant difference in the mean of the automatic transmission type and the manual transmission type.* In other words, the choice between automatic transmission and manual transmission affects mpg. Either one of the two types does better.

Main: Regression Analysis

Build the models and refine with checking coefficients

Let's select predictors.

```
# analyse the variables
all_var <- aov(mpg ~ ., data = mtcars);summary(all_var)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## cyl          2    825     412   61.86 2.7e-09 ***
## disp         1     58      58    8.65  0.0081 **
## hp           1     19      19    2.78  0.1113
## drat         1     12      12    1.79  0.1963
## wt           1     56      56    8.37  0.0090 **
## qsec         1      2       2    0.23  0.6377
## vs           1      0       0    0.05  0.8336
## am           1     17      17    2.49  0.1306
## gear         1      4       4    0.56  0.4618
## carb         1      2       2    0.29  0.5948
## Residuals   20    133       7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The variables with high relevance are cyl, wt, disp, and drat (in descending rank) because they have a p-value smaller than 0.05. After a series of ANOVA analysis(Appendix II),cyl and wt were chosen as the predictors in addition to am.

```
# final linear regression model
fit <- lm(mpg ~ am + cyl + wt, data = mtcars);summary(fit)
```

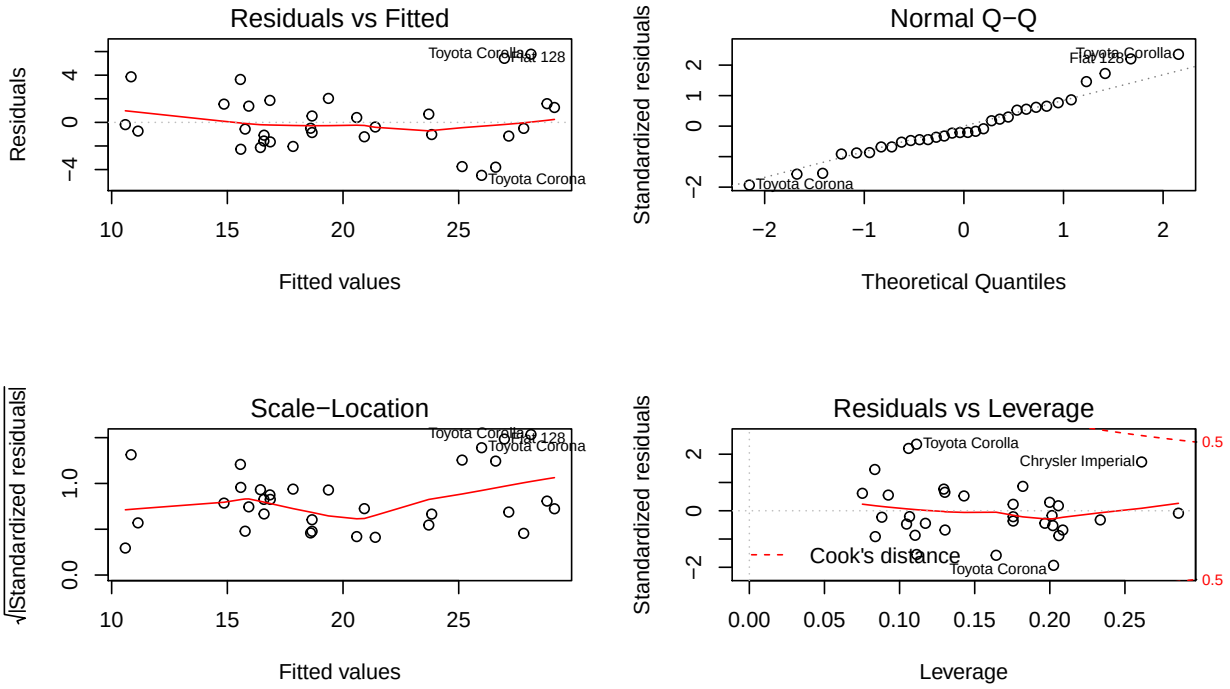
```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.490 -1.312 -0.504  1.416  5.776
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.754      2.813   12.00  2.5e-12 ***
## amManual       0.150      1.300    0.12   0.9089
## cyl6          -4.257      1.411   -3.02   0.0055 **
## cyl8          -6.079      1.684   -3.61   0.0012 **
## wt            -3.150      0.908   -3.47   0.0018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.6 on 27 degrees of freedom
## Multiple R-squared:  0.838,  Adjusted R-squared:  0.813
## F-statistic: 34.8 on 4 and 27 DF,  p-value: 2.73e-10
```

The R-square value suggested this model explained over 80% of the variance. The p-value of 2.73e-10—showed the confidence level was improved.

Residual Plot and Diagnostics

Let's check the residuals.

```
par(mfrow = c(2,2));plot(fit)
```



- *Residuals vs. Fitted plot*: No obvious band interval or trend observed, matching the independence condition.
- *Normal Q-Q plot*: There is no apparent outliners (points distant from the regression line). Some points fall below the line and some fall above the line. However the sample size is too small to drop some of the data. The residuals are likely to be normally distributed.
- *Scale-Location plot*: Even interval band above and below the regression line.
- *Residual vs Leverage plot*: It exhibited convergence trend despite a outliner at the top right corner.

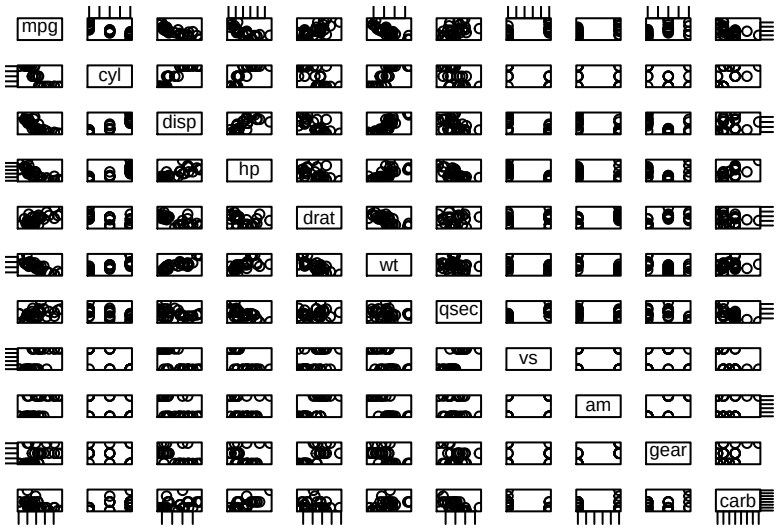
Conclusion

- Cars with Manual transmission get more miles per gallon mpg compared to cars with Automatic transmission. (0.150 adjusted by cyl, and wt).
- When the number of cylinder changes from 4 to 6, mpg will descrease by a factor of 4.257(adjusted by cyl, and wt).
- When the number of cylinder changes from 4 to 8, mpg will descrease by a factor of 6.079(adjusted by cyl, and wt).
- mpg will decrease by 3.150 (adjusted by cyl, and wt) for every 1000 lb increase in wt.

To improve the model, more data is needed to justify the normal distribution hypothesis and residual conditions.

Appendix I: Check categorical variables

```
pairs(mtcars, oma = c(0,0,0,0))
```



Appendix II: Select the model

```
simple_model <- lm(mpg~am, data = mtcars) # Fit mpg and am only
full_model <- lm(mpg ~ ., data = mtcars) # Fit all variables
multi_model1 <- lm(mpg ~ am + cyl, data = mtcars) # Fit 1 more significant variable
multi_model2 <- lm(mpg ~ am + cyl + wt, data = mtcars) # Fit 2 more
multi_model3 <- lm(mpg ~ am + cyl + wt + disp, data = mtcars) # Fit 3 more
multi_model4 <- lm(mpg ~ am + cyl + wt + disp + drat, data = mtcars) # Fit 4 more
anova(simple_model,multi_model1,multi_model2,multi_model3,multi_model4,full_model) # Compare
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + wt
## Model 4: mpg ~ am + cyl + wt + disp
## Model 5: mpg ~ am + cyl + wt + disp + drat
## Model 6: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df RSS Df Sum of Sq     F  Pr(>F)
## 1     30 721
## 2     28 264  2       456 34.23 3.5e-07 ***
## 3     27 183  1        82 12.23  0.0023 **
## 4     26 183  1         0  0.01  0.9042
## 5     25 183  1         0  0.03  0.8660
## 6     20 133  5        49  1.48  0.2402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The change from simple-model to multi-model2 gave the smallest p-value, suggesting the greatest improvement in confidence level. Therefore the final regression model considered am, cyl, and wt.