Regression Models Course Project

Ivan Astafiev

April 3, 2016

Executive Summary

Motor Trend dataset 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 calcuate this difference. Regression Linear model was created which inslude MPG as an outcome and Weight, HorsePower, Number of cylinders and Transmission Type as a predictors.

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 for Transmission type predictor equals only to 0.21, what do not allow us to reject null hypothesis. So it can not be said with confidence that transmission type has a statistical dependency to fuel efficiency expressed in (MPG)

DataSet exploration

Data Set: mtcars Detailed data description presented at Annex 1.

We interested in relationship of **mpg** (Miles Per Gallon) and **am** (Transmission (0 = automatic, 1 = manual).

Modeling

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

Model selection process presented at **Annex 2.**

Final Model

Variable	Type	Description
am	factor	Transmission type (0=automatic, 1 = manual)
wt	numeric	Weight (1000 lbs)
cyl	factor	Number of cylinders (to be included as Factor of 4,6,8 cylinder)

hp	numeric	Gross horsepower

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

Model predicting force: Adjusted R-squared = *0.8401* Residual Plot shows that there not clear dependencey visible. (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
```

- Increase of weight for 1000 lb cause a 2.5 miles decrease in distance driven per 1 gallon
- Usage of Manual Transmission cause a 1.8 miles increase in distance driver per 1 gallon
- Usage of 6cyl engine cause a 3 miles decrease in distance driven per 1 gallon
- Usage of 8cyl engine cause a 2.16 miles decrease in distance driven per 1 gallon
- P value for Transmission type do not allow to reject null hypothesis(H0), so we cant guarantee that automatic/manual transmission has a statistical dependable influence on fuel usage

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)

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

Anex 1. MTCARS DATA Descritpion

Description

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).

Format

A data frame with 32 observations on 11 variables.

- mpg Miles/(US) gallon
- cyl Number of cylinders
- disp Displacement (cu.in.)
- hp Gross horsepower
- drat Rear axle ratio
- wt Weight (1000 lbs)
- qsec 1/4 mile time
- vs V/S
- am Transmission (0 = automatic, 1 = manual)
- gear Number of forward gears
- carb Number of carburetors

Annex 2. 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

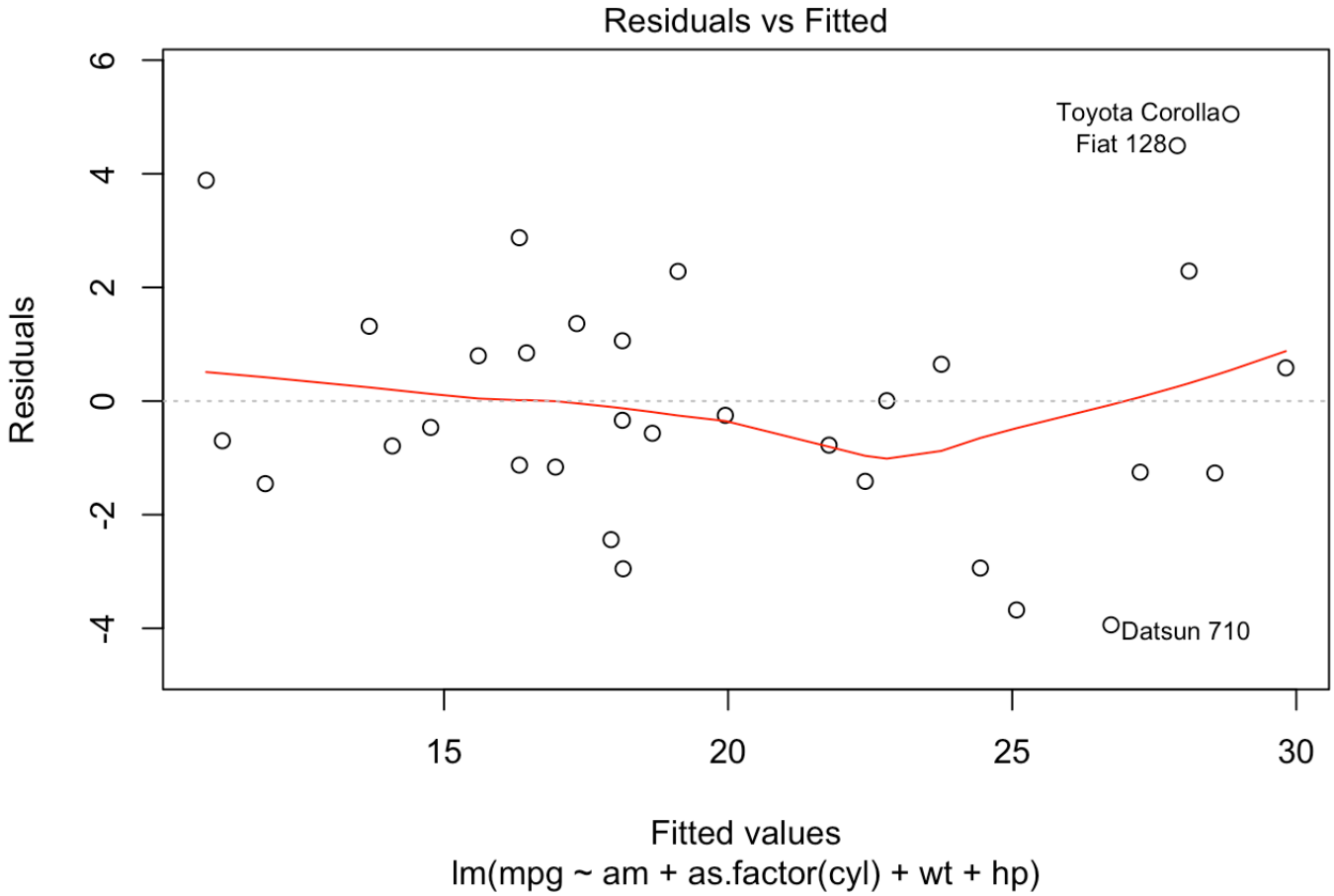
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + hp
## Model 5: mpg ~ am + cyl + disp + hp + wt
## Model 6: mpg ~ am + cyl + disp + hp + wt + qsec
## Model 7: mpg ~ am + cyl + disp + hp + wt + qsec + 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