Motor Trend Data Analysis - Regression Models Course Project

Aroge F. Akindele

December 14, 2016

Executive Summary

This project analyses the `mtcars` dataset from the *Motor Trend* US magazine. The relationship between the `mpg` variable and the other variables is examined and their effects. Particularly, we want to know which of the transmission types is good for the `mpg` feature. We also try to quantitaviely describe the relationship. The t confidence interval is observed for the variation in the transmission types and we were able to ascertain that, the different transmission types were significant to the data. Different model were then fitted to get a best fit. These models were analysed and a suitable model was achieved. We were able to determine the relationship between a car with manual and automatic transmission types leaving other variables constant. Specifically the model implied, given that weight and 1/4 mile time are held constant, manually transmitted vehicles are 14.079 + (-4.141)*wt more in the `mpg` values than automatic ones. Looking at the value, we deduce that a light manual transmission and a heavy automatic transmission car have higher mpg values.

Exploring Data Analysis

We try to explore the data set to gain some insights as well as prep the data for further analysis.

```r
data("mtcars")
head(mtcars, 2)
```

```
##               mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

```r
mtc <- mtcars
#convert the tyes of the required variables
mtc$am <- factor(mtc$am)
mtc$cyl <- factor(mtc$cyl)
mtc$vs <- factor(mtc$vs)
mtc$gear <- factor(mtc$gear)
mtc$carb <- factor(mtc$carb)
```

```r
fit <- lm(mpg~am, data = mtc)
coef(summary(fit))
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

It is observed that the the estimate for `am1`(manual transmission) as shown in the output above is in comparison with the intercept (`am0`), which is the automatic transmission.

We may take a null hypothesis that the effect of transmission on `mpg` is independent of transmission type. So we just proceed to compare automatic with anual since we have a binary column. The confidence interval for the `am1` coefficient is also calculated below:

```
confint(fit)
```

```
##                2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## am1          3.64151 10.84837
```

From the above, we get a significantly low p-value for the manual transmission (`am1`) of 2.850207e-04 with reference to automatic. The confidence interval does not contain zero and so we reject the null hypothesis that there is no effect in the type of transmission on `mpg`

Regression Analysis

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

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

From the result, the model has an adjsuted Adjusted R-squared: 0.779 but none of the variables are statistically significant, with p-values all greater than .05 For a best model selection, the step function is used;

```
modelBest <- step(fullfit, k = log(nrow(mtc)), trace = F)
summary(modelBest)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtc)
##
## 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 reveals that `wt`, `qsec`, `am1` have been chosen as features for the best fit. It also shows an improved value of 0.8336 for the adjusted R-squared. Furthermore, all of the coefficients are significant at the 0.05 significant level.

Implementing the Nested mode testing

```
fit1 <- lm(mpg~wt, data= mtc);fit2 <- update(fit1, mpg~wt+qsec)
fit3 <- update(fit2, mpg~wt+qsec+am)
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + qsec
## Model 3: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 278.32
## 2     29 195.46  1    82.858 13.7048 0.0009286 ***
## 3     28 169.29  1    26.178  4.3298 0.0467155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indeed, each of the additional parameter shows significance at the 0.05 level. Looking at the pairs plot (Appendix 2), it indicates a relationship between the `wt` and the `am` variables. We may want to add this to our model to cater for this interaction. So we have:

```r
fit4 <- lm(mpg~ wt+qsec+am+wt:am, data = mtc); summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt:am, data = mtc)
##
## 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
```

With our new model fit (fit4), we see an improved model which explains about 88% of the variation in the data. The estimates of the coeficcients tell that, with `wt` (weight in 1000lb) and `qsec` (1/4 mile time) kept constant, a car with manual transmission is 14.079 + (-4.141)*wt greater than that with an automatic transmission.

Residual Analysis

1. The first plot in the residual plots (Appendix 3) doesn't seem to show off any obvious pattern which implies we may take the residuals as randomness in the data.
2. The Normal Q-Q plot also show the plot fairly lying across the dotted line, implying the residual distibution is fairly normal
3. The Scale-Location plot doesn't show off any systematic pattern as well
4. The Residuals vs. Leverage plot plot also follows the dotted line closely, also all values fall within the 0.5 bands.

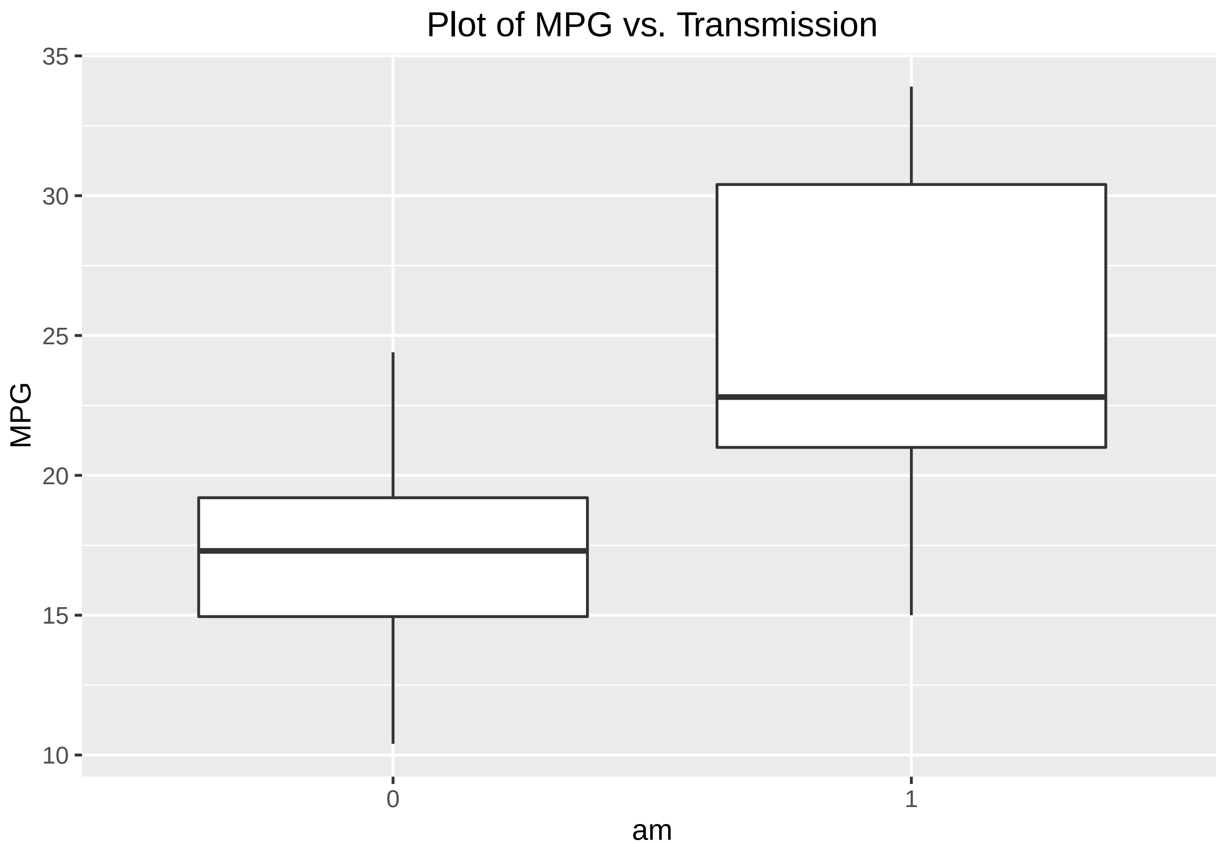
Conclussion

From the fore going it is obvious that the `fit4` appears to be the best fit having gotten the key predictors using the step function as well as including the perceived interation between the `wt` and `am` variables. This model explains about 88% of the variation in the data with minimal features and the residual plots show no systematic variations.

Appendix

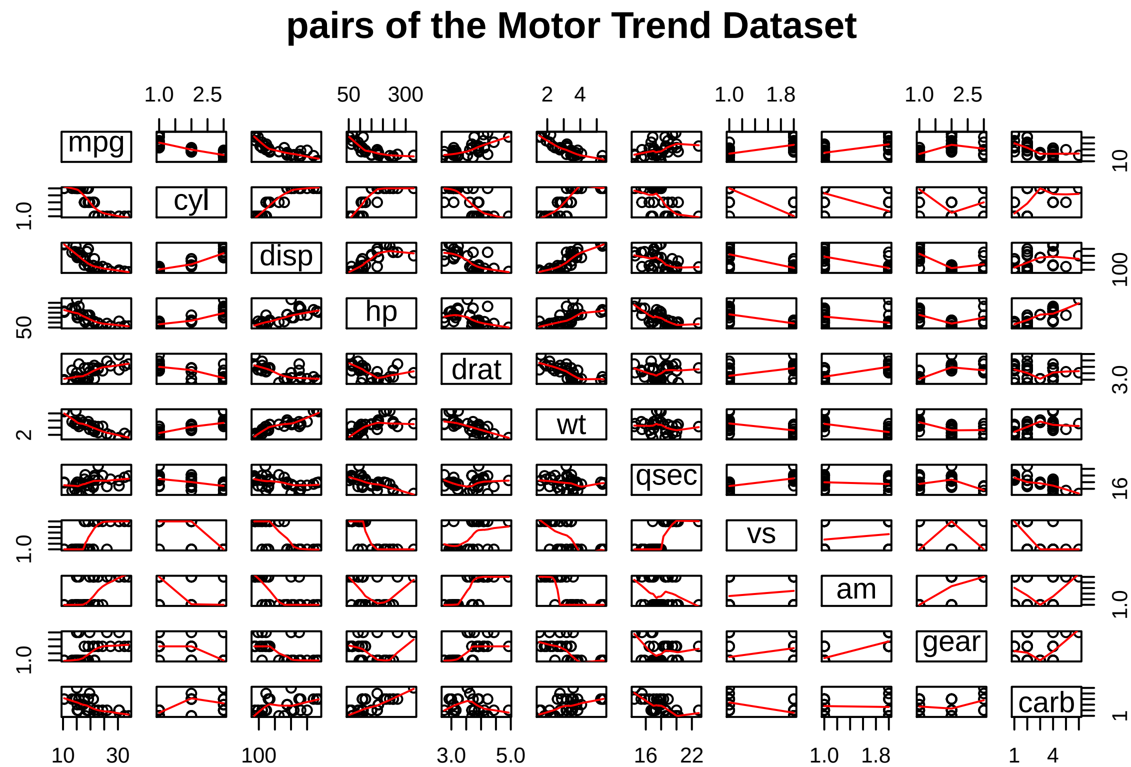
1. Plot of MPG vs. Transmission

```r
compareTompg <- function(variable){
ggplot(data = mtc)+geom_boxplot(mapping = aes_string(variable, "mpg"))+
    labs(x= variable, y= "MPG", title= "Plot of MPG vs. Transmission")}
compareTompg("am")
```



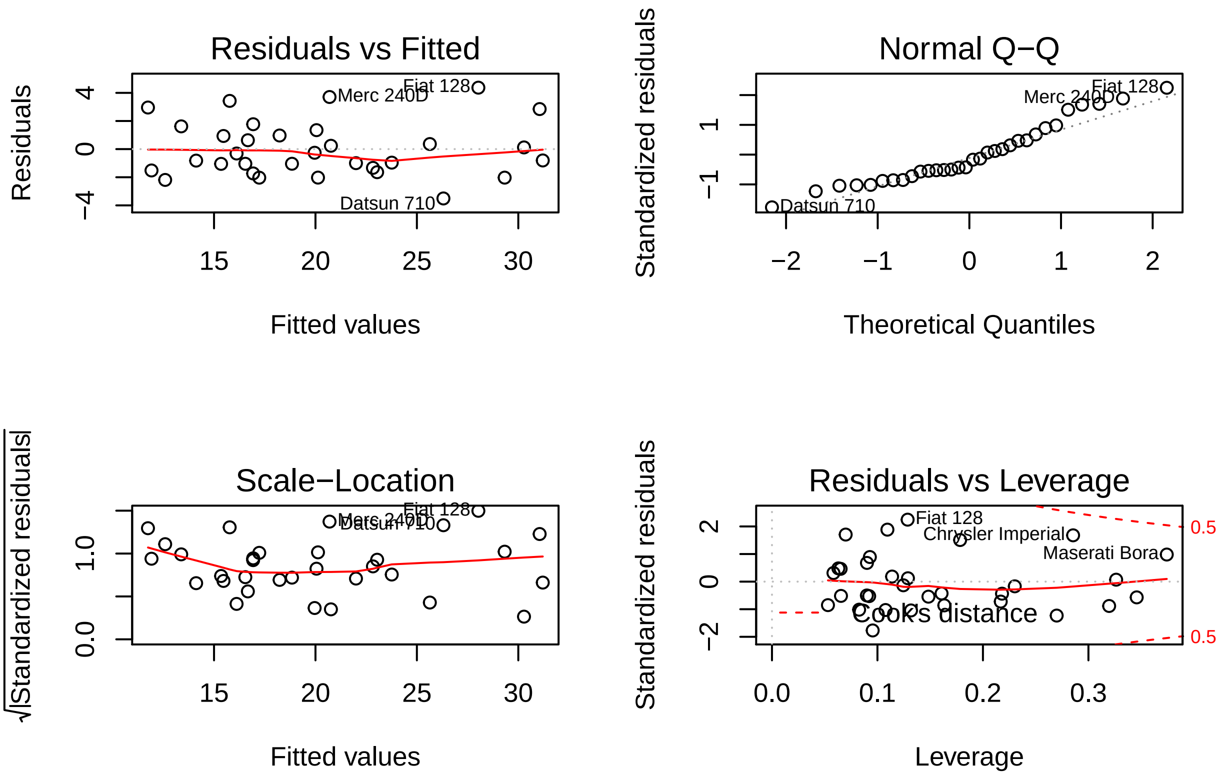
2. Pairs plot for the dataset

```r
pairs(mtc, panel = panel.smooth,main= "pairs of the Motor Trend Dataset")
```



3. Residual Plots

```
par(mfrow=c(2,2)); plot(fit4)
```


Residuals vs Fitted
Residuals
4
0
−4
15
20
25
30
Fitted values
Fiat 128
Merc 240D
Datsun 710
Normal Q−Q
Standardized residuals
1
−1
−2
−1
0
1
2
Theoretical Quantiles
Fiat 128
Merc 240D
Datsun 710
Scale−Location
√|Standardized residuals|
1.0
0.0
15
20
25
30
Fitted values
Fiat 128
Merc 240D
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