```
# options(rpubs.upload.method = 'internal')
```

Factors Affecting MPG

Executive Summary

The goal of this is to explore the relationship between a set of variables and miles per gallon (MPG) (outcome) using the mtcars dataset. The questions of particular interest are: is an automatic or manual transmission better for MPG and quantify the MPG difference between automatic and manual transmissions. The research described below suggests that manual transmission does provide better mileage, however this result is not statistically significant. Likely the result could be improved with bigger dataset; mtcars dataset imcludes only 31 observation.

Exploratory Analysis

The simplistic approach to questions regarding the diffenece between automatic and manual transmission would be to run a single valiable regrerssion

The resulted model (mpg = 17.147+ 7.245*am) suggests that manual transmission causes extra 7.245 mpg, on average. However this model can explain only 36% of the variance (can be seen from the output of the summary(fit0) command)

More prudent approach would be to consider all variables, select the best model and see if the “am”“ variable should be included into model or not.

First, let's look at all variable. Just for the sake of exploratory analysis, let's run regression with all available variables:

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

Of course, this is not a good model: with 10 variables and 31 observationswe have severe overfit. Also, as we can see from the figure 1 in the Appendix, many variablres are heavily correclated.

However let's take a look at one of the lines of the output of the summary command:

Multiple R-squared: 0.869, Adjusted R-squared: 0.8066

We see that this model explains 87% (R-squared: 0.869) so we should aim for our model to be not to far form this number. Another variable too look at is Adjusted R-squared (0.8066) that adjusts R squared for the number of variables. Our model should exceeed 0.8066 as much as possible.

Regression Model

Our plan is to start with the variables that are evidently must be there based on our knowlenge of the subject, Ttry to add othe variables (becides "am"“) and check if their presence improve the model and, finally, add the "am” variable to the model and see if it improves the model.

The common sense suggests that mpg depends on the weight of the car (wt variable) and its “power”: there are 2 replated varaible hp (Gross horsepower) and qsec (¼ mile time) and they are highly correclated. Comparing models shows that hp provides slightly better fit (in order to save space we only show the model, R-squared and Adjusted R-squared):

fit1 <- lm(mpg ~ wt, data=mtcars); R-squared:0.7528,Adjusted R-squared:0.7446

fit2a <- lm(mpg ~ wt+hp, data=mtcars); R**2: 0.8268, Adjusted R**2: 0.8148

fit2b <- lm(mpg ~ wt+qsec, data=mtcars); R**2: 0.8264, Adjusted R**2: 0.8144

fit3 <- lm(mpg ~ wt+hp+qsec, data=mtcars); 0.8348, Adjusted R-squared: 0.8171

anova(fit2a, fit3) shows P value 0.2546 so we decided to keep only hp variable

Now we are trying adding variables and keep monitoring values R-squared and Adjusted R-squared. Apparently adding the cyl varable (fit5 model) provides some improvement while adding other variables don't: adding disp resulted in R**2= 0.8268 and Adj R**2=0.8083; adding gear: 0.8356 / 0.8112; adding vs: 0.8329 /0.815; adding carb: 0.837 / 0.789; adding drat: 0.8369 /0.8194.

fit5 <- lm(mpg ~ wt+hp+as.factor(cyl), data=mtcars); R2: 0.8572, Adj R**2: 0.8361

anova(fit2a,fit5) gives P value 0.07364.

Now will add the “am” variable to be best moodel selected in the previous step

```
fit_final <- lm(mpg ~ wt + hp + as.factor(am) +
as.factor(cyl), data = mtcars)
```

Figure 2 in the Appendix describes the model in more detail.The resulted R-squared, 0.8659, this is very close to one with use of all available variables and Adjusted R-squared is the best we were able to achieve, 0.8401.

Figure 3 provides more information regarding residuals.

Conclusion

This model suggests that manual transmission leads to better milege than automatic one by 1.8 mpg (recall that am = 0 = automatic, 1 = manual). However the P-value is approxinately 20% so, based on the available data, we can not be even confident that manual transmission is really better even though it appears pretty likely. (80% confidence level). Evidently we need bigger data set to make a conclusion with condfidence (reach statistical significance).

Appendix

Figure 1. Relations between variables

```
pairs(x = mtcars[, names(mtcars)], panel = panel.smooth,
main = "Relations between MTCars Variables")
```

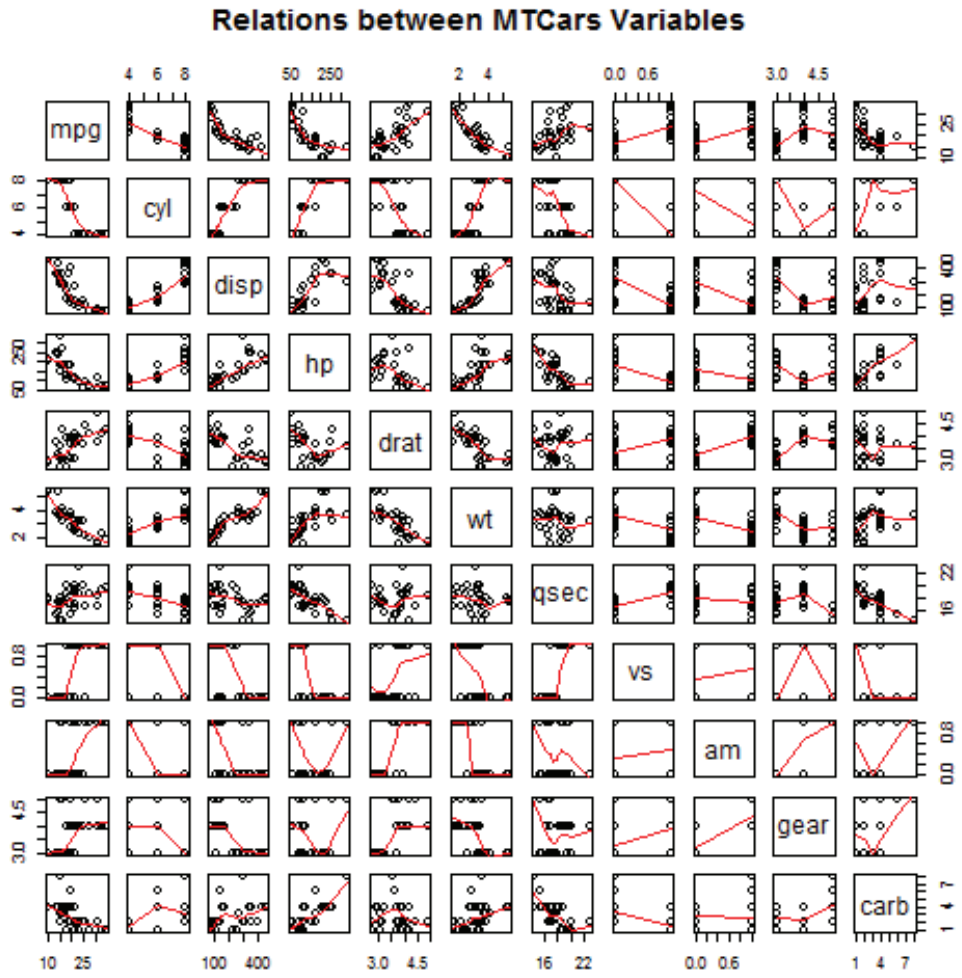


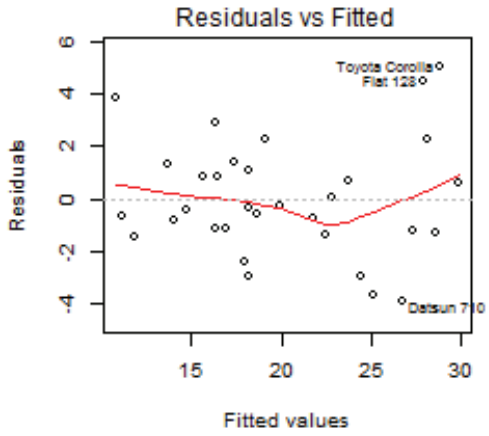
Figure 2. Final Model

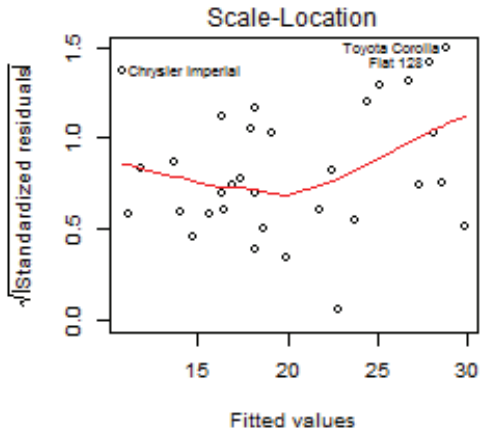
```
summary(fit_final)
```

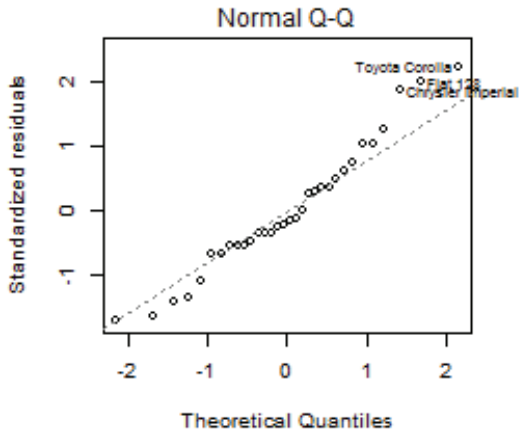
```
##
## Call:
## lm(formula = mpg ~ wt + hp + as.factor(am) +
as.factor(cyl),
##     data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.939 -1.256 -0.401  1.125  5.051
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      33.7083     2.6049   12.94  7.7e-13 ***
## wt               -2.4968     0.8856   -2.82   0.0091 **
## hp               -0.0321     0.0137   -2.35   0.0269 *
## as.factor(am)1    1.8092     1.3963    1.30   0.2065
## as.factor(cyl)6  -3.0313     1.4073   -2.15   0.0407 *
## as.factor(cyl)8  -2.1637     2.2843   -0.95   0.3523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.84
## F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

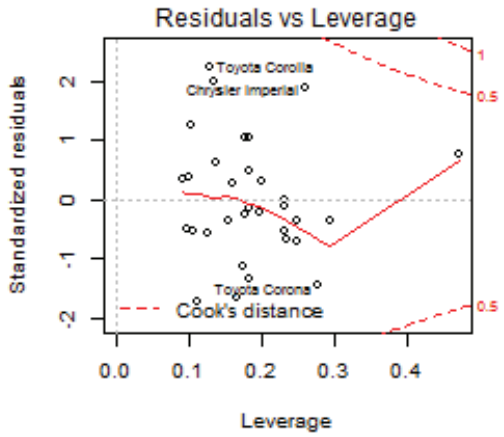
Figure 3. Residuals for Final Model

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


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710


Scale-Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128


Normal Q-Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial


Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance