Regression Models Project.

Sergey Rakhmatullin

Executive Summary

This report is the analys of `mtcars` dataset performed to find out relationship between MPG and a bunch of other variables of dataset and fit a regression model. Another report objective is to answer the question whether an automatic or manual transmission better for MPG and give a statistical evidence.

Exploratory Data analysis

First step is to analyse dependance of MPG on am (tranmission). On Plot 1. in Appendix 1. you can find boxplot that shows that cars with automatic transmissions tend to have less MPG.
Next we are going to do quick variable analysis using pairs() function. See Plot 2. in Appendix 1. We excluded factor variables to make plot more readable. We can conclude that all of this variations have significant correlation with MPG and could be included in our model as regressors.

Model Selection Strategy

In this paragraph we are using both 'Step-by-step' and automated strategies to obtain similar results. Step-by-step strategy is placed to Appendix 2 to make the main report concise.
As an automated method we use AIC algrorithm with both forward selection and backward elimination.

```r
full<-lm(mpg~., mtcars)
best<-step(full, direction = 'both', trace=0)
summary(best)$call
```

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
```

As we mentioned before the correlation between variables in `mtcars` dataset is significant so we can possibly upgrade our model with adding interaction between highly correlated regressors.
First step, lets find correlations

```r
corr<-cbind(cor(mtcars$wt,mtcars$qsec), cor(mtcars$wt,mtcars$am), cor(mtcars$am,mtcars$qsec))
colnames(corr)<-c('wt~qsec','wt~am','am~qsec')
corr
```

```
##          wt~qsec      wt~am    am~qsec
## [1,] -0.1747159 -0.6924953 -0.2298609
```

Second, add the interaction term `wt:am` and compare with our 'best' model, also compare with the model that has `am` term excluded.

```r
fit1<-lm(mpg ~ wt + qsec, mtcars)
fit3<-lm(mpg ~ wt + qsec + am + wt:am, mtcars)
anova(fit1, best, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + qsec
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ wt + qsec + am + wt:am
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     29 195.46
## 2     28 169.29  1    26.178  6.0268 0.020819 *
## 3     27 117.28  1    52.010 11.9740 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So we can see that adding `am` term and `wt:am` interaction term have good impact on RSS.

Residual Analysis

Please refer to Appendix 1 to see Residual plots.
Left plots (Residuals vs. Fitted, Scale-Location) have randomly distributed residuals and shows no consistent pattern. Normal Q-Q plot shows that our residuals are normally distributed and Residuals vs. Leverage plot shows no obvious outliers. So we can interpretate this results that residuals for our model are statistical errors with close to normal distribution no dependacy from variables.

Error Estimation

Also lets perform t-test for groups with different transmission

```
t.test(mpg~am, mtcars)
```

```
##
##  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
```

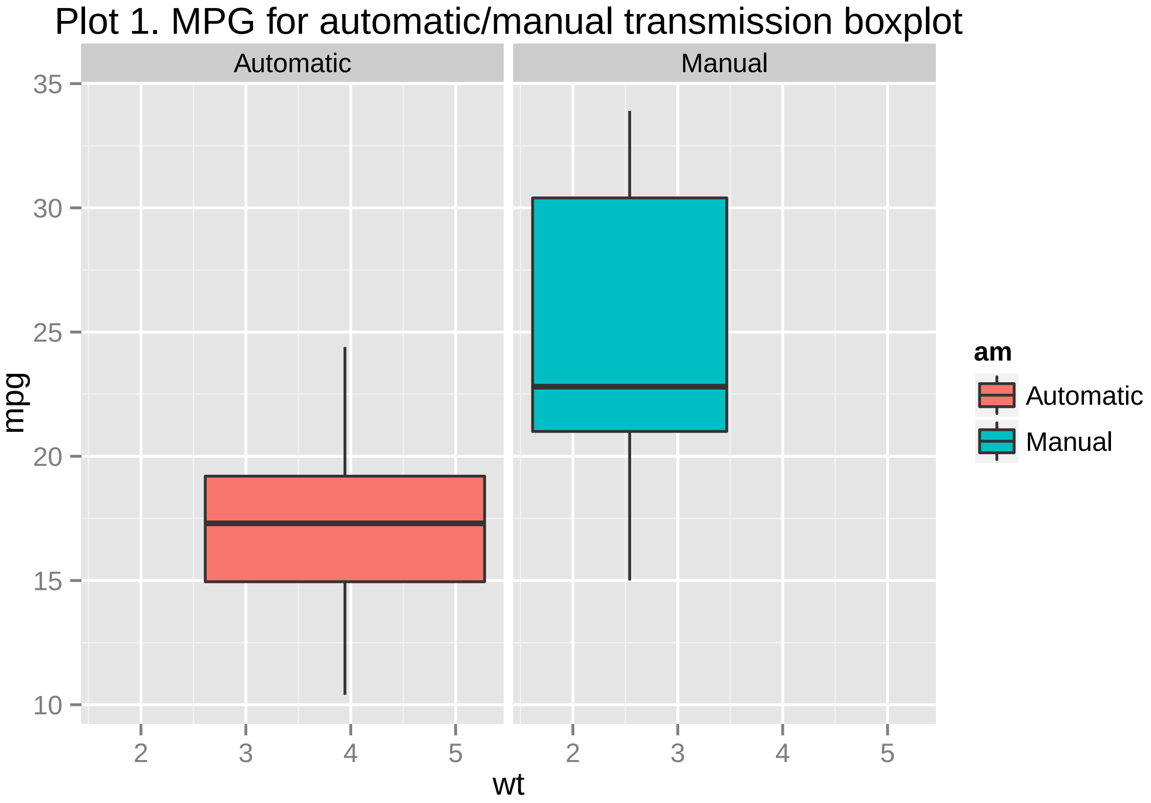
We have to reject the null hypothesis that we have no difference in means for automated and manual transmission.

Conclusions

We could see from previous study that transmission `am` variable:
1) can be used as a regressor for linear model fitting and thus have 'linear' influence on the MPG
2) highly correlated with `wt` regressor so one should be aware and take interaction into account.
Also, we can clearly observe and statistically confirmed that mean and confidence interval of group with automatic transmission is lower in MPG that these of the group with manual transmission if other factors are neglected. So we can conclude that cars with manual transmittion tend to be better for MPG.

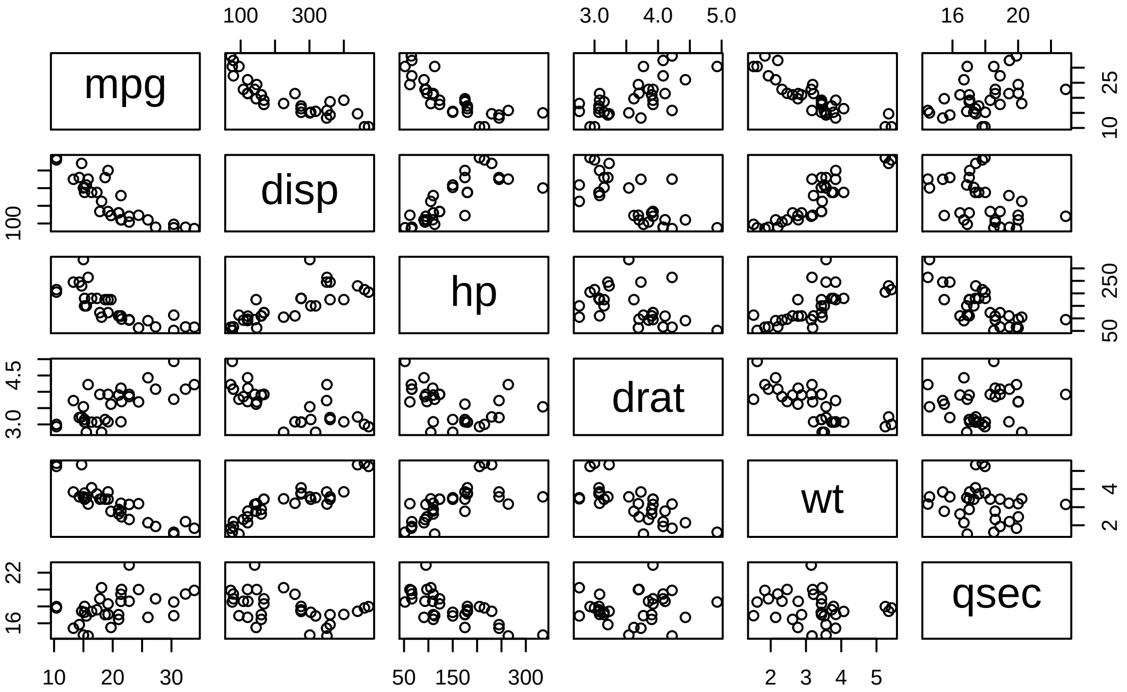
Appendix 1. Exploratory data plots

```
library(ggplot2)
require(datasets)
mtcars$am<-factor(mtcars$am)
levels(mtcars$am) <-c('Automatic','Manual')
d1<-ggplot(mtcars, aes(y=mpg, x=wt, group=am))
d1<-d1+geom_boxplot(aes(fill=am)) +facet_grid(.~am)
d1 + ggtitle('Plot 1. MPG for automatic/manual transmission boxplot')
```

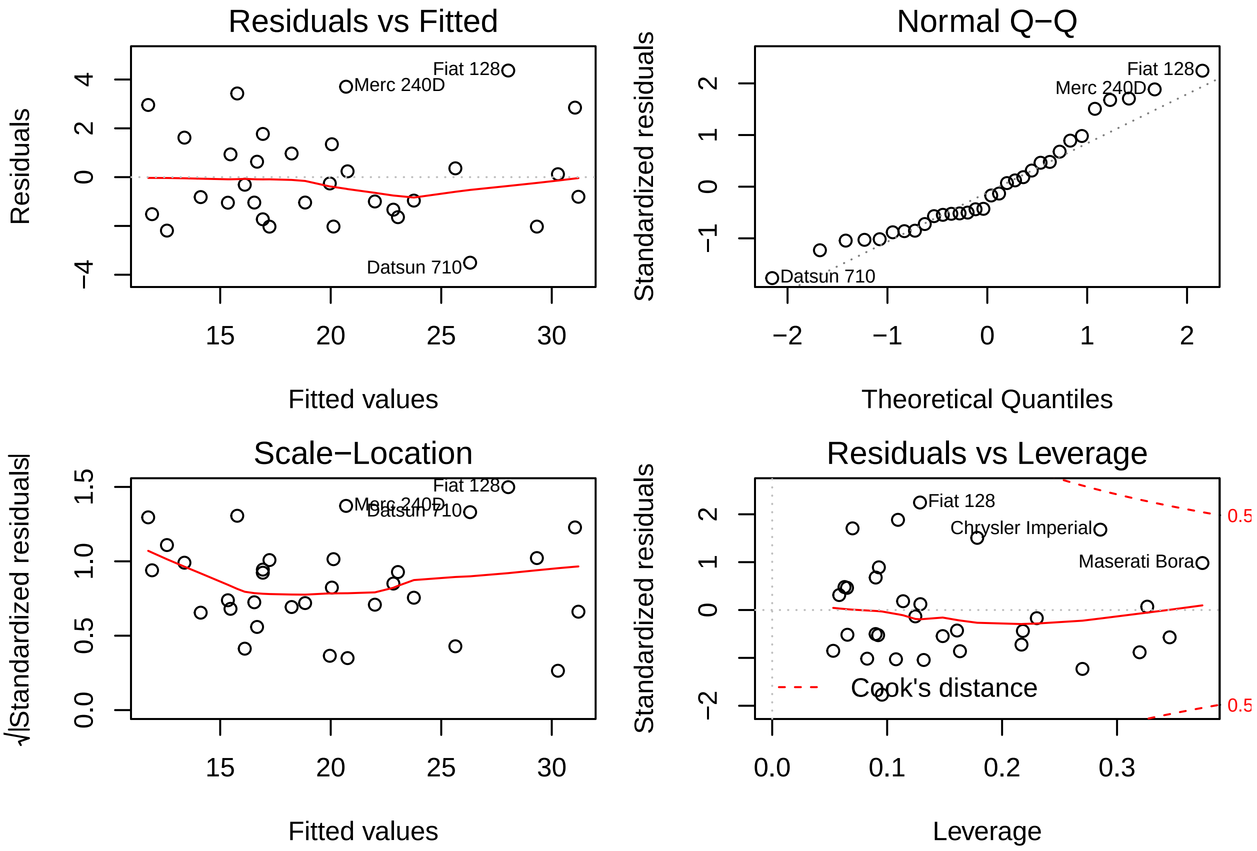


```
pairs(mtcars[c(1,3,4,5,6,7)], main='Plot 2. Variable correlation')
```

Plot 2. Variable correlation



```r
par(mfrow=c(2,2), mar=c(4,4,2,1))
plot(fit3)
```



Appendix 2. Step-by-step model selection

1. Fitting the highest correlated with mpg regressor

Lets find the most correlated with MPG variable:

```
remove(mtcars)
require(datasets)
cr<-c()
for (i in 1:dim(mtcars)[2]) cr<-cbind(cr, cor(mtcars$mpg, mtcars[,i]))
colnames(cr)<-names(mtcars)
cr
```

```
##      mpg       cyl       disp         hp      drat         wt     qsec
## [1,]   1 -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684
##             vs        am      gear       carb
## [1,] 0.6640389 0.5998324 0.4802848 -0.5509251
```

```
fit0<-lm(mpg ~ wt, mtcars)
```

This is the `wt` variable, so we fit this variable as the first regressor to our model fit0.

2. Adding the highest correlated with residuals to fit0 regressor

The next step we exclude `wt` influence by taking residuals to model fit0. The next regressor will be highest correlated variable to this residuals

```
cr<-c()
for (i in 1:dim(mtcars)[2]) cr<-cbind(cr, cor(residuals(fit0), mtcars[,i]))
colnames(cr)<-names(mtcars)
cr
```

```
##            mpg        cyl       disp         hp      drat            wt
## [1,] 0.4971591 -0.3484239 -0.1550555 -0.4115374 0.1267524 -7.631868e-17
##           qsec        vs           am        gear       carb
## [1,] 0.5372327 0.3672086 -0.002046779 -0.05191431 -0.3618736
```

```
fit1<-lm(mpg ~ wt + qsec, mtcars)
```

This is the `qsec` variable. Lets use it as second regressor and make nested model fit1.

3. Adding the highest correlated with residuals to fit1 regressor

The next step we exclude both `wt` and `qsec` influence by taking residuals to model fit1. The next regressor will be highest correlated variable to this residuals

```
cr<-c()
for (i in 1:dim(mtcars)[2]) cr<-cbind(cr, cor(residuals(fit1), mtcars[,i]))
colnames(cr)<-names(mtcars)
cr
```

```
##           mpg        cyl          disp          hp      drat            wt
## [1,] 0.416634 -0.1152085 -0.0008298392 -0.09886947 0.1732503 -3.274668e-17
##               qsec          vs        am      gear        carb
## [1,] -2.234436e-17 0.009965737 0.2295583 0.1460763 -0.04727076
```

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

Finally we have `am` variable. Lets use it as third regressor and make nested model fit2. We could proceed this strategy implementation but we know that most of our variables highly correlated and we have a risk of having standard error inflation sooner or later. Of course we always could test both RSS and standard error every step but R have good automated algorithm for this task. Now the good time to analyze results and compare it with some of the machine learning algorithm.

4. Comparing nested models

So regressors that we found is `wt`, `qsec` and `am`. To find out whether they are significant to our model lets compare nested models, subsequently adding variables.

```
anova(fit0,fit1,fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + qsec
## Model 3: mpg ~ wt + qsec + am
## Model 4: mpg ~ wt + qsec + am + wt:am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 278.32
## 2     29 195.46  1    82.858 19.0761 0.0001663 ***
## 3     28 169.29  1    26.178  6.0268 0.0208187 *
## 4     27 117.28  1    52.010 11.9740 0.0018086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
rbind(summary(fit0)$coef[2,],
      summary(fit1)$coef[2,],
      summary(fit2)$coef[2,],
      summary(fit3)$coef[2,])
```

```
##       Estimate Std. Error    t value     Pr(>|t|)
## [1,] -5.344472  0.5591010  -9.559044 1.293959e-10
## [2,] -5.047982  0.4839974 -10.429771 2.518948e-11
## [3,] -3.916504  0.7112016  -5.506882 6.952711e-06
## [4,] -2.936531  0.6660253  -4.409038 1.488947e-04
```

We can see that RSS is reduced by adding variables (wt, qsec, am) to our model. Standard error is higher for fit2 because of the correlation of ‘wt’ and ‘am’.
So the conclusion we can make so far: All three variables is significant predictors for the chosen model.