Regression_Analysis

Karthik Kalimuthu

August 4, 2016

## Executive Summary
In this report, we will analyze `mtcars` data set and explore the relationship between a set of variables and miles per gallon (MPG). The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). We use regression models and exploratory data analyses to mainly explore how **automatic** (am = 0) and **manual** (am = 1) transmissions features affect the **MPG** feature. The t-test shows that the performance difference between cars with automatic and manual transmission. And it is about 7 MPG more for cars with manual transmission than those with automatic transmission. Then, we fit several linear regression models and select the one with highest Adjusted R-squared value. So, given that weight and 1/4 mile time are held constant, manual transmitted cars are 14.079 + (-4.141)*weight more MPG (miles per gallon) on average better than automatic transmitted cars. Thus, cars that are lighter in weight with a manual transmission and cars that are heavier in weight with an automatic transmission will have higher MPG values.

Exploratory Data Analysis

First, we load the data set `mtcars` and change some variables from `numeric` class to `factor` class.

```r
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
```

```r
dim(mtcars)
```

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

```r
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
```

Then, we do some basic exploratory data analyses. Please refer to the **Appendix: Figures** section for the plots. According to the box plot, we see that manual transmission yields higher values of MPG in general. And as for the pair graph, we can see some higher correlations between variables like "wt", "disp", "cyl" and "hp".

Inference

At this step, we make the null hypothesis as the MPG of the automatic and manual transmissions are from the same population (assuming the MPG has a normal distribution). We use the two sample T-test to show it.

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

```
## [1] 0.001373638
```

```
result$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Since the p-value is 0.00137, we reject our null hypothesis. So, the automatic and manual transmissions are from different populations. And the mean for MPG of manual transmitted cars is about 7 more than that of automatic transmitted cars.

Regression Analysis

First, we fit the full model as the following.

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

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## am1          1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

This model has the Residual standard error as 2.833 on 15 degrees of freedom. And the Adjusted R-squared value is 0.779, which means that the model can explain about 78% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significant level.

Then, we use backward selection to select some statistically significant variables.

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

```
## Start:  AIC=101.32
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS     AIC
## - carb  5   13.5989 134.00  87.417
## - gear  2    3.9729 124.38  95.428
## - cyl   2   10.9314 131.33  97.170
## - am    1    1.1420 121.55  98.157
## - qsec  1    1.2413 121.64  98.183
## - drat  1    1.8208 122.22  98.335
## - vs    1    3.6299 124.03  98.806
## - disp  1    9.9672 130.37 100.400
## <none>              120.40 101.321
## - wt    1   25.5541 145.96 104.014
## - hp    1   25.6715 146.07 104.040
##
## Step:  AIC=87.42
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 81.662
## - cyl   2   12.5642 146.57 83.353
## - disp  1    0.9934 135.00 84.187
## - drat  1    1.1854 135.19 84.233
## - vs    1    3.6763 137.68 84.817
## - qsec  1    5.2634 139.26 85.184
## - am    1   11.9255 145.93 86.679
## <none>              134.00 87.417
## - wt    1   19.7963 153.80 88.360
## - hp    1   22.7935 156.79 88.978
##
## Step:  AIC=81.66
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - cyl   2   10.4247 149.45 77.045
## - drat  1    0.9672 139.99 78.418
## - disp  1    1.5483 140.57 78.551
## - vs    1    2.1829 141.21 78.695
## - qsec  1    3.6324 142.66 79.022
## <none>              139.02 81.662
## - am    1   16.5665 155.59 81.799
## - hp    1   18.1768 157.20 82.129
## - wt    1   31.1896 170.21 84.674
##
## Step:  AIC=77.04
## mpg ~ disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     0.645 150.09 73.717
## - drat  1     2.869 152.32 74.187
## - disp  1     9.111 158.56 75.473
## - qsec  1    12.573 162.02 76.164
## - hp    1    13.929 163.38 76.431
## <none>              149.45 77.045
## - am    1    20.457 169.91 77.684
## - wt    1    60.936 210.38 84.523
##
## Step:  AIC=73.72
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 70.956
## - disp  1     8.545 158.64 72.023
## - hp    1    13.285 163.38 72.965
## <none>              150.09 73.717
## - am    1    20.036 170.13 74.261
## - qsec  1    25.574 175.67 75.286
## - wt    1    67.572 217.66 82.146
##
## Step:  AIC=70.96
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 68.844
## - hp    1    12.572 166.01 70.011
## <none>              153.44 70.956
## - qsec  1    26.470 179.91 72.583
## - am    1    32.198 185.63 73.586
## - wt    1    69.043 222.48 79.380
##
## Step:  AIC=68.84
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 67.170
## <none>              160.07 68.844
## - qsec  1    20.225 180.29 69.186
## - am    1    25.993 186.06 70.193
## - wt    1    78.494 238.56 78.147
```

```
##
## Step:  AIC=67.17
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 67.170
## - am    1    26.178 195.46 68.306
## - qsec  1   109.034 278.32 79.614
## - wt    1   183.347 352.63 87.187
```

```
summary(stepModel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am1           2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

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. According to the scatter plot, it indicates that there appear to be an interaction term between “wt” variable and “am” variable, since automatic cars tend to weigh heavier than manual cars. Thus, we have the following model including the interaction term:

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

`##   ## Call: ## lm(formula = mpg ~ wt + qsec + am + wt:am, data = mtcars)   ##   ## Residuals: ##   Min   1Q Median   3Q   Max   ## -3.5076 -1.3801 -0.5588 1.0630 4.3684   ##   ## Coefficients: ##   Estimate Std. Error t value Pr(>|t|)   ## (Intercept)   9.723   5.899   1.648 0.110893   ## wt   -2.937 0.666 -4.409 0.000149 ***   ## qsec   1.017   0.252   4.035 0.000403 ***   ## am1   14.079   3.435   4.099 0.000341 ***   ## wt:am1   -4.141   1.197 -3.460 0.001809 **   ## ---   ## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1   ##   ## Residual standard error: 2.084 on 27 degrees of freedom   ## Multiple R-squared: 0.8959,   Adjusted R-squared: 0.8804   ## F-statistic: 58.06 on 4 and 27 DF, p-value: 7.168e-13` 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 88% 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)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

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)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 4: mpg ~ wt + qsec + am + wt:am
##   Res.Df    RSS  Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29   2    551.61 34.3604 2.509e-06 ***
## 3     15 120.40  13     48.88  0.4685    0.9114
## 4     27 117.28 -12      3.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(amIntWtModel)
```

```
##                  2.5 %    97.5 %
## (Intercept) -2.3807791 21.826884
## wt          -4.3031019 -1.569960
## qsec         0.4998811  1.534066
## am1          7.0308746 21.127981
## wt:am1      -6.5970316 -1.685721
```

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

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

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

Thus, the result shows that when “wt” (weight lb/1000) and “qsec” (1/4 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 1/4 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:

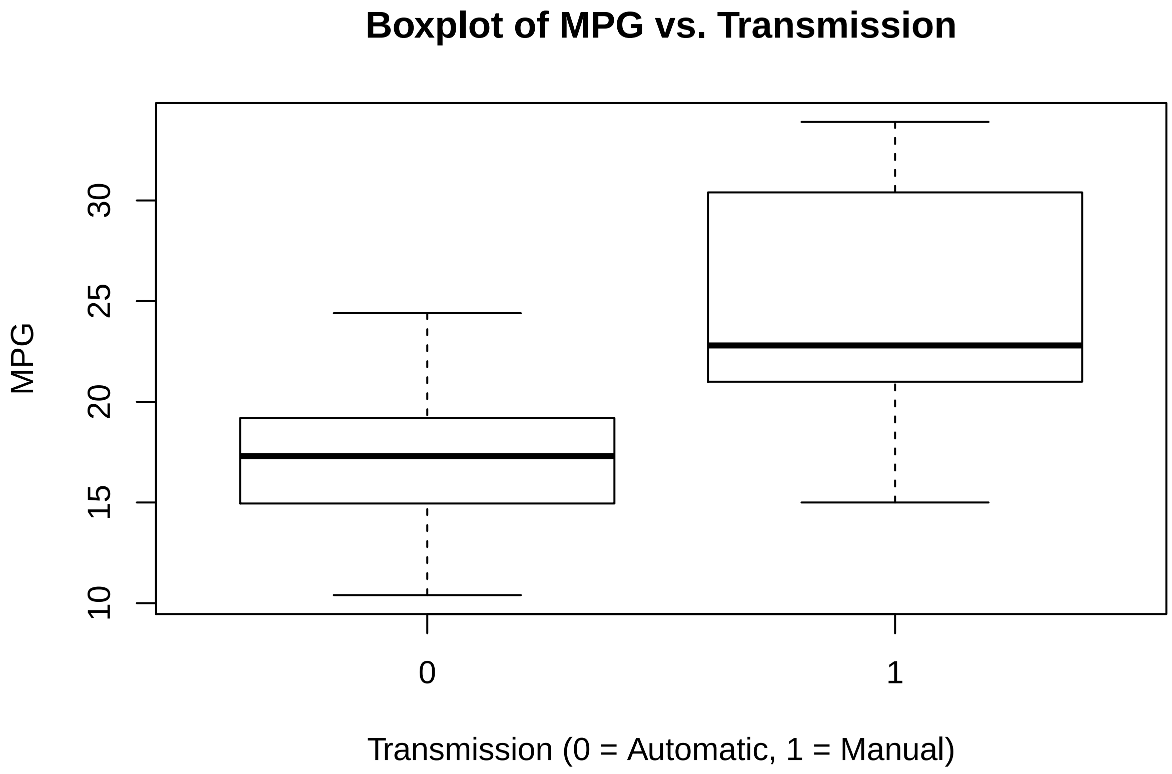
`r   sum((abs(dfbetas(amIntWtModel)))>1)`

`## [1] 0` Therefore, the above analyses meet all basic assumptions of linear regression and well answer the questions.

Appendix: Figures

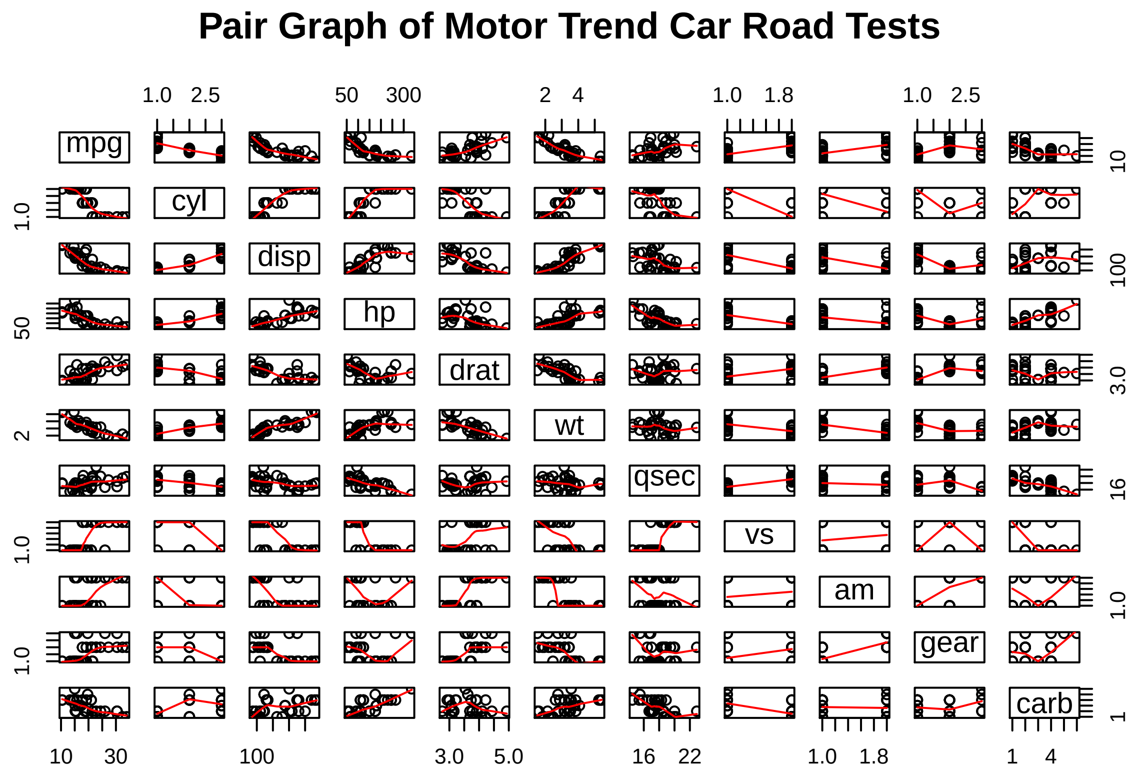
1. Boxplot of MPG vs. Transmission

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



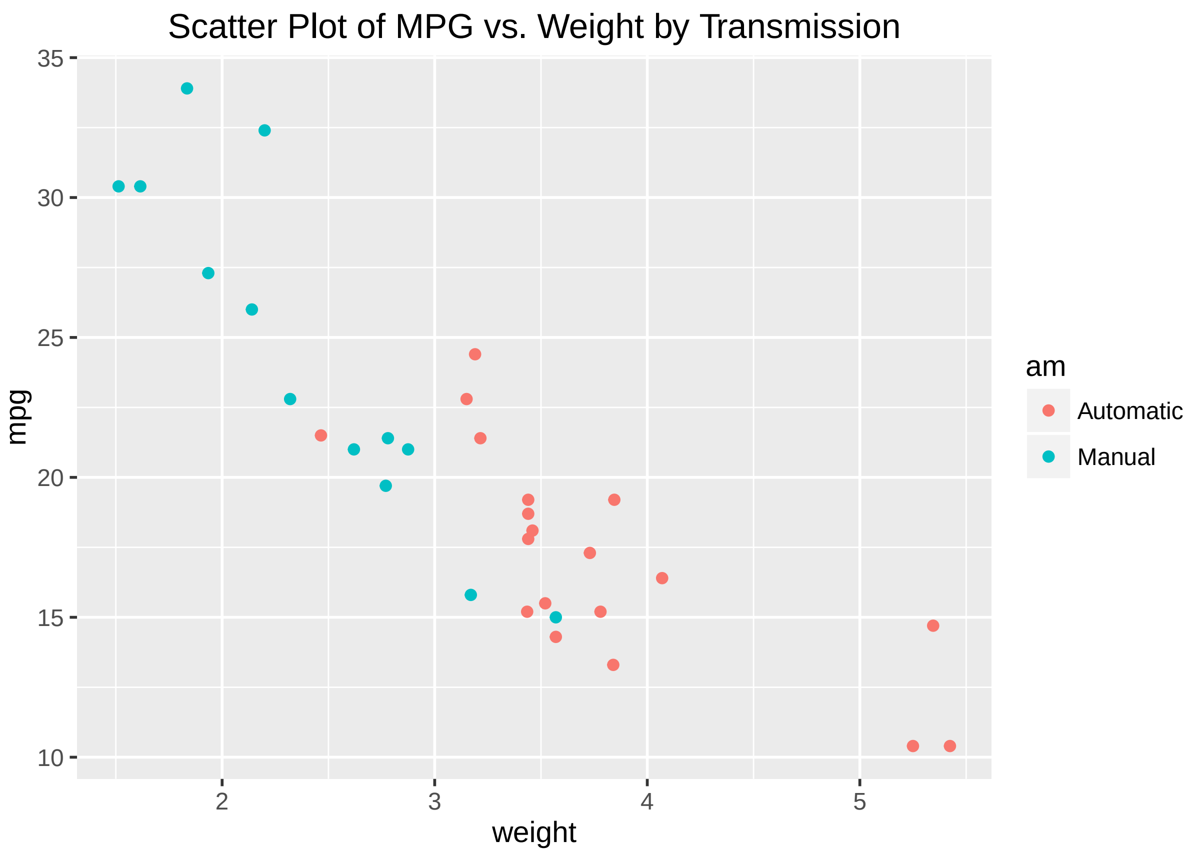
2. Pair Graph of Motor Trend Car Road Tests

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



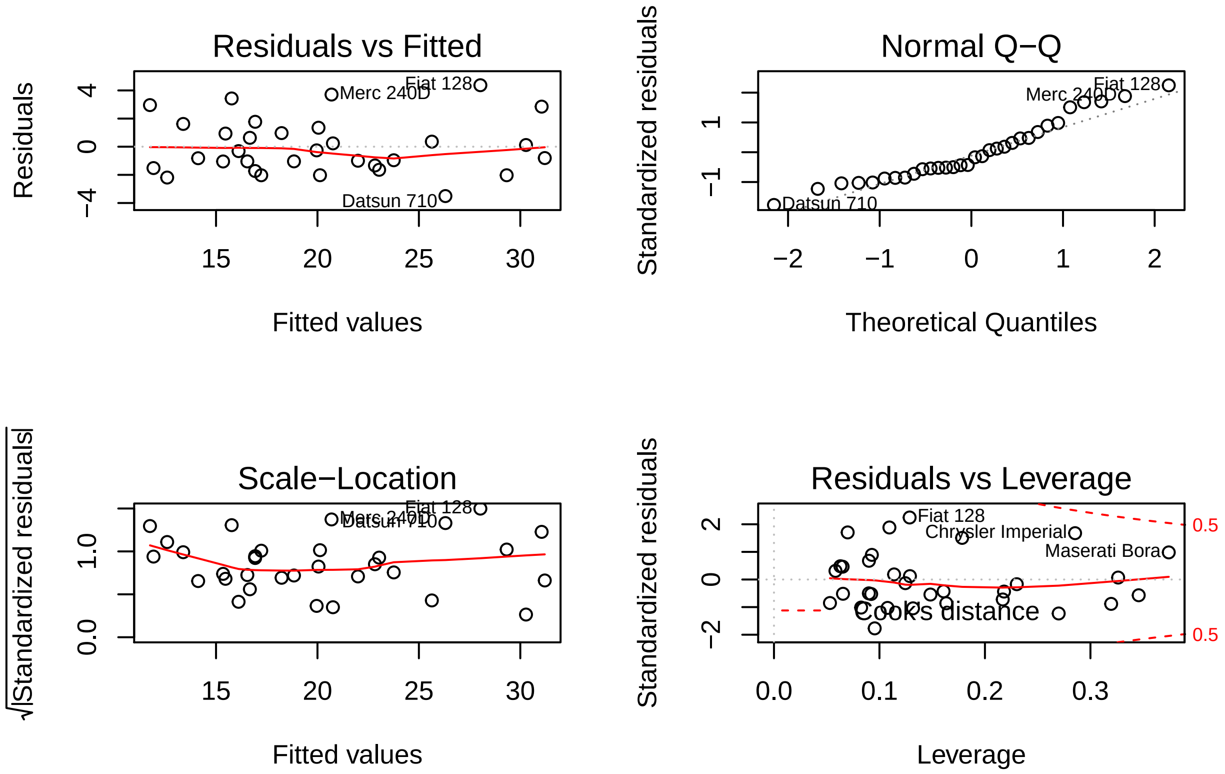
3. Scatter Plot of MPG vs. Weight by Transmission

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