Relationship between a set of variables and miles per gallon

Summary

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 are interested in two question: "Is an automatic or manual transmission better for Mpg?" and "Quantifying how different is the Mpg between automatic and manual transmissions?" After the data analysis I can say there is a significant relationship between the fuel consumption and the transmission type so the manual transmission is better for Mpg. After experimenting with linear models I found that third model describes Mpg variable the best.

Calculating the mean of each transmission types

```r
data(mtcars)
automean<-mean(mtcars$mpg[mtcars$am=="0"])
manualmean<-mean(mtcars$mpg[mtcars$am=="1"])
```

The mean for manual transmission is 24.3923 Mpg and 17.1474 Mpg for automatic transmission. This is a significant difference in the means. Manual transmissions have a higher value so based on mean of Mpg it is better (cheaper) to have manual transmission.

Regression model of transmission type affecting the car's fuel consumption

Let's make a simple linear regression model with Mpg as depending variable and am as explanatory variable. We get following model: Mpg = 17.147 + 7.245 * am. Since manual transmission is denoted with am = 1, we can conclude that we can make more Mpg with manual transmission.

See the boxplot as fig 1 for visual explanation.

```r
fit1 <- lm(mpg~am, data=mtcars)
summary(fit1)
```

```
##
## 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 ***
## am              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
```

Finding the best model

Let's fit a model with all explanatory variables. R-squared = 0.869

```
fitAll<-lm(mpg~.,mtcars)
summary(fitAll)
```

```
##
## 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
## am            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
```

P values are high so we can use only those variables with low p values, for example p, wt, qsec and am. Actually, we got 0.8579 R-squared for model lm(mpg~hp+wt+qsec+am,mtcars) which is lower than fitAll model, but with a little experimenting, I got the following model with 0.8942 R-squared:

```
fit2<-lm(mpg~hp*wt+qsec+am,mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ hp * wt + qsec + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.394 -1.441 -0.271  1.253  4.150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.93948   10.26506    3.50  0.00169 **
```

```
## hp          -0.09776    0.02954   -3.31  0.00274 **
## wt          -7.92300    1.75109   -4.52  0.00012 ***
## qsec         0.59727    0.39232    1.52  0.13998
## am           0.91131    1.40071    0.65  0.52101
## hp:wt        0.02516    0.00841    2.99  0.00602 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.14 on 26 degrees of freedom
## Multiple R-squared:  0.894,  Adjusted R-squared:  0.874
## F-statistic:   44 on 5 and 26 DF,  p-value: 7.18e-12
```

Comparing models

We got 2 models, second one combining variables describes Mpg the best.

See fig 2 and 3 for visual explanation.

```
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ hp * wt + qsec + am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     26 119  4       602 32.8 8.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P value is lower than 0.05 which means that model fit2 is significant improvement over model fit1

Appendix

Figure 1

```
m<-transform(mtcars, am=factor(am))
levels(m$am)[1] <- "Automatic"
levels(m$am)[2] <- "Manual"
plot(m$am , m$mpg, main="MPG vs Transmission", xlab="Transmission", ylab="MPG", )
abline(lm(mpg ~ am, m), col="red")
```

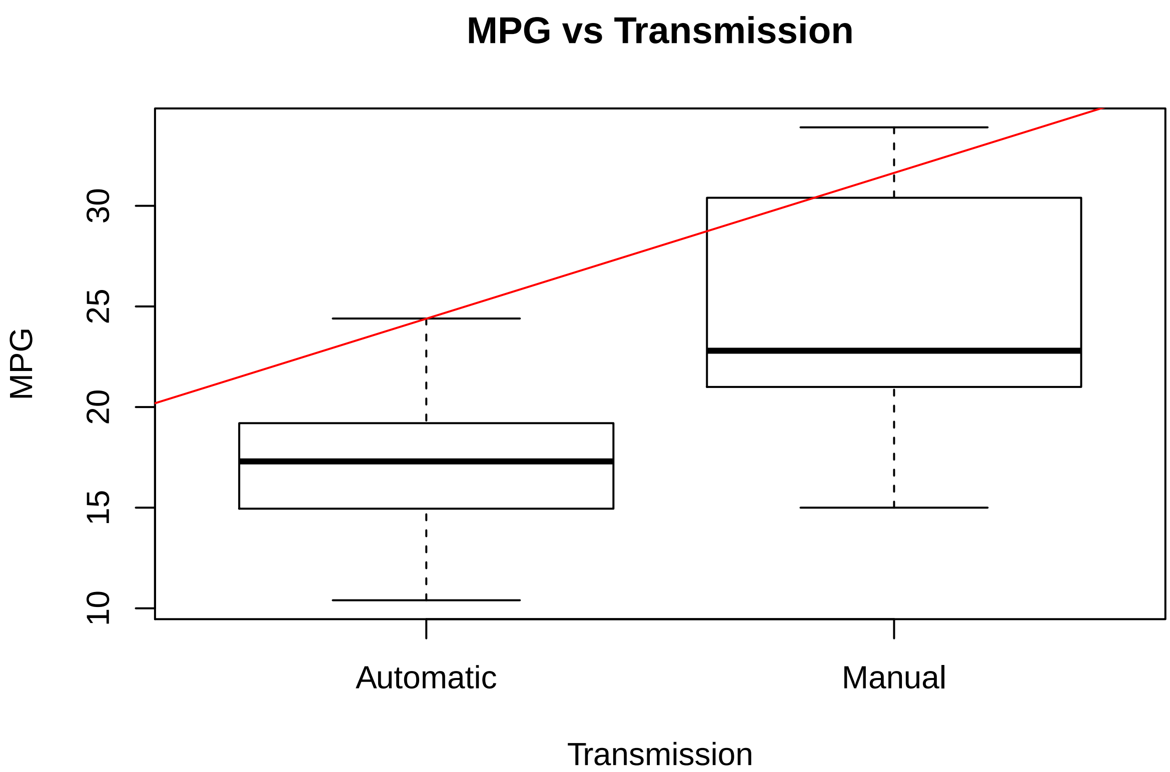


Figure 2

```
layout(matrix(c(1,2,3,4),2,2))
plot(fit1)
```

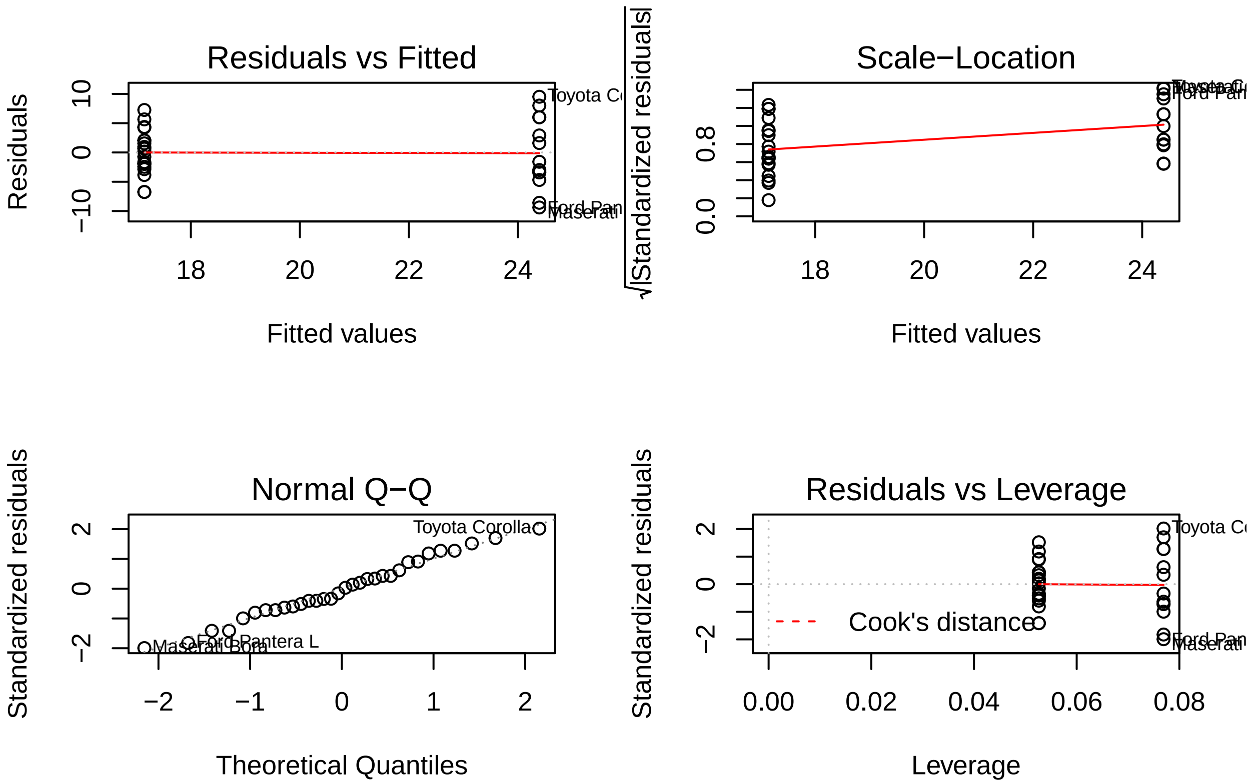
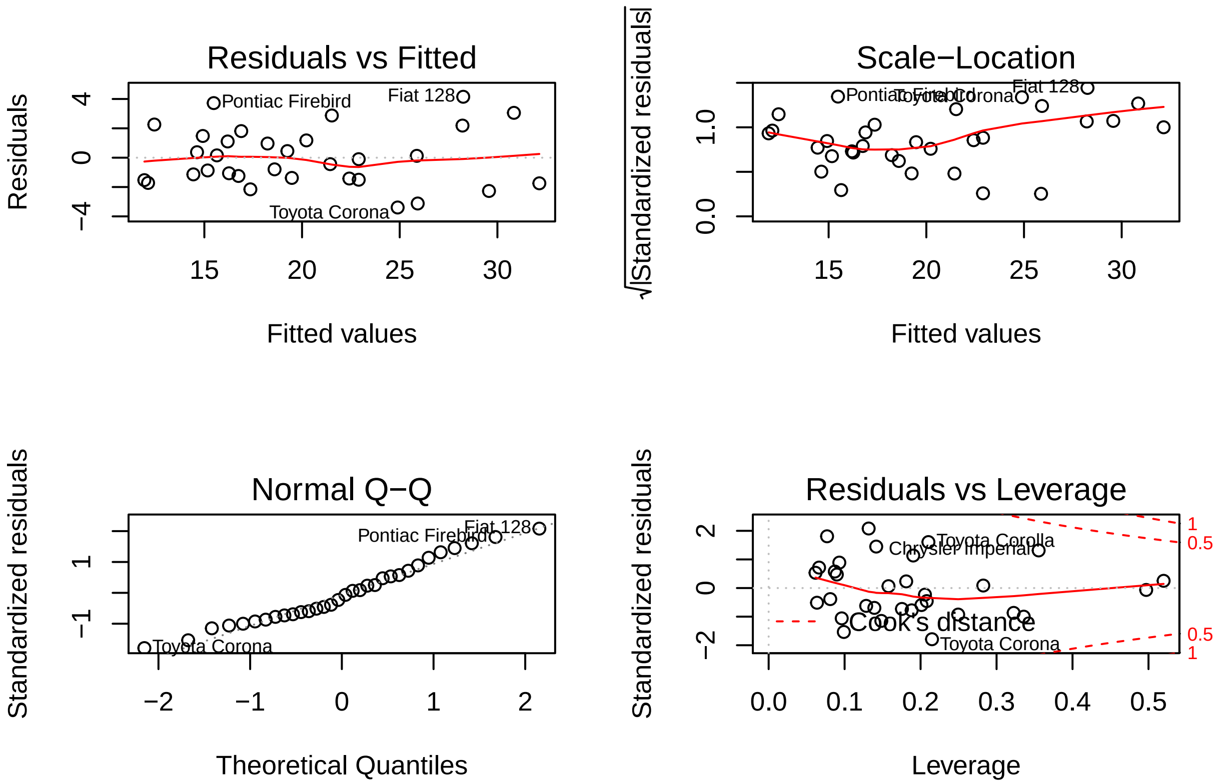


Figure 3

```
layout(matrix(c(1,2,3,4),2,2))
plot(fit2)
```


Residuals vs Fitted
Residuals
Fitted values
Pontiac Firebird
Fiat 128
Toyota Corona
Scale−Location
√|Standardized residuals|
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Cook's distance
0.5
1