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)
```

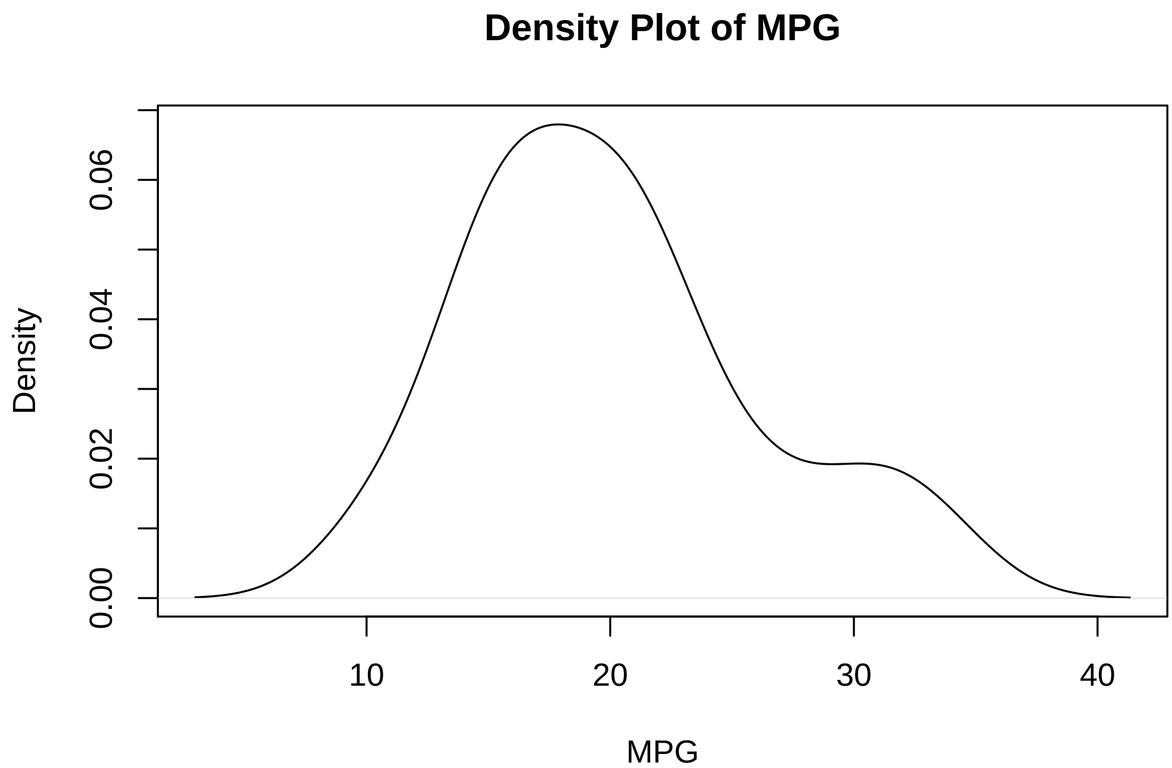
With help(mtcars) (See Appendix 1) and the scatter plot of mtcars (Appendix 2), it's shown that mtcars$am is the category variable specifying whether the transmission is automatic or manual.

```
# Turn the variable for transmission type into a factor
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
```

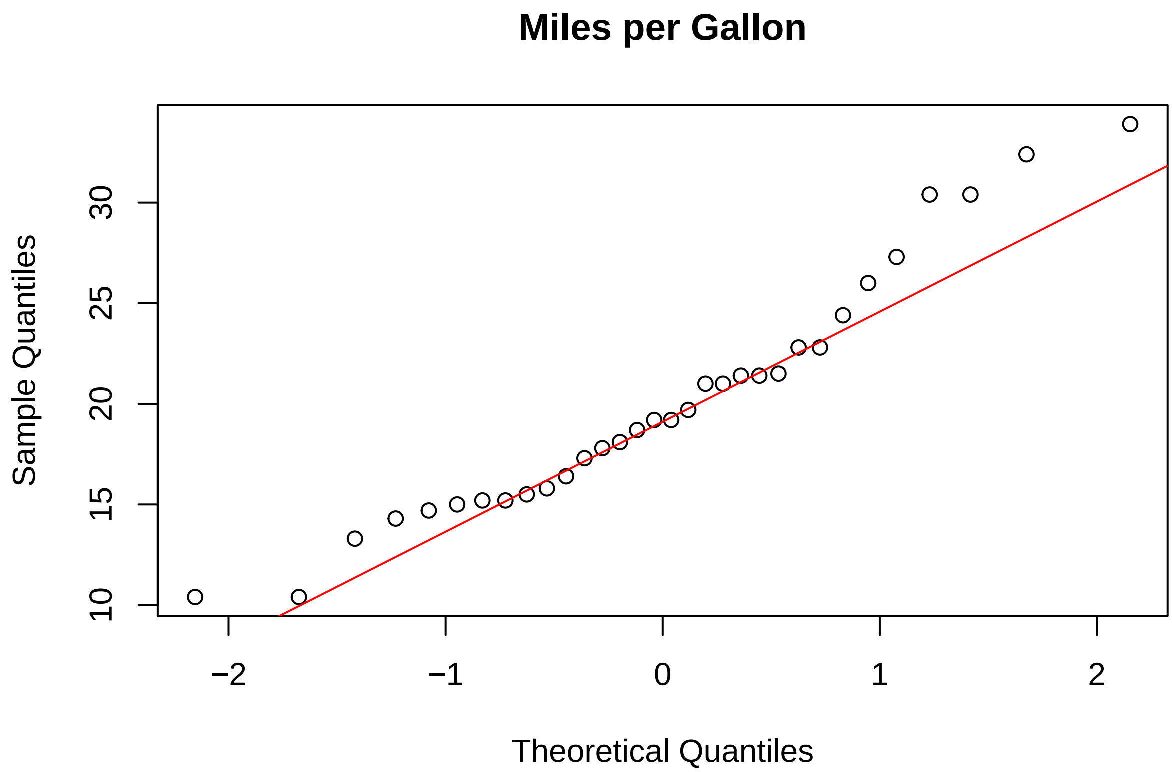
Exploratory Data Analysis

Let's have a look at the dependent variable before the linear regression analysis. ### Inference: Valid data

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



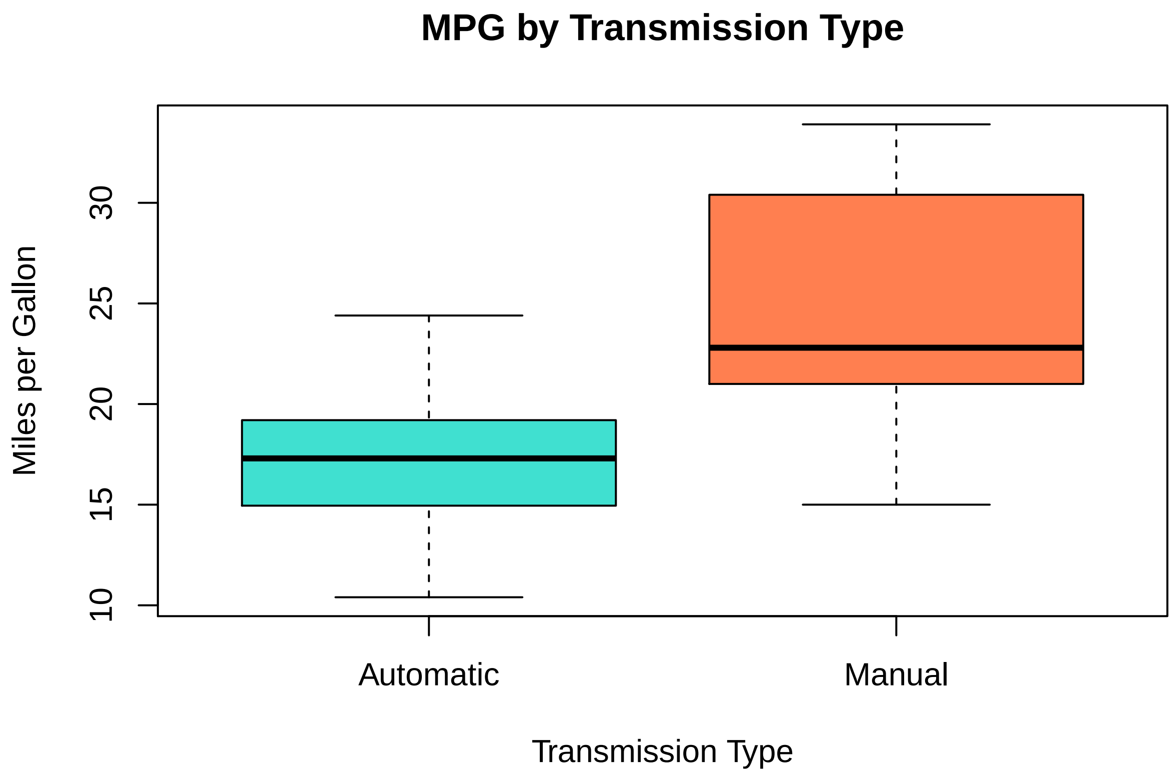

```r
qqnorm(mtcars$mpg, main="Miles per Gallon")
qqline(mtcars$mpg, col=2)
```



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.

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.

Regression Analysis

Build the models and refine with checking coefficients

Let's get the simple linear regression model considering mpd and the transmission type only. This can be used as reference for model refinement.

```
# Fit mpg and am only
simple_model <- lm(mpg~am, data = mtcars)
summary(simple_model)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.15       1.12   15.25  1.1e-15 ***
## amManual        7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

Let's consider all the variables.

```
# Fit all variables
full_model <- lm(mpg ~ ., data = mtcars)
summary(full_model)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3.45  -1.60  -0.12   1.22   4.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.3034    18.7179    0.66    0.518
## cyl          -0.1114     1.0450   -0.11    0.916
## disp          0.0133     0.0179    0.75    0.463
## hp           -0.0215     0.0218   -0.99    0.335
## drat          0.7871     1.6354    0.48    0.635
## wt           -3.7153     1.8944   -1.96    0.063 .
## qsec          0.8210     0.7308    1.12    0.274
```

```
## vs            0.3178     2.1045    0.15    0.881
## amManual      2.5202     2.0567    1.23    0.234
## gear          0.6554     1.4933    0.44    0.665
## carb         -0.1994     0.8288   -0.24    0.812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.807
## F-statistic: 13.9 on 10 and 21 DF,  p-value: 3.79e-07
```

```
# Compare the 2 models
anova(simple_model,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: 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     21 147  9       573 9.07 1.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the analysis of variance suggested the 2 models had significant difference. Presumably considering more variables can improve the confidence level of the model, which was reflected on a smaller p-value.

It was unsure that considering some of the variables with higher relevance will further improve the model. To select such variablces as predictors, an analysis of variance model is performed.

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

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## cyl          1    818     818  116.42  5e-10 ***
## disp         1     38      38    5.35 0.0309 *
## hp           1      9       9    1.33 0.2610
## drat         1     16      16    2.34 0.1406
## wt           1     77      77   11.03 0.0032 **
## qsec         1      4       4    0.56 0.4617
## vs           1      0       0    0.02 0.8932
## am           1     14      14    2.06 0.1659
## gear         1      1       1    0.14 0.7137
## carb         1      0       0    0.06 0.8122
## Residuals   21    147       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,cyl and wt were chosen as the predictors in addition to am.

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

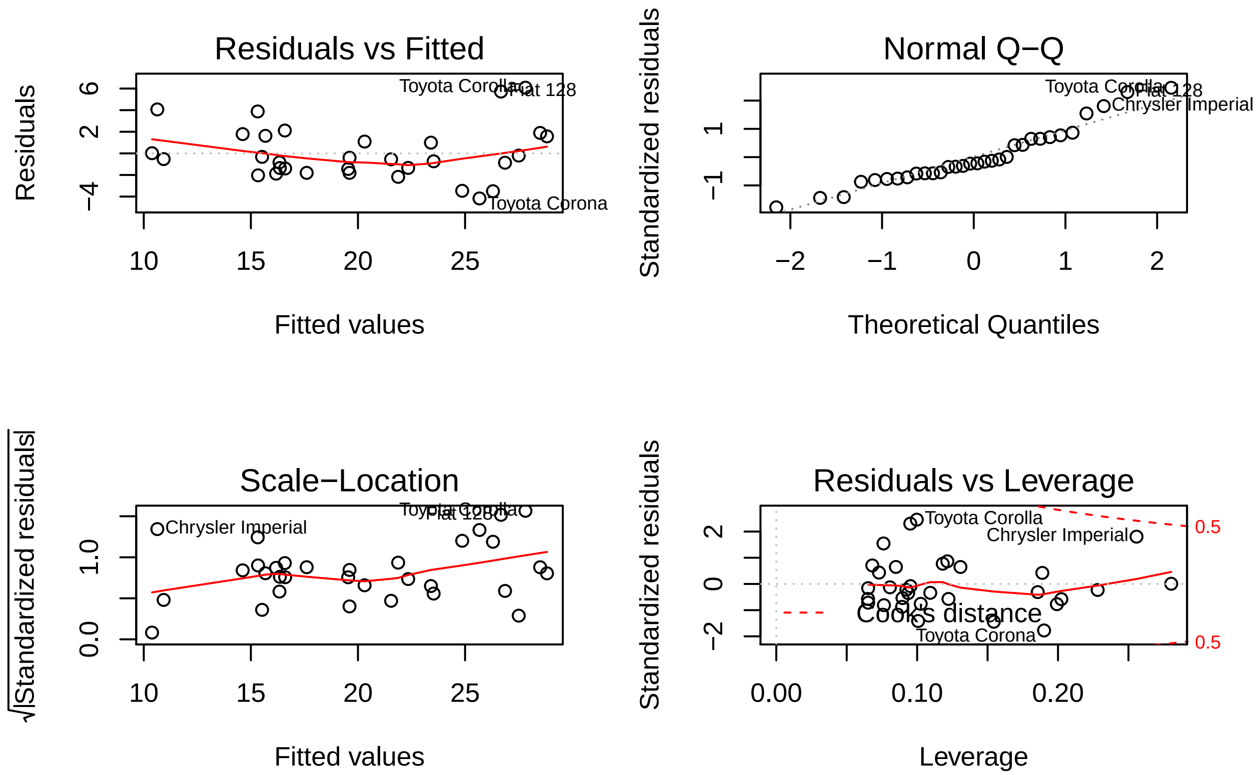
```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.173 -1.534 -0.539  1.586  6.081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.418      2.641   14.92  7.4e-15 ***
## amManual       0.176      1.304    0.14   0.8933
## cyl           -1.510      0.422   -3.58   0.0013 **
## wt            -3.125      0.911   -3.43   0.0019 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.61 on 28 degrees of freedom
## Multiple R-squared:  0.83,   Adjusted R-squared:  0.812
## F-statistic: 45.7 on 3 and 28 DF,  p-value: 6.51e-11
```

The R-square value suggested this model explained over 80% of the variance. The p-value of 6.51e-11 showed the confidence level was improved.

Residual Plot and Diagnostics

Let's check the residuals.

```r
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

```r
summary(full_model)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -3.45  -1.60  -0.12   1.22   4.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.3034    18.7179    0.66    0.518
## cyl          -0.1114     1.0450   -0.11    0.916
## disp          0.0133     0.0179    0.75    0.463
```

```
## hp           -0.0215     0.0218   -0.99    0.335  
## drat          0.7871     1.6354    0.48    0.635  
## wt           -3.7153     1.8944   -1.96    0.063 .
## qsec          0.8210     0.7308    1.12    0.274  
## vs            0.3178     2.1045    0.15    0.881  
## amManual      2.5202     2.0567    1.23    0.234  
## gear          0.6554     1.4933    0.44    0.665  
## carb         -0.1994     0.8288   -0.24    0.812  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.807
## F-statistic: 13.9 on 10 and 21 DF,  p-value: 3.79e-07
```

Cars with Manual transmission get more miles per gallon mpg compared to cars with Automatic transmission. (0.17 adjusted by cyl, and wt). mpg will decrease by 3.12 (adjusted by cyl, and wt) for every 1000 lb increase in wt. mpg will descrease by a factor of 1.51(adjusted by cyl, and wt).

To improve the model, more data is needed to justify the normal distribution hypothesis and residual conditions. Further study on the incremental changes of the independent variables can be performed.

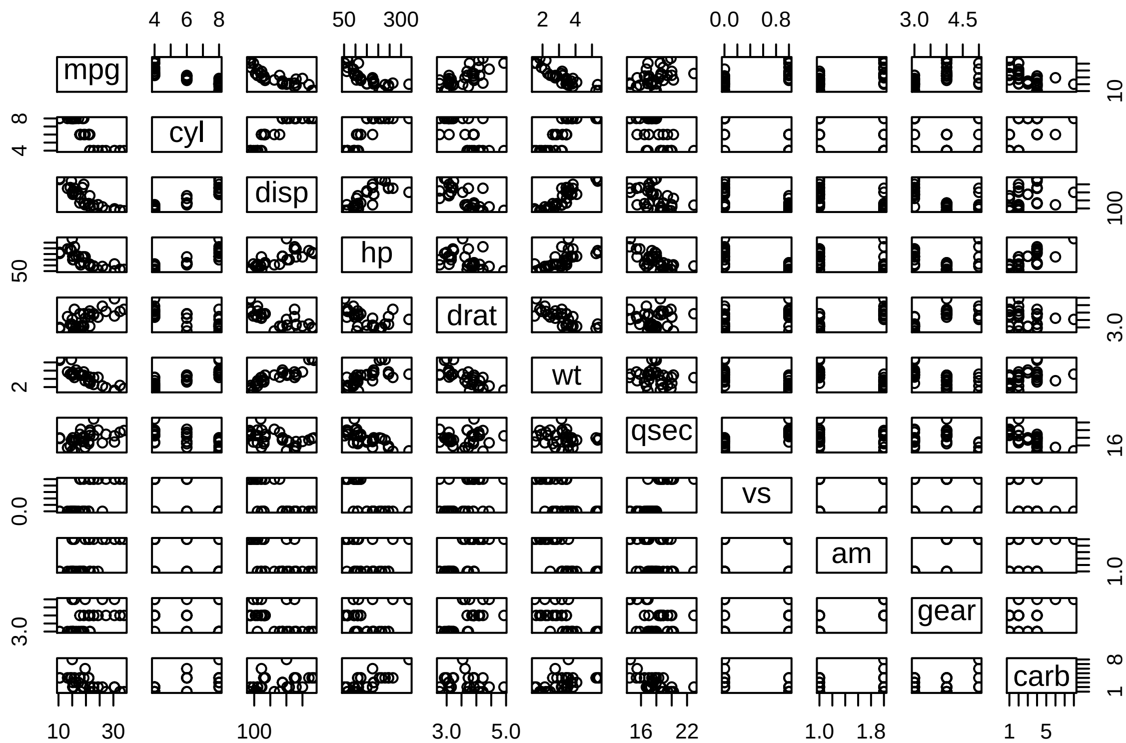
Appendix

1. Fields in mtcars

```
*1.  mpg   Miles/(US) gallon
*2.  cyl   Number of cylinders
*3.  disp  Displacement (cu.in.)
*4.  hp    Gross horsepower
*5.  drat  Rear axle ratio
*6.  wt    Weight (lb/1000)
*7.  qsec  1/4 mile time
*8.  vs    V/S
*9.  am    Transmission (0 = automatic, 1 = manual)
*10.  gear  Number of forward gears
*11.  carb  Number of carburetors
```

2. Scatterplot of mtcars

```r
pairs(mtcars)
```



3. Selecting the predictors

```
# Fit the 1 most significant variables in additional to am
multi_model1 <- lm(mpg ~ am + cyl, data = mtcars)
# Fit the 2 most significant variables in additional to am
multi_model2 <- lm(mpg ~ am + cyl + wt, data = mtcars)
# Fit the 3 most significant variables in additional to am
multi_model3 <- lm(mpg ~ am + cyl + wt + disp, data = mtcars)
# Fit the 4 most significant variables in additional to am
multi_model4 <- lm(mpg ~ am + cyl + wt + disp + drat, data = mtcars)
# Check the model summary
summary(multi_model1)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.686 -1.717 -0.266  1.884  6.814
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.522      2.603   13.26   7.7e-14 ***
```

```
## amManual       2.567      1.291    1.99    0.056 .
## cyl           -2.501      0.361   -6.93  1.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.06 on 29 degrees of freedom
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.742
## F-statistic: 45.7 on 2 and 29 DF,  p-value: 1.09e-09
```

```r
summary(multi_model2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.173 -1.534 -0.539  1.586  6.081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.418      2.641   14.92  7.4e-15 ***
## amManual       0.176      1.304    0.14   0.8933
## cyl           -1.510      0.422   -3.58   0.0013 **
## wt            -3.125      0.911   -3.43   0.0019 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.61 on 28 degrees of freedom
## Multiple R-squared:  0.83,   Adjusted R-squared:  0.812
## F-statistic: 45.7 on 3 and 28 DF,  p-value: 6.51e-11
```

```r
summary(multi_model3)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt + disp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.058
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.8983     3.6015   11.36  8.7e-12 ***
## amManual      0.1291     1.3215    0.10   0.9229
## cyl          -1.7842     0.6182   -2.89   0.0076 **
## wt           -3.5834     1.1865   -3.02   0.0055 **
## disp          0.0074     0.0121    0.61   0.5451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.64 on 27 degrees of freedom
```

```
## Multiple R-squared:  0.833,  Adjusted R-squared:  0.808
## F-statistic: 33.6 on 4 and 27 DF,  p-value: 4.04e-10
```

```r
summary(multi_model4)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt + disp + drat, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.383 -0.473  1.323  6.060
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.29638    7.53839    5.48  9.6e-06 ***
## amManual     0.17298    1.53004    0.11   0.9109
## cyl         -1.79400    0.65054   -2.76   0.0105 *
## wt          -3.58704    1.21050   -2.96   0.0064 **
## disp         0.00737    0.01232    0.60   0.5546
## drat        -0.09363    1.54878   -0.06   0.9523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.69 on 26 degrees of freedom
## Multiple R-squared:  0.833,  Adjusted R-squared:  0.801
## F-statistic: 25.9 on 5 and 26 DF,  p-value: 2.53e-09
```

```r
# Compare the models to the full model
anova(multi_model1,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df RSS Df Sum of Sq   F Pr(>F)
## 1     29 271
## 2     21 148  8       124 2.2   0.07 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(multi_model2,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + wt
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     28 191
## 2     21 148  7      43.6 0.89   0.53
```

```r
anova(multi_model3,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + wt + disp
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     27 188
## 2     21 148  6      40.9 0.97   0.47
```

```r
anova(multi_model4,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + wt + disp + drat
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     26 188
## 2     21 148  5      40.9 1.16   0.36
```

```r
# Compare the models to the simple model
anova(multi_model1,simple_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl
## Model 2: mpg ~ am
##   Res.Df RSS Df Sum of Sq  F  Pr(>F)
## 1     29 271
## 2     30 721 -1      -450 48 1.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(multi_model2,simple_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + wt
## Model 2: mpg ~ am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     28 191
## 2     30 721 -2      -530 38.8 8.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(multi_model3,simple_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + wt + disp
```

```
## Model 2: mpg ~ am
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     27 188
## 2     30 721 -3      -532 25.4  5e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(multi_model4,simple_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + wt + disp + drat
## Model 2: mpg ~ am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     26 188
## 2     30 721 -4      -532 18.4 2.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(full_model,simple_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     21 147
## 2     30 721 -9      -573 9.07 1.8e-05 ***
## ---
## 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.