Regression Models: Course Project

Fuel Consumption analysis - Filomena Ciccarelli 30th April 2016

Executive Summary

This report looks at a data set of a collection of cars (*mtcars*) and explores the relationship between a set of variables and miles per gallon (MPG). Specifically, it analyses whether manual transmission cars are better than automatic cars for MPG and quantify the MPG difference between the two transmission systems. The data used for the analysis was extracted from the *1974 Motor Trend US magazine*, and comprises fuel consumption (MPG) and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

The analysis seems to suggest that the types of transmission (Automatic or Manual) play a key role in fuel consumption.However, other car factors such as horse power, weight, and number of cylinders also have a an impact on MPG, albeit with different levels of magnitude.The report concludes that manual transmission cars consume more fuel than automatic cars when all other factors are held constant.

The analysis report and the associated code can be found on GitHub.

Exploratory Data Analysis

```r
library(ggplot2)
library(plyr)
library(dplyr)
library(magrittr)
library(knitr)
data(mtcars)
dim(mtcars)
attach(mtcars)
```

The data has 0 NA values and 32 x 11 dimensions as indicated above.

```r
str(mtcars)
```

The data structure indicates that it is necessary to convert some variables from *numeric* to *factor* class.

```r
mtcars$am<-factor(mtcars$am)
mtcars$cyl<-as.factor(mtcars$cyl)
mtcars$gear<-factor(mtcars$gear)
mtcars$vs<-as.factor(mtcars$vs)
mtcars$carb<-factor(mtcars$carb)
```

The variation of fuel consumption by transmission type indicates that manual transmission yields higher values of MPG (see Appendix ”*FuelConsumptionbyTransmissionType*” plot) i.e. cars with a manual transmission seem to have lower fuel consumption than automatic cars. However, the pairs graph in the Appendix showes high correlation with the variables number of cylinders *cyl*, gross horse power *hp*, weigth *wt* and car displacement *disp*. In the next section we carry out the regression and statistical significance analysis.

Regression and Statistical Significance Analysis

We assume that the data *mtcars* is randomly sampled. Also, as the data is non-paired it means that the indipendence condition between the 32 groups (different car models) is met.

Effect of Transmission type on fuel consumption - Inference analysis

In this section we assess the statistical significance that fuel consumption varies by transmission. We use the two sample T-test to show this (both samples have n<30) ($H_0 : \mu_{automatic} = \mu_{manual}$. $H_A : \mu_{automatic} \neq \mu_{manual}$):

```
ttest<-t.test(mpg~am,data=mtcars)
print(ttest)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Since the p-value 0.0014 < 0.005, we reject the null hypothesis as the data provides strong evidence that fuel consumption differs between car transmission systems. The 95% confidence levels indicates that, based on the *mtcars* data, cars with for automatic transmission drive between *3.2* and *11.3* less MPG than manual transmission cars. The average MPG for the latter is circa 7 more than the average MPG for autiomatic transmission cars.

Effect of Multiple variables on Fuel Consumption - Regression analysis

Building up from the finding in the previous paragraph, in this section we asses whether the difference in fuel type by transmission types holds when the other variables are included in the regression model. If we were to fit the simple model observed above with MPG as outcome and Transmission predictor value:

```
amfit<-lm(mpg ~ am,data=mtcars)
```

The model has an Adjusted R-squared value standard error of *0.3385* (see Appendix) which means that the model can only explain about 34% (the remaining variance is not explained by the model) of the variance of the variable MPG. This indicates that other variables need to be added to the model.

We now explore the full model

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

If we look at the Adjusted R-squared value standard error for this model, this is *0.779* (see Appendix) which means that the full model can actually explain c. 78% of the MPG variance. However, the p-value of all the coefficient estimates are statistically not significant (greater than the significant level 0.05) which indicates that not all the variables are necessary.

Starting from the full model above, we use the *step* function with the backward direction to select some statistically significant variables.

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

The analysis of the variance table gives the model *mpg~wt + qsec + am*. The Adjusted R-squared value standard error for this model, this is *0.850* (see Appendix) which means that the model can explain about 85% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level. The variables weight *wt*, time for 1/4 mile *qsec* and transmission *am* as regressors provide a good fit fo the MPG variation. As a final step, we now check if there is an interaction between the regressor variables. In the Appendix we show that cars with automatic transmission (am=0) weigh more than the automatic transmission cars (am=1) and therefore have a higher petrol consumption. This indicates that there may be an interaction term between weight *wt* and transmission type *am* in the regression model:

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

The Adjusted R-squared value standard error for this model is *0.880* (see Appendix) which means that the model, with the interaction term between weigth and transmission, can explain about 88% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

Conclusion on the Model selection

We select **mpg ~ wt + qsec + am + wt : am** model for the MPG outcome as it provides the best fit in terms of minimum standard error and statistical significance.

```
##              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
```

The result shows that when *wt* weight (lb/1000) and *qsec* (1/4 mile time) remain constant, cars with manual transmission (am1) add *14.079 + (-4.141) * wt* more miles per gallon on average than cars with automatic transmission.

Residual Analysis and diagnostic

The residual plots of the chosen model are in the Appendix. The following observations can be made:

1. Residual vd. Fitted: supports the indipendency assumption (no visible pattern)
2. Normal Q-Q: residuals lie close to the line (normal distributed)
3. Scale-Location: points are randomly distributed
4. Residuals vs. Leverage: all values are within the 0.5 band (no outliers)

The sum of the abs values of the dfbetas is 20.9 >1 which indicates that no observation has impacted the estimate of a regression coefficient.

Appendix

Exploratory Data Analysis

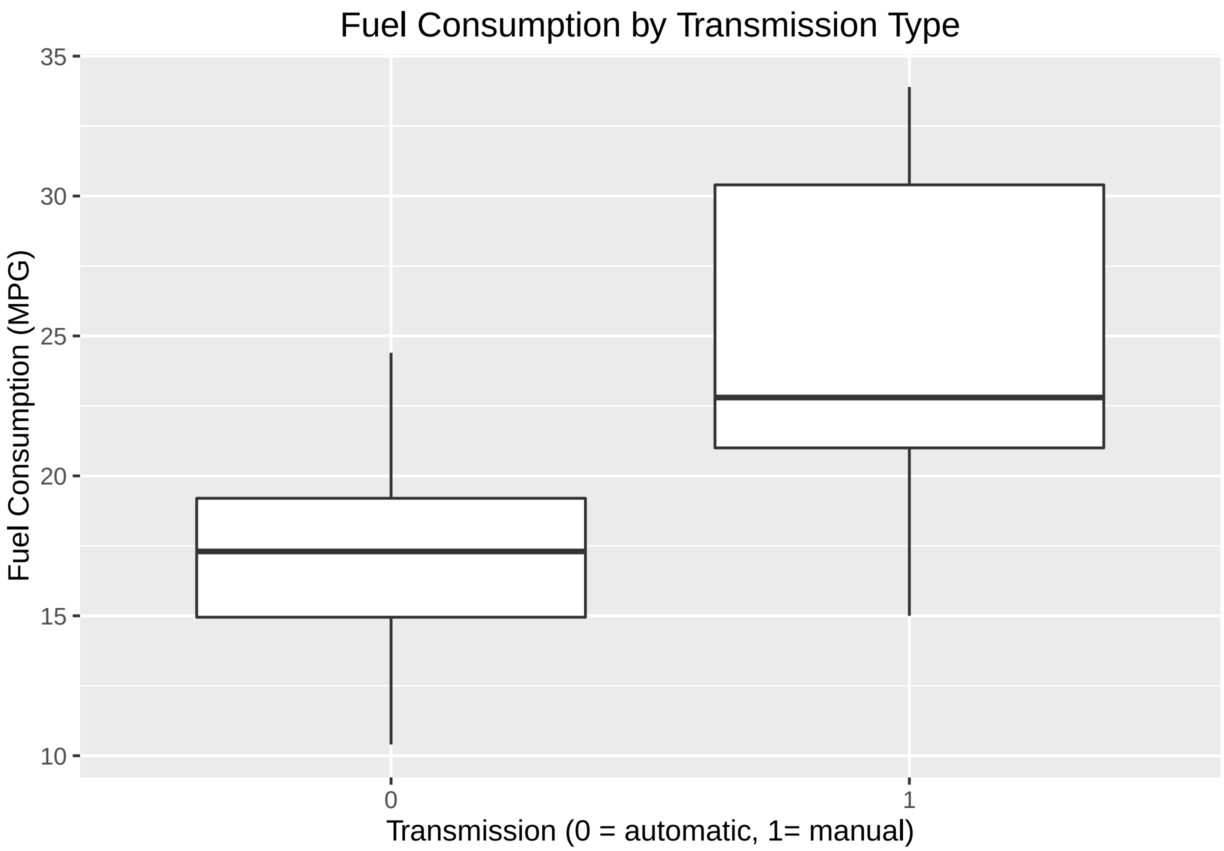
Number of cars by transmission type

```
table(mtcars$am)
```

```
##
##  0  1
## 19 13
```

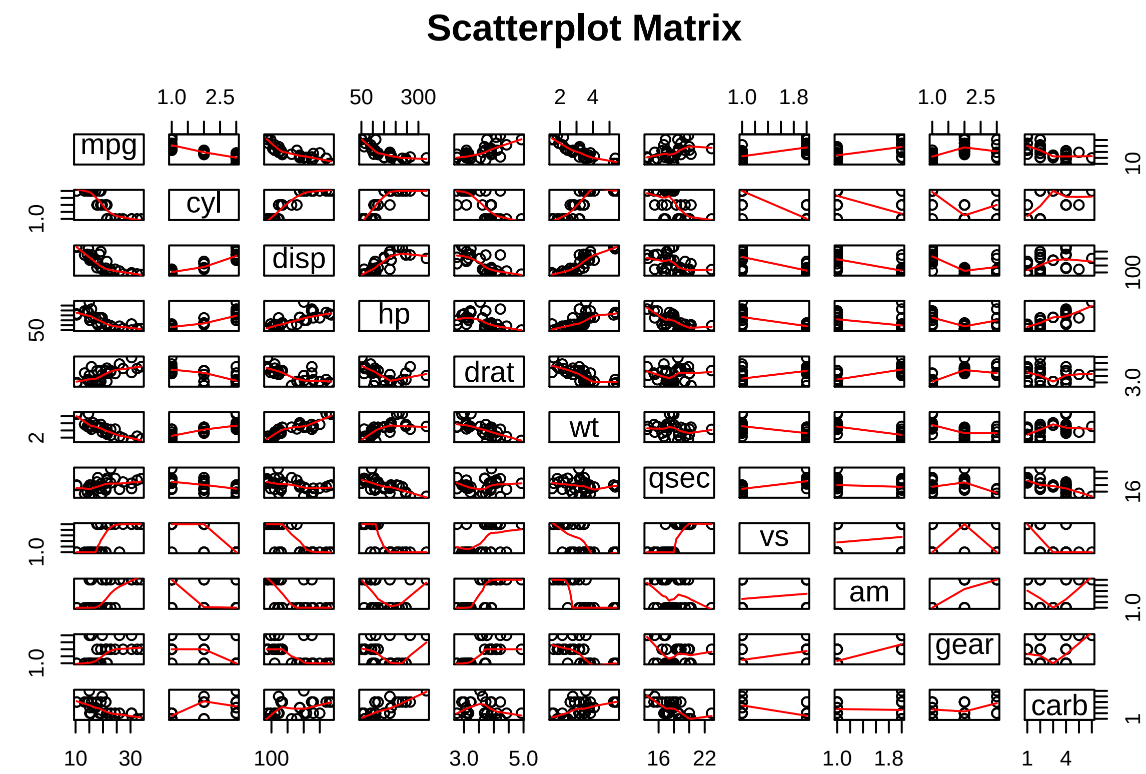
Fuel consumption by transmission type

```
ggplot(mtcars,aes(am, mpg)) + geom_boxplot() +
        xlab("Transmission (0 = automatic, 1= manual)") +
        ylab("Fuel Consumption (MPG)") +
        ggtitle("Fuel Consumption by Transmission Type")
```



###Pairs chart

```
pairs(mtcars, panel=panel.smooth, main="Scatterplot Matrix")
```



The charts on the top line shows the variation of *mpg* (y-axis) with the other variables (x-axis). Variables such as horse power *hp*, weight *wt*, and number of cylinders *cyl* also seem to have an impact on MPG (with different levels of magnitude).

Regression and Statistical Significance Analysis

Transmission only fit

```
amfit<-lm(mpg ~ am,data=mtcars)
summary(amfit)
```

```
##
## 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
```

Full model

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

```
##
## 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
```

Step fit

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

```
summary(stepModel)
```

```
##
```

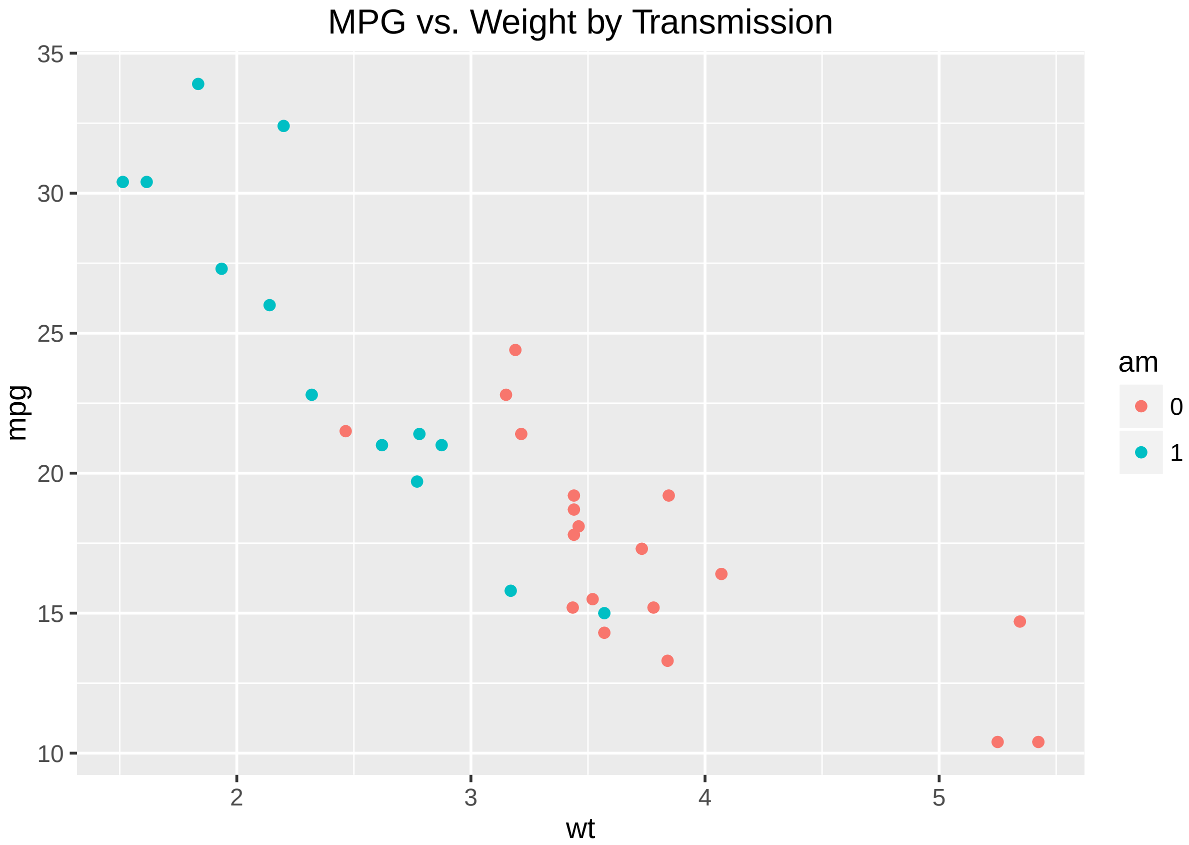
```
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -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
```

```r
anova(stepModel)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## wt         1 847.73  847.73 140.2143 2.038e-12 ***
## qsec       1  82.86   82.86  13.7048 0.0009286 ***
## am         1  26.18   26.18   4.3298 0.0467155 *
## Residuals 28 169.29    6.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Scatter Plot for Interaction Model

```r
ggplot(mtcars,aes(x=wt,y=mpg,color=am))+geom_point()+ggtitle("MPG vs. Weight by Transmission")
```



Cars with automatic transmission (am=0) weigh more than the automatic transmission cars (am=1) and therefore have a higher petrol consumption. #### Interaction Model

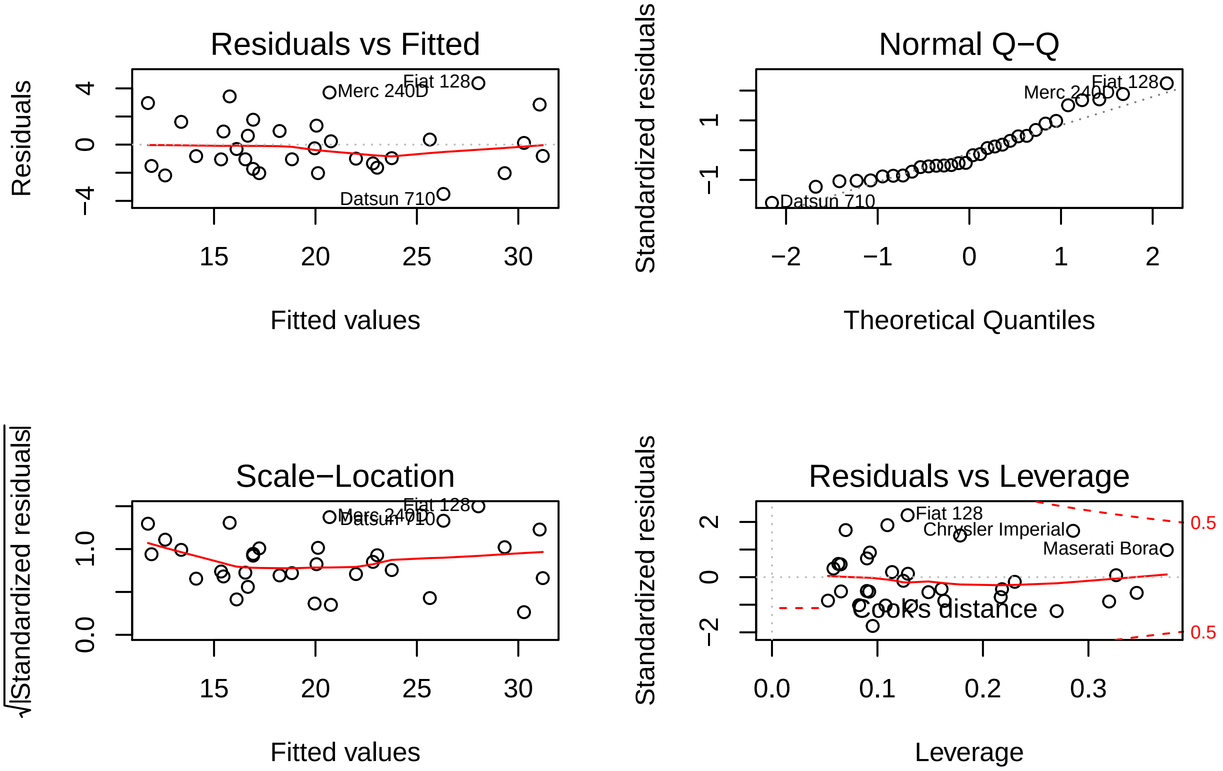
```
IntModel<-lm(mpg ~ wt + qsec + am + wt:am, data = mtcars)
```

```
summary(IntModel)
```

```
##
## 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
```

Residual Plots

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



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