Peer-graded Assignment: Regression Models Course Project

Simon Grasdal

2024-12-15

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
data(mtcars)
MTcars <- mtcars
```

Executive Summary

This is a report analyzing the mtcars dataset provided by R Studio. The purpose is to explore the relationship(s) that exist between transmission type and the Miles Per Gallon (MPG) extracted from the 1974 Motor Trend US magazine. The sample is comprised of 32 cars (1973-74 models) and explored 10 aspects of each vehicle including but not limited to: number of cylinders (cyl), weight (wt), and horsepower (hp). In this analysis, regression models are used to explore how **automatic** (am = 0) and **manual** (am = 1) transmissions impact **MPG**. 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, several linear regression models are fitted and one with highest Adjusted R-squared value is selected. So, given that weight and 1/4 mile time are held constant, manual transmitted cars are 14.079 + (-4.141) **weight** more **MPG** 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.

Analysis

Exploratory Data Analysis

```
MTcars[1:5,]
```

```
##                   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
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

```
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
```

Our cursory look at the data tells us that there is a difference between the two transmission types. In the boxplot depicted in **Appendix Fig. 1** we can see that the manual transmission vehicles have higher **MPG** values. Likewise the pairwise correlations shown in **Appendix Fig.2** show correlations between other such variables like weight, displacement, and horsepower.

Inference

To test if the automatic and manual transmissions are sufficiently different we use a two-sample t-test. The null hypothesis for this test is that there is no significant difference between the two transmissions types. The alternate hypothesis states that there is a difference.

```
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
```

Our result indicates that there is a significant difference in **MPG** between the two kinds of transmission types ($T = 24.39$, $p < .05$) thus rejecting the null hypothesis. This means that there is a significant difference in **MPG** between automatic and manual transmission types.

Regression

First, we must fit the full model without any interference.

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

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## 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:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

The results indicate 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. However, we can see that **weight** (wt) is quite close to significance ($p = .06$). To confirm our suspicions we run a stepwise model of the regression.

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

```
## Start:  AIC=87.02
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - cyl   1    0.0799 147.57 83.572
## - vs    1    0.1601 147.66 83.590
## - carb  1    0.4067 147.90 83.643
## - gear  1    1.3531 148.85 83.847
## - drat  1    1.6270 149.12 83.906
## - disp  1    3.9167 151.41 84.394
## - hp    1    6.8399 154.33 85.006
## - qsec  1    8.8641 156.36 85.423
## - am    1   10.5467 158.04 85.765
## <none>              147.49 87.021
## - wt    1   27.0144 174.51 88.937
##
## Step:  AIC=83.57
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 80.165
## - carb  1    0.5201 148.09 80.219
```

```
## - gear  1    1.8211 149.40 80.499
## - drat  1    1.9826 149.56 80.534
## - disp  1    3.9009 151.47 80.942
## - hp    1    7.3632 154.94 81.665
## - qsec  1   10.0933 157.67 82.224
## - am    1   11.8359 159.41 82.575
## <none>              147.57 83.572
## - wt    1   27.0280 174.60 85.488
##
## Step:  AIC=80.16
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 76.847
## - gear  1    2.1437 149.99 77.160
## - drat  1    2.2139 150.06 77.175
## - disp  1    3.6467 151.49 77.479
## - hp    1    7.1060 154.95 78.201
## - am    1   11.5694 159.41 79.110
## - qsec  1   15.6830 163.53 79.925
## <none>              147.84 80.165
## - wt    1   27.3799 175.22 82.136
##
## Step:  AIC=76.85
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 73.717
## - drat  1     1.932 150.46 73.795
## - disp  1    10.110 158.64 75.489
## - am    1    12.323 160.85 75.932
## - hp    1    14.826 163.35 76.426
## <none>              148.53 76.847
## - qsec  1    26.408 174.94 78.618
## - wt    1    69.127 217.66 85.610
##
## 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 ***
## am            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.

According to the scatter plot (**Appendix Fig. 3**), 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. Thereby, following model including the interaction term is generated:

```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 ***
## am            14.079      3.435   4.099 0.000341 ***
## wt:am         -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, which is pretty good.

Next, the simple model is fitted with MPG as the outcome variable and Transmission as the predictor variable.

```r
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 ***
## am             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 other variables should be added to the model.

Finally, the final model is selected: mpg ~ wt + qsec + am + wt:am.

```
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 39.2687 8.025e-08 ***
## 3     21 147.49  7     21.79  0.4432    0.8636
## 4     27 117.28 -6     30.22
## ---
## 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
## am           7.0308746 21.127981
## wt:am       -6.5970316 -1.685721
```

```
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
## am          14.079428  3.4352512  4.098515 0.0003408693
## wt:am       -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.

Analysis of Residuals

According to the residuals plots shown in **Appendix Fig. 4**, the following assumptions can be verified concerning our model. 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.

Appendix

Figure 1

Boxplot of MPG vs. Transmission Type

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

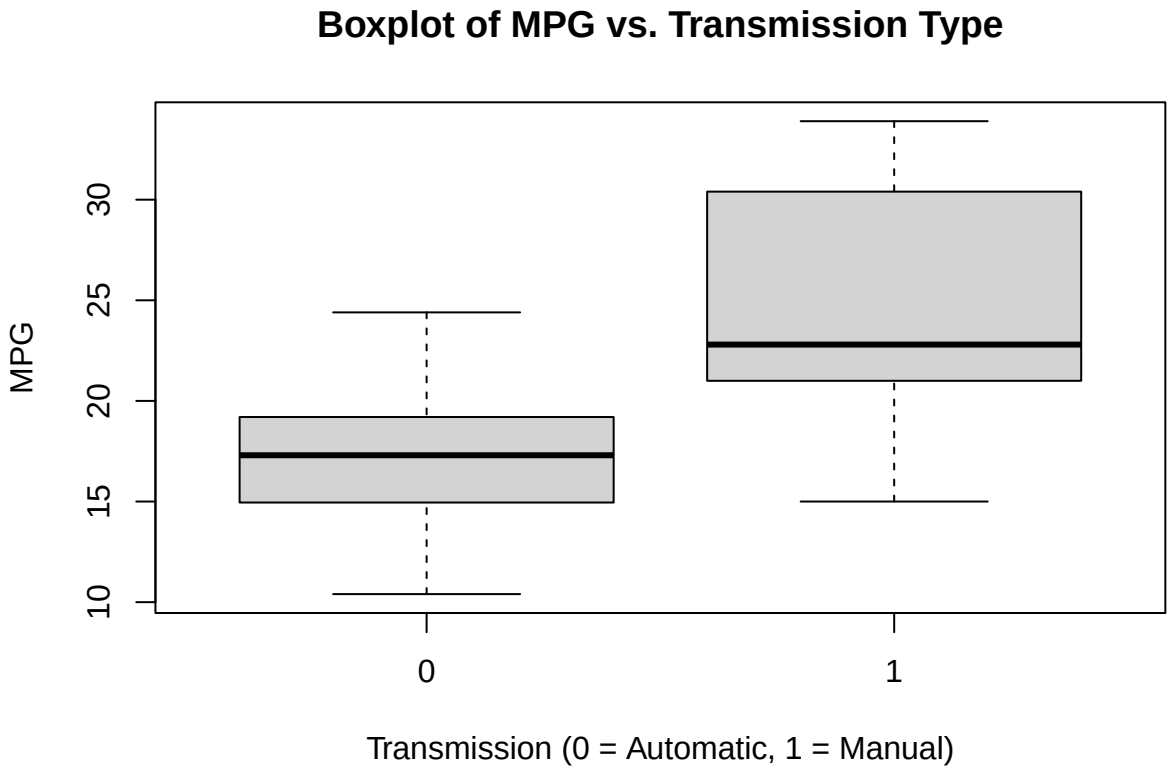
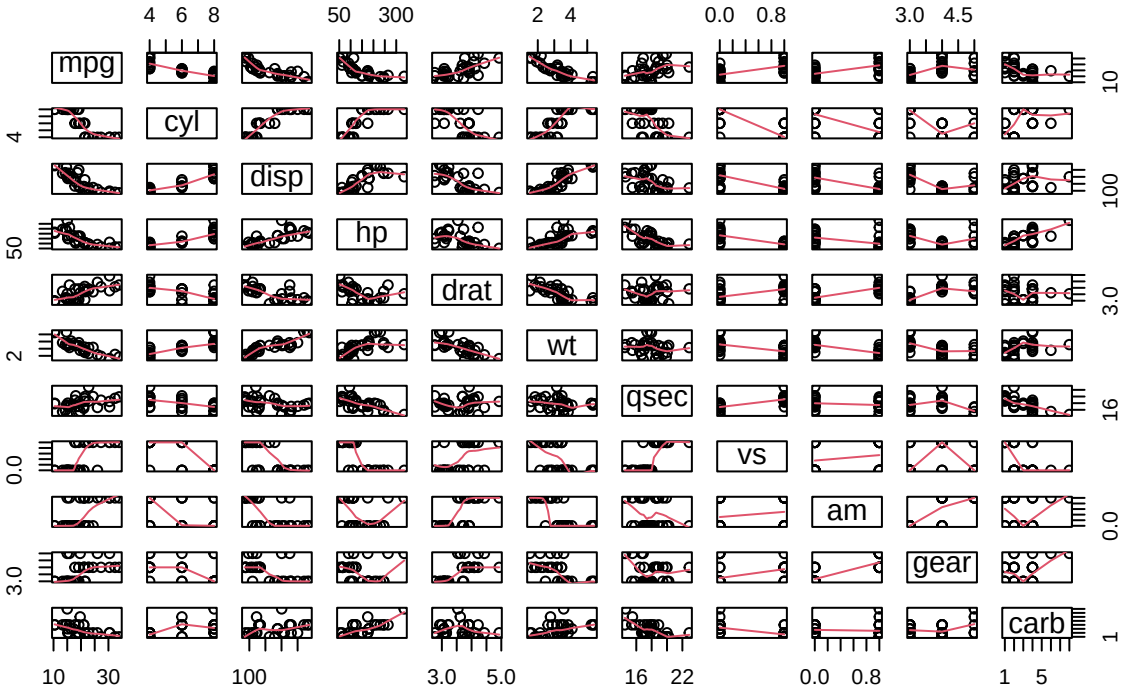


Figure 2

Pairwise Correlation Matrix

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




```
fig.align="center"
```

Figure 3

Scatterplot of MPG vs. Weight by Transmission

```
ggplot(MTcars, aes(x=wt, y=mpg, group=am, color=am, height=5, width=5)) + geom_point() + scale_colour_di
```

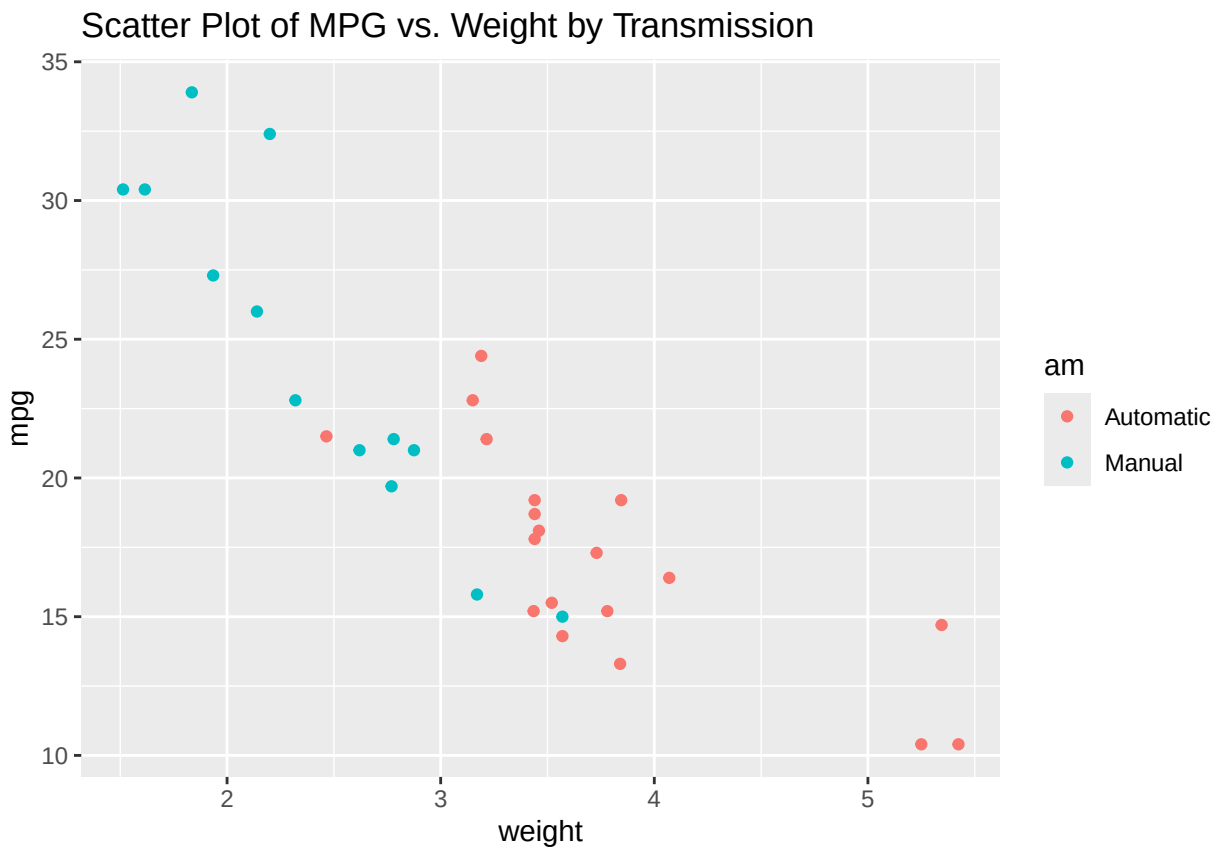
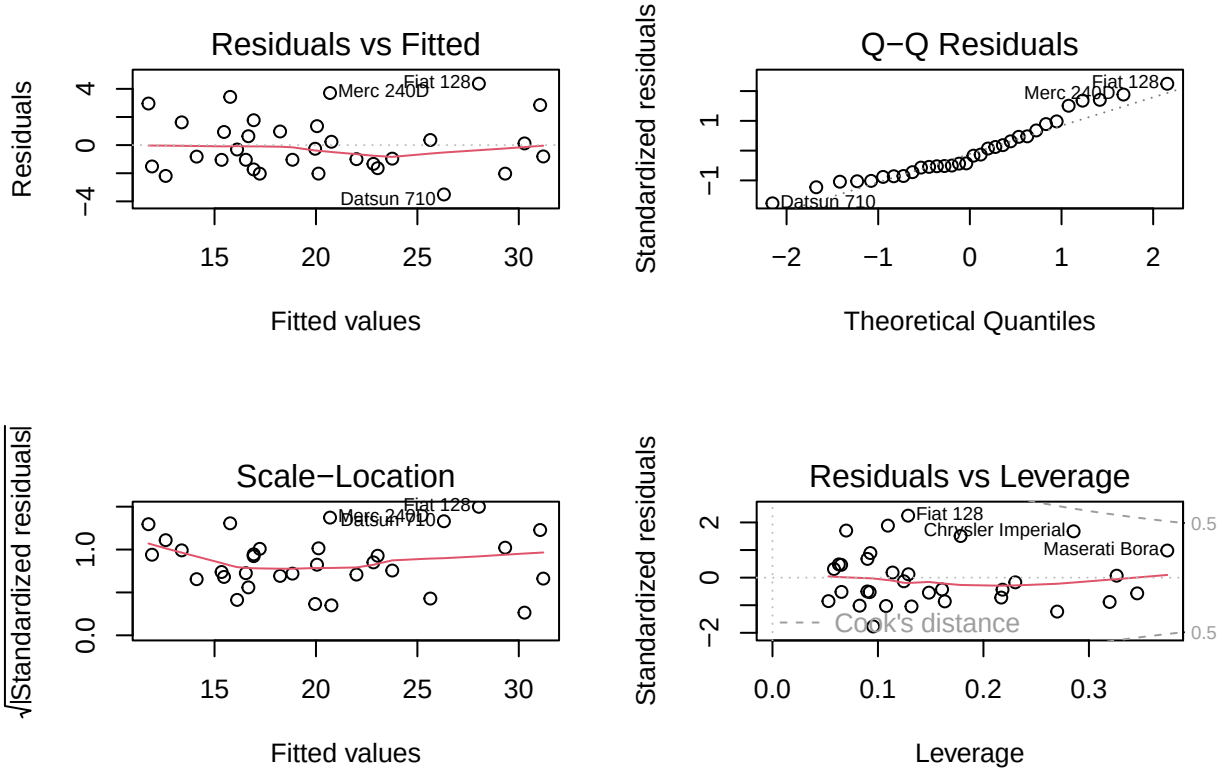


Figure 4

Residuals

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


Residuals vs Fitted
Residuals
Fitted values
Merc 240D
Fiat 128
Datsun 710
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Fiat 128
Merc 240D
Datsun 710
Scale−Location
√|Standardized residuals|
Fitted values
Fiat 128
Merc 240D
Datsun 710
Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5