Regression Models Course Project

Ivan Astafiev

April 3, 2016

Executive Summary

Dataset from the 1974 Motor Trend US magazine was explored to understand a relationship between transmission type (automatic/manual) and fuel usage in MilesPerGalon (MPG).

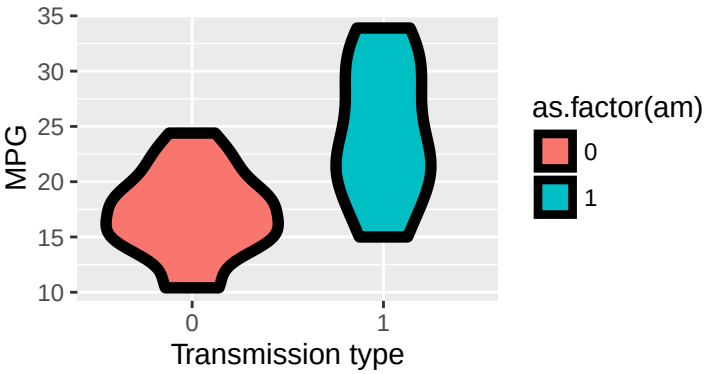
The main questions was to distinguish which transmission type is more fuel efficient and try to calculate this difference.

Regression Linear model was created which include MPG as an outcome and Weight, Horse Power, Number of cylinders and Transmission Type as a predictors.

According to regression model results, manual transmission shows better fuel efficiency than automatic gear. Usage of Manual transmission compared to Automatic brings 1.8 miles increase in distance driven per 1 Gallon (MPG).

However, p-value of a Transmission factor in final model equaled to 0.21. This fact do not allow to reject null hypothesis that type of transmission has a statistical significant influence on a MPG. More data and deeper research is necessary.

Simple Mean analisys



Hypothesis test to check that difference in MPG means for automatic and manual transmissions statistical significant.

```
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
```

T-test results on simple means brings p-value at 0.001374. This is below 0.05 and confirms alternative hypothesis that difference in means statistical significant. But according to dataset analysis there can be present other confounding variables which can be evaluated in Regression model analysis.

Regression Model analisys

Multivariable Regression Model is selected as a method to answer research questions. Several models was fitted and optimal model was selected using a step-by-step method.

Regression model analysis allows to fix several confounding variables which has a big impact on MPG. These variables are: Weight, Horse Power and number of cylinders.

Model selection process presented at **Annex 1.**

Final Model

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

Model predicting force: Adjusted R-squared = *0.8401*

Residual Plot analysis shows that there not clear dependency visible on a residual plot. (see Annex 2. Residuals for more details)

Coefficients Interpretation

```
round(summary(final_fit)$coef,2)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        33.71       2.60   12.94     0.00
## am                  1.81       1.40    1.30     0.21
## as.factor(cyl)6    -3.03       1.41   -2.15     0.04
## as.factor(cyl)8    -2.16       2.28   -0.95     0.35
## wt                 -2.50       0.89   -2.82     0.01
## hp                 -0.03       0.01   -2.35     0.03
```

- Usage of Manual Transmission cause a 1.8 miles increase in distance driver per 1 gallon

However, P-value eqaled to 0.21 do not allow to reject null hypothesis that type of transmission has a statistical significant influence on a MPG. More data and deeper research is necessary.

Research answers

Question 1 - Is an automatic or manual transmission better for MPG?

ANSWER = Manual transmission is better for MPG

Question 2 - Calculate MPG difference between automatic and manual transmissions

ANSWER = Manual transmission brings 1.8 miles increase in distance driven per 1 gallon (MPG)

Annex 1. Model Selection

Multivariable Regression Model is selected as a method to answer research questions.

Following process selected:

- Create a several models with step by step adding variables
- Compare models using Anova function to detect valuable variables
- Constract final model with the most dependent variables

Variance and R-squared parameters used for selection of best model.

Base model

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

Model Selection

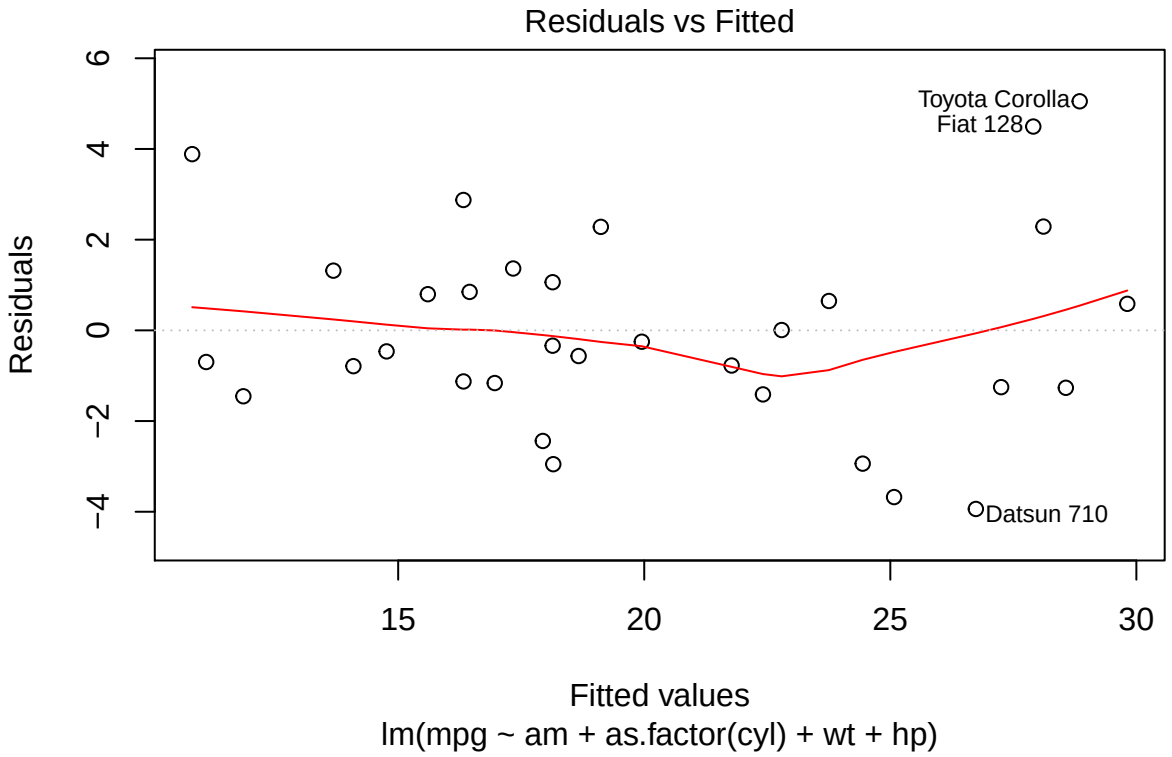
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + hp
## Model 5: mpg ~ am + cyl + disp + hp + wt
## Model 6: mpg ~ am + cyl + disp + hp + wt + qsec
## Model 7: mpg ~ am + cyl + disp + hp + wt + qsec + as.factor(vs)
## Model 8: mpg ~ am + cyl + disp + hp + wt + qsec + as.factor(vs) + as.factor(gear)
## Model 9: mpg ~ am + cyl + disp + hp + wt + qsec + as.factor(vs) + as.factor(gear) +
##     as.factor(carb)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 57.8111 7.243e-07 ***
## 3     28 252.08  1     19.28  2.4796   0.13376
## 4     27 216.37  1     35.71  4.5929   0.04687 *
## 5     26 163.12  1     53.25  6.8477   0.01804 *
## 6     25 150.99  1     12.13  1.5598   0.22862
## 7     24 150.76  1      0.23  0.0299   0.86472
## 8     22 149.21  2      1.55  0.0997   0.90567
## 9     17 132.19  5     17.02  0.4377   0.81609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova results show that according to P value the following variables should be included in final model because of confirmed statistical inference: wt, cyl, hp

Final model

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

Residuals



Residuals spread on a plot in general without any clear dependency. However there are some dependency in a line which can be caused by outfitted values of Toyota Corolla and Fiat 128. This must be further investigated