Motor Trend Analysis

Transmission type impact on car fuel consumption

by Vajo Lukic

Executive Summary

The goal of this analysis is to explore the effect of transmission type (automatic and manual) on a car fuel consumption. This analysis will provide answers to following questions: - Is an automatic or manual transmission better for MPG? - How different is the MPG between automatic and manual transmissions?

Conclusion: a simple analysis showed that the type of transmission is significant for the fuel consumption (mpg), where manual transmission had on average 7.24 mpg higher consumption than automatic. Futher analysis showed different picture, where transmission type was not significant variable anymore, and some other variables like: weight, horsepower and number of cilinders turned out to be more significant than transmission. In this final model, manual transmission contributes 1.81 mpg to higher consumption, with assumption that all other variables are equal to zero.

Uncertanty and limitations

It is important to emphasize that analyzed sample was quite small with only 32 observations, which has additional impact on further statistical inference. It was also almost imposible to fit this analysis on only 2 pages, as requested.

Exploratory Data Analysis

The data we are about to analyze 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). Here are the steps we have performed to process the data:

Load the data set and make some variables names a bit more descriptive:

```
data(mtcars)
names(mtcars) <- c("mpg", "cylinders", "displacement", "horsepower", "axleratio", "weight",
"qmiletime","vs","transmission", "gears","carburetors")
```

Convert some variables to factors and Change the names of values for variable “transmission”. This will make further analysis easier. Finally, display the structure of transformed data set:

```
mtcars$transmission <- factor(mtcars$transmission)
mtcars$cylinders <- factor(mtcars$cylinders)
mtcars$gears <- factor(mtcars$gears)
mtcars$carburetors <- factor(mtcars$carburetors)
mtcars$vs <- factor(mtcars$vs)
levels(mtcars$transmission) <- c("automatic", "manual")
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg         : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cylinders   : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ displacement: num  160 160 108 258 360 ...
```

```
##  $ horsepower  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ axleratio   : num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ weight      : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qmiletime   : num  16.5 17 18.6 19.4 17 ...
##  $ vs          : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ transmission: Factor w/ 2 levels "automatic","manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gears       : Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carburetors : Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Testing for normality

We want to know if our sample for "mpg" (fuel consumption) is from normaly distributed population. The test result depends on p-value. When $p < 0.05$, then population is likely not normaly distributed. When $p > 0.05$ there is no such evidence.

```
shapiro.test(mtcars$mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.9476, p-value = 0.1229
```

In our case, large p-value indicates that population is likely to be normaly distributed.

Comparing the means

We want to know if there is any difference at all, in fuel consumption for two types of transmision. Again, p-value will provide an answer. $p < 0.05$ indicates that means are likely different. $p > 0.05$ provides no such evidence.

```
t.test(mtcars$mpg ~ mtcars$transmission)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$transmission
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
## mean in group automatic    mean in group manual
##                   17.15                   24.39
```

Low p-value (less than 0.05) indicates difference for these two groups. Also the numbers representing the mean fuel consumption for manual and automatic type clearly are diffrent. See "Diagram 1." in appendix for further analysis of means comparison.

Model Selection

Selecting the best model

In order to select the best model, we need to find out which variables have biggest impact on fuel consumption, beside transmission type. We will use "Backward stepwise regression", which starts with all predictors and removes those which are not statistically significant.

```
full.model <- lm(mpg ~ ., data = mtcars)
reduced.model <- step(full.model, direction="backward", k=2, trace=0)
summary(reduced.model)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + horsepower + weight + transmission,
##     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 ***
## cylinders6          -3.0313     1.4073   -2.15   0.0407 *
## cylinders8          -2.1637     2.2843   -0.95   0.3523
## horsepower          -0.0321     0.0137   -2.35   0.0269 *
## weight              -2.4968     0.8856   -2.82   0.0091 **
## transmissionmanual   1.8092     1.3963    1.30   0.2065
## ---
## 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
```

Let's analyse the summary and see if this model is statistically significant. We'll start from the end because there we can find the most important statistics: - Model's p-value of less than 0.05 also indicates that this model likely is significant - R-squared as a measure of model's quality, represents a fraction of outcome's variance explained by the model. In this case the model explains 0.8401 (adjusteed value) or 84.01% of variance – Model is significant if any of coeficientts are non-zero. Clearly this is true, therefore this model is significant. - The model is suggesting "weight", "horsepower" and "cylinder" as significant variables. - However, transmission is marked as not significan in this model

Let's test significance of suggested model (now renamed to "fit.model") comparing it with the basic model:

```
basic.model <- lm(mpg ~ transmission, data = mtcars)
fit.model <- lm(mpg ~ cylinders + horsepower + weight + transmission, data = mtcars)
anova(basic.model, fit.model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ transmission
## Model 2: mpg ~ cylinders + horsepower + weight + transmission
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     26 151  4       570 24.5 1.7e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analysis of variance (ANOVA) resulted in p-value much lower than 0.05, which indicates that "fit.model" is statisticaly significant.

Residuals analysis

Points in "Residuals vs Fitted" (Diagram 2. in appendix) are randomly scattered with no obvious pattern. Points in Q-Q plot are mor-or-less on the line, indicating that residuals are normaly distributed.

Appendix

Diagram 1. - Boxplot displaying mpg by transmission type

```
boxplot(mpg ~ transmission, data = mtcars, xlab = "Transmission type", ylab = "Miles per gallon")
```

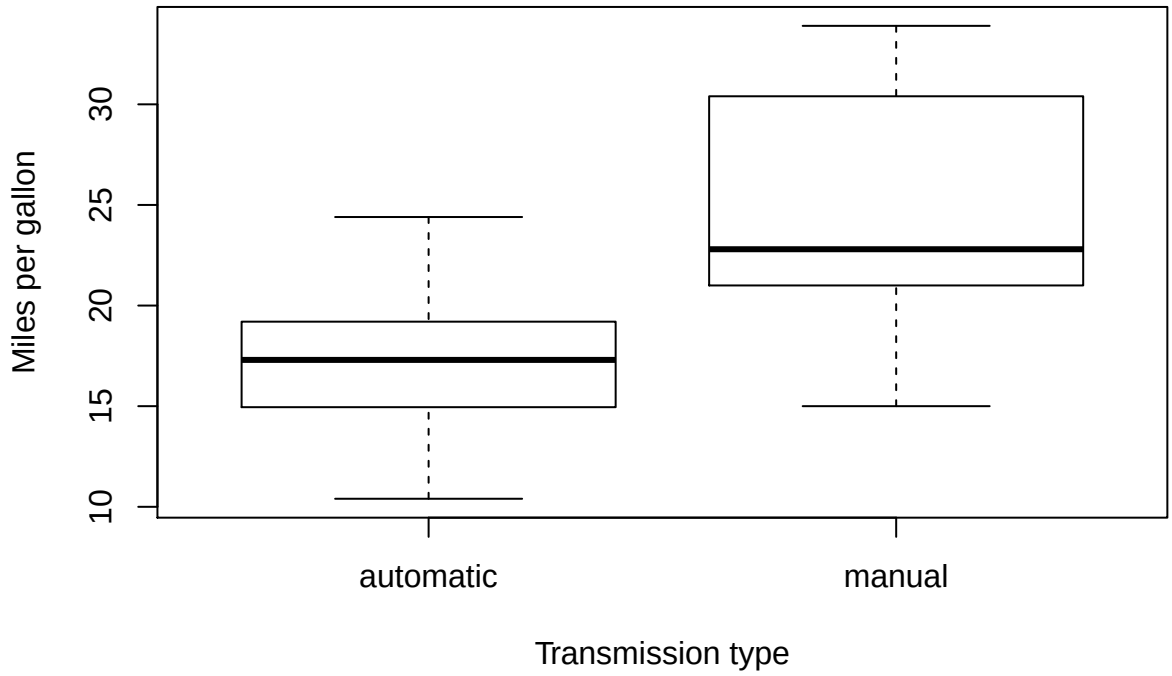
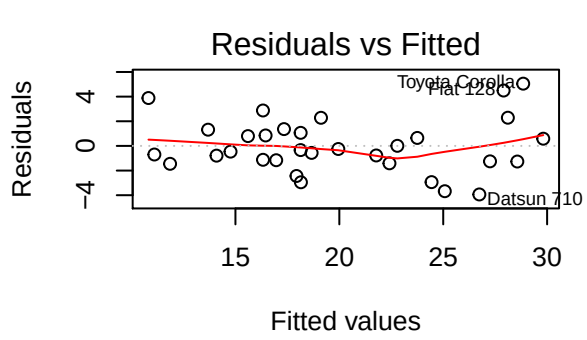
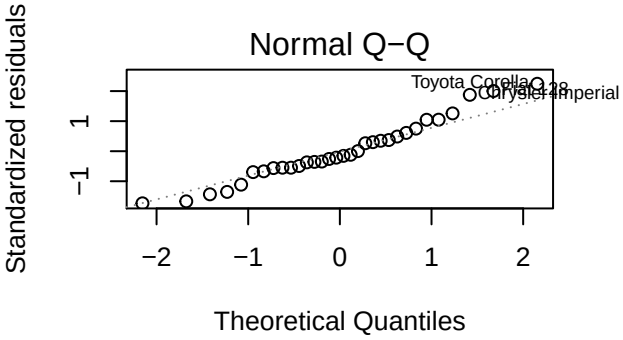
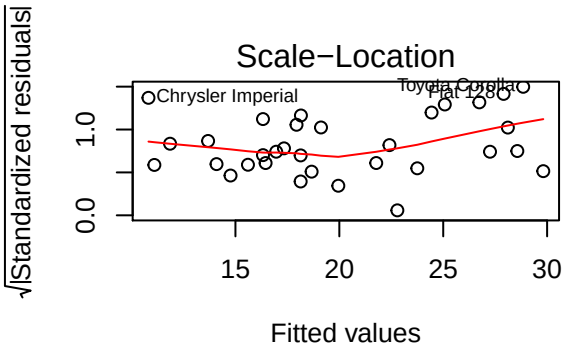


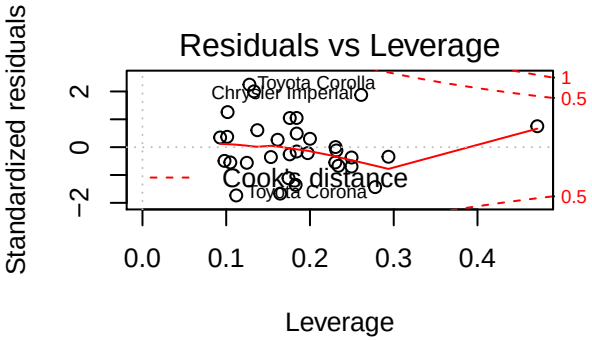
Diagram 2. - Plot for residuals of the "fit.model" - suggested by the regression analysis

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


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
15
20
25
30
−4
0
4


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Chrysler Imperial
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
15
20
25
30
0.0
1.0


Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
0.0
0.1
0.2
0.3
0.4
−2
0
2