Regression Models Project

Galina Malovichko

06/18/2015

Summary

In this report, we analyze `mtcars` data set and use different variables to build linear regression model of miles per gallon (MPG). We are mostly interested in how automatic (`am = 0`) and manual (`am = 1`) transmissions features affect the `mpg` feature. Best model aquired includes `am`, `wt` and `qsec` as variables and accounts for 88% of MPG variance.

Exploratory Data Analysis

First, we load the data set `mtcars` and convert some variables into `factor` class.

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

Then we create a box plot of MPG values depending on the transmission type (see Figure 1 in the Appendix). At the first glance it seems that manual transmission yields higher MPGs. Pair plot (Figure 2) shows that some other variables, such as `cyl`, `disp`, `hp`, `drat`, `wp` and `vs` strongly correlate with MPG, so might also be predictors.

Regression Models

For the first model, we neglect all variables except transmission and build a linear model of MPG agains this single variable.

```r
model.one <- lm(mpg ~ am, data = mtcars)
summary(model.one)
```

Transmission accounts for 34% of MPG variance. The intercept and slope coefficients show that automatic transmissions has average MPG of 17.147, and that manual transmissions have average MPG $17.147 + 7.245 = 24.39$. As expected, these values match means in the ox plot from Figure 1. Both coefficient have significance 0.001. Residual standard error is 4.902 on 30 degrees of freedom.

Next naive model is a linear regression of `mpg` vs all the other variables.

```r
model.all <- lm(mpg ~ ., data = mtcars)
summary(model.all)
```

This model accounts for 78% of MPG variance. Residual standard error is 2.833 on 15 degrees of freedom, but none of the coefficients have significance level of 0.05 or higher.

Now we want to use backward selection to select only statistically significant variables.

```
model.step <- step(model.all, k = log(nrow(mtcars)))
summary(model.step)
```

This model has formula `mpg ~ wt + qsec + am`. Its residual standard error is 2.459 on 28 degrees of freedom and it explains about 83% of MPG variance. All of the slope coefficients are significant at 0.05 or 0.001 level.

Going back to pair plots in Figure 2 from the appendix, we can inspect the correlations between these three significant variables. Pairs `wt` and `qsec`, `qsec` and `am` seem to be uncorrelated, but there is a relationship between `wt` and `am`: automatic cars tend to be heavier on average. To improve our model, we will add the interaction term between these two variables:

```
model.best <- lm(mpg ~ wt + qsec + am + wt:am, data = mtcars)
summary(model.best)
```

This model accounts for 88% of MPG variance and has the residual standard error as 2.084 on 27 degrees of freedom. All of the coefficients are significant at 0.05 significant level. This model seems best so far.

Here is model comparison with `anova`.

```
anova(model.one, model.all, model.step, model.best)
```

Model `mpg ~ wt + qsec + am + wt:am` has the least resiual sum of squares.

```
summary(model.best)$coef
```

```
##              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
```

From the model coefficients we can see that if `wt` (weight) and `qsec` (1/4 mile time) are kept constant, manual transmission car yields `14.08 - 4.14 * wt` higher MPG than automatic car of the same `wt` and `qsec`. Unfortunately, confidence intervals of these coefficients are quite wide, so we are not very certain of their values:

```
confint(model.best)
```

```
##                  2.5 %    97.5 %
## (Intercept) -2.3807791 21.826884
## wt          -4.3031019 -1.569960
## qsec         0.4998811  1.534066
## am1          7.0308746 21.127981
## wt:am1      -6.5970316 -1.685721
```

Residual Analysis

Residual plots for the best model are given in Figure 3 from the Appendix.
1. The Residuals vs. Fitted plot shows no pattern, i.e. residuals and variable values are independent.
2. The points in Normal Q-Q plot lie closely to the line, i.e. residuals are normally distributed.
3. The Scale-Location line is roughly horizontal, i.e. variances and variable valuesis are independent.
4. All Residuals vs. Leverage points lie within the 0.5 bands, i.e. there are no suspiciour outliers.

Appendix

Figure 1. MPG vs. transmission

```
boxplot(mpg ~ am, xlab = "Transmission (0 = automatic, 1 = manual)", ylab = "MPG",
        main = "MPG vs. transmission")
```

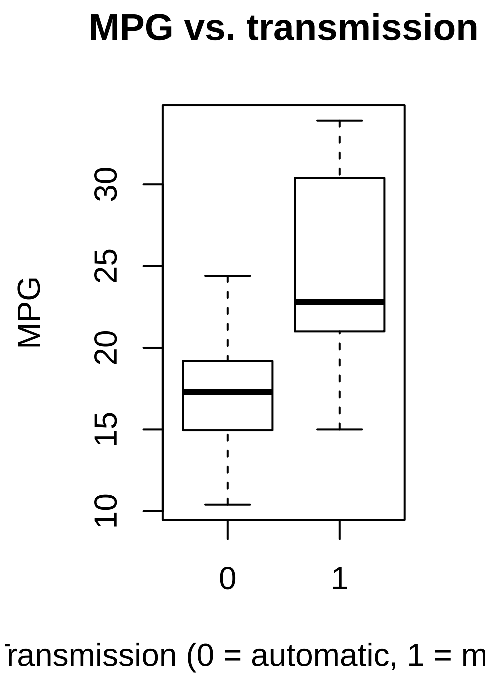


Figure 2. Pair plots of mtcars variables

```
pairs(mtcars, panel = panel.smooth, main = "Pair plots of mtcars variables")
```

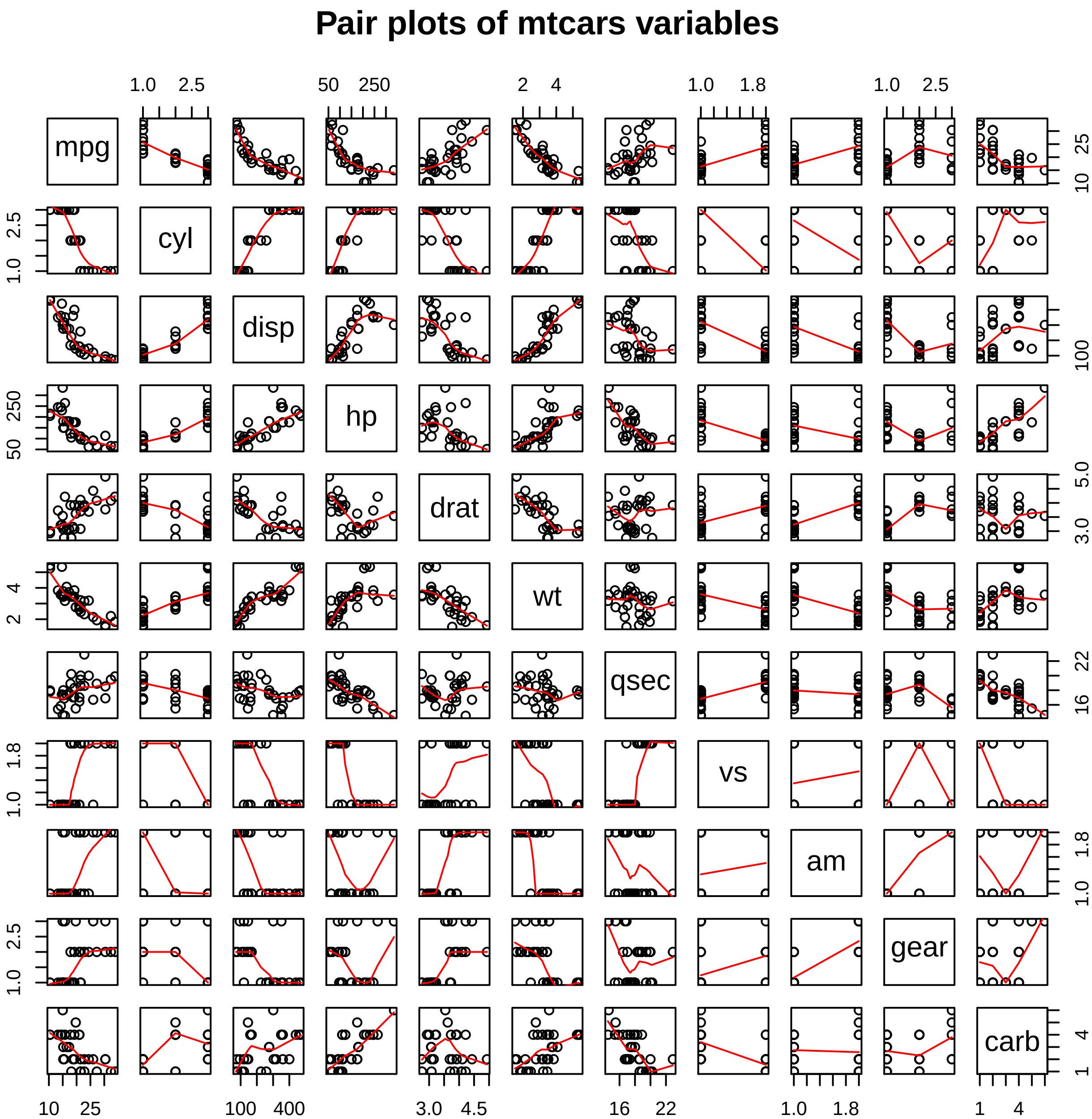
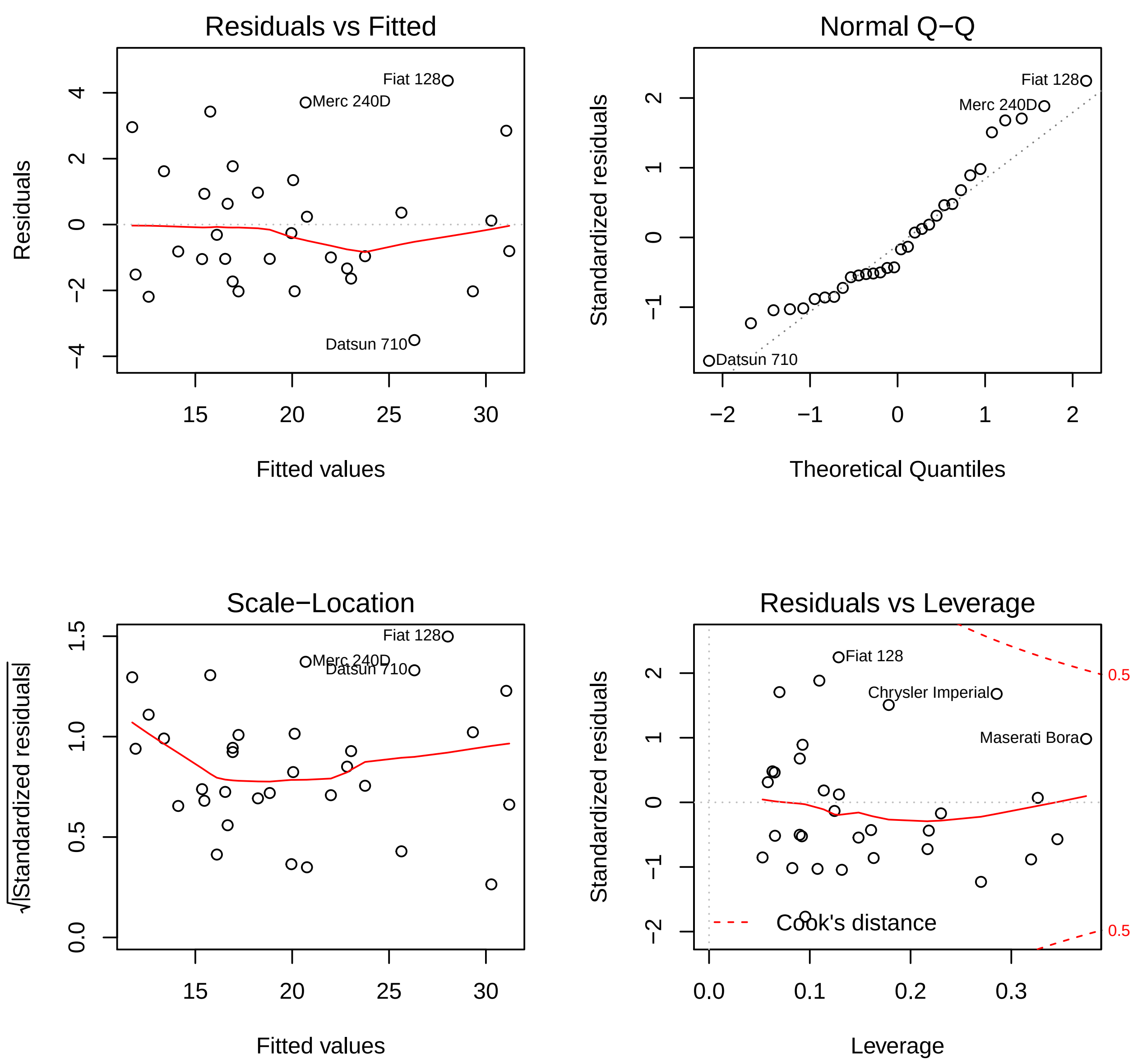


Figure 3. Residual Plots

```r
par(mfrow = c(2, 2))
plot(model.best)
```


Residuals vs Fitted
Fiat 128
Merc 240D
Datsun 710
Residuals
Fitted values
Normal Q−Q
Fiat 128
Merc 240D
Datsun 710
Standardized residuals
Theoretical Quantiles
Scale−Location
Fiat 128
Merc 240D
Datsun 710
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5
Standardized residuals
Leverage