Automobile transmission type impact on fuel efficiency

Pawel Daniluk

2016-12-14

Summary

This report is a summary of regression analysis, for the Regression Model Course. The goal of the analysis is to verify and quantify the impact of transmission type on fuel efficiency, measured as MPG.

The analysis found, that initial evidence of significant difference in fuel economy between two types of transmission did not held under further investigation.

The obtained results suggests, that under standard confidence level of 5%, transmission type has statistically insignificant impact on MPG. The difference in fuel efficiency between two types of transmission are results of impact of the other variables.

This report was written using Vim editor and Nvim-R plugin, supporting `rmarkdown`.

The original markdown file for this report can be found in the github repository.

Data description

The data for analysis is well know set of 32 automobiles, comparing their fuel economy and various aspects of their design. The data was extracted from 1974 Motor Trend US magazine. The detailed description of the data set can be found in R Documentation.

The research problem is to verify, weather the transmission type has significant impact on fuel consumption.

However, the fuel consumption may be codependent from various other factors such as vehicle weight and displacement, therefore complete analysis will take into account other significant variables.

All of the variables are recorded as numeric, while some of them seems to be categorical, rather than continuous. `R` requires, categorical variables to be converted to `factor` data type. This will allow for correct plotting and modeling of those variables.

The `dplyr` package is used for converting and selecting variables. Some variables were omitted, based on their low correlation with MPG in order to conserve report space.

```
mtc2 <- dplyr::select(mtcars, mpg, cyl, disp, hp, wt, am) %>%
    dplyr::mutate(cyl=factor(cyl), am=factor(am))
```

Model assumptions verification

Selection of appropriate statistical tool set for analysis depends on distribution of dependant variable. The linear regression is a standard choice, when exploring the impact of intercorrelated variables.

One of the principal requirement of linear regression is that dependant variable is normally distributed. The distribution can be assessed with help of Shapiro-Wilk Test for Normality.

In addition, the Shapiro-Wilk test for normality was performed. The null hypothesis of the test is that the variable follows a normal distribution.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.94756, p-value = 0.1229
```

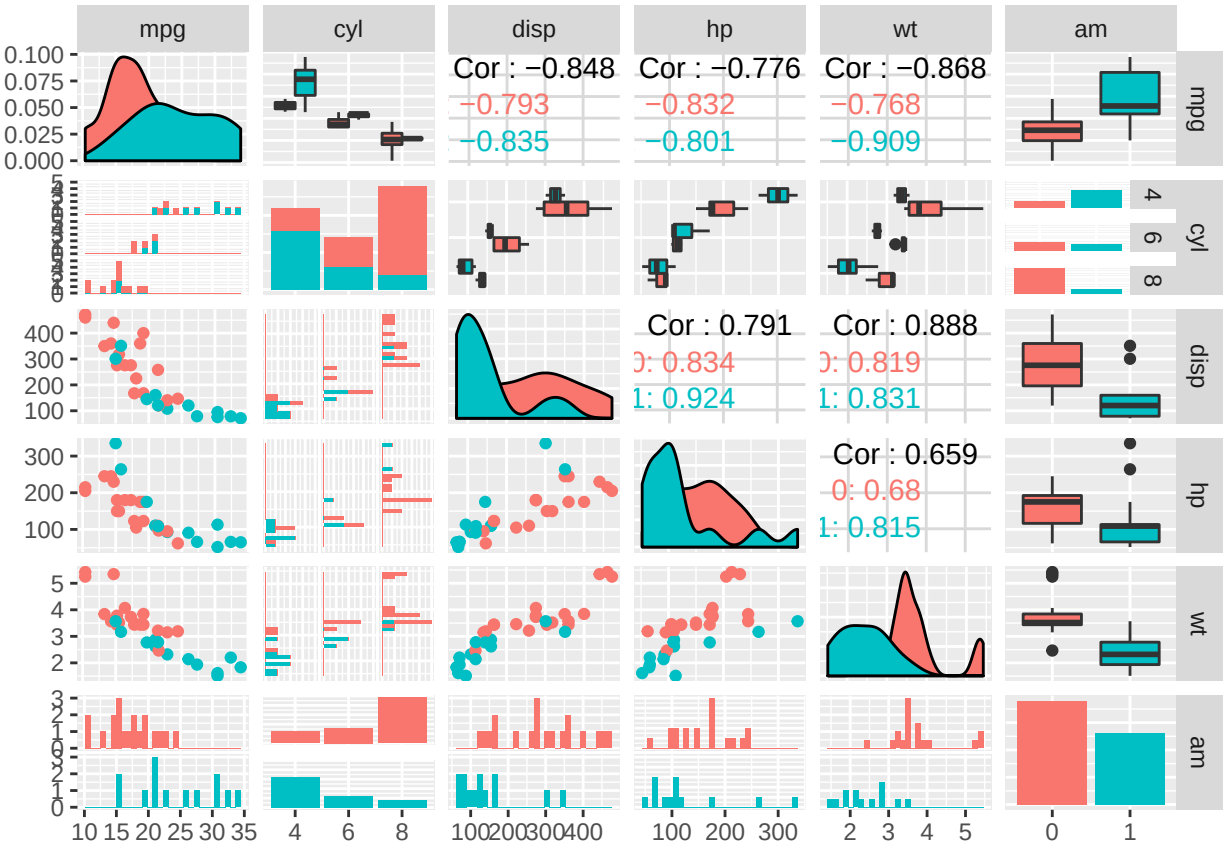
Since the obtained p-value of S-W test is 0.12, which means that the null hypothesis cannot be rejected udder assumption of standard 5% confidence level.

It can be assumed, that linear regression is an appropriate toll for the analysis, since the dependant variable follows the Gaussian distribution.

Exploratory analysis

First, stage of the analysis is to consider correlations, between variables. The useful tool for exploratory data analysis is `ggpairs` function from `GGally` package, which allows for plotting variable correlations in the data set.

```
ggpairs(mtc2, aes(color=am))
```



The initial intuition seems to be confirmed by the plotted data, as it can be clearly observed, that MPG is highly correlated with other automobile characteristics, as well as with the 'Transmission type' which is in focus of this study.

In order to get the correct confusions, the fuel consumption must be analyzed, taking into possible intercorrelations between explanatory variables.

Model selection

The basic model selection strategy is nested model testing, as discussed in the course. The goal is to balance the information explained by model on one hand and on the other, not to over inflate the residual variance with correlated exploratory variables.

The nested model testing is a technique compares increasingly complex models, in order to test weather new variables parameters are statistically significant or not.

Since the 'Type of transmission' is the variable of interest of this analysis, it will be included in

The other variables will be added, depending on their correlation with the dependant variable. Since some of the variables are factors, the spearman correlation seems to be more appropriate for comparing continuous and categorical variables.

```
# 1st row of the output: correlation of MPG with other variables
cor(mtcars, use="complete.obs", method="spearman")[1,]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.9108013 -0.9088824 -0.8946646  0.6514555 -0.8864220
##       qsec         vs         am       gear       carb
##  0.4669358  0.7065968  0.5620057  0.5427816 -0.6574976
```

It appears, that `cyl`, `disp`, `hp` and `wt` are highly correlated with MPG, while the variable of interest `am` is only moderately correlated.

The following listing represents the applied model selection strategy.

```
# Fitting nested models
fit1 <- lm(mpg~am, data=mtc2)
fit2 <- update(fit1, mpg~am+cyl)
fit3 <- update(fit1, mpg~am+cyl+disp)
fit4 <- update(fit1, mpg~am+cyl+disp+hp)
fit5 <- update(fit1, mpg~am+cyl+disp+hp+wt)

anova(fit1, fit2, fit3, fit4, fit5)
```

```
## 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
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 37.9300 2.678e-08 ***
## 3     27 230.46  1     34.04  5.6572  0.025339 *
## 4     26 183.04  1     47.42  7.8820  0.009541 **
## 5     25 150.41  1     32.63  5.4236  0.028246 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After fitting the series of nested models, it can be observed, that only the `cyl`, `disp`, `hp` and `wt` are contributing significant informations to the model.

In order to conclude the model selection procedure, the Variance Inflation needs to be assessed, since the explanatory variables are correlated.

```
vif(fit5)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## am    2.590898  1        1.609627
## cyl   9.765272  2        1.767751
## disp 12.901490  1        3.591864
## hp    4.736101  1        2.176258
## wt    6.821979  1        2.611892
```

The `disp` variable is contributing the most to the Inflation of Variation of fitted model. It should be removed, as it does not contribute new information to the model.

Selected model summary

The following listing displays model statistics and parameter estimates.

```
# Fitting selected  model
fit6 <- lm(mpg~am+cyl+hp+wt, data=mtc2)
summary(fit6)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtc2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## am1          1.80921    1.39630   1.296  0.20646
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ---
## 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:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

The model is well fitted, with 84% of adjusted R^2. However, it appears, that with standard confidence level of 5%, the `am` variable is statistically insignificant.

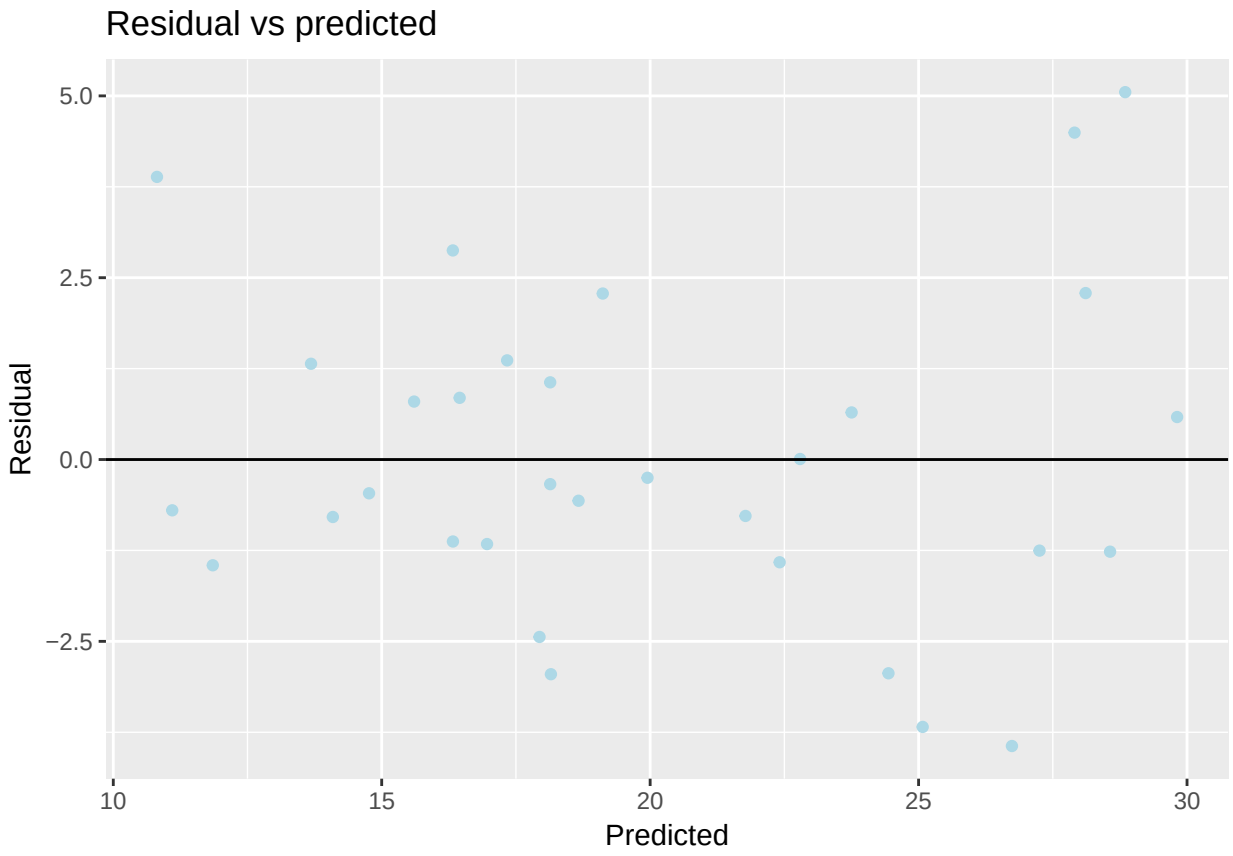
The same can be observed for the cyl=8 binary variable, which indicates, that there is no statistically significant difference in MPG between cyl=6 and cyl=8.

The obtained results indicate that transmission type has statistically insignificant impact on fuel economy, taking into consideration other factors.

Model diagnostics

In order to perform basic model diagnostics, that residual vs. fitted values is usually plotted.

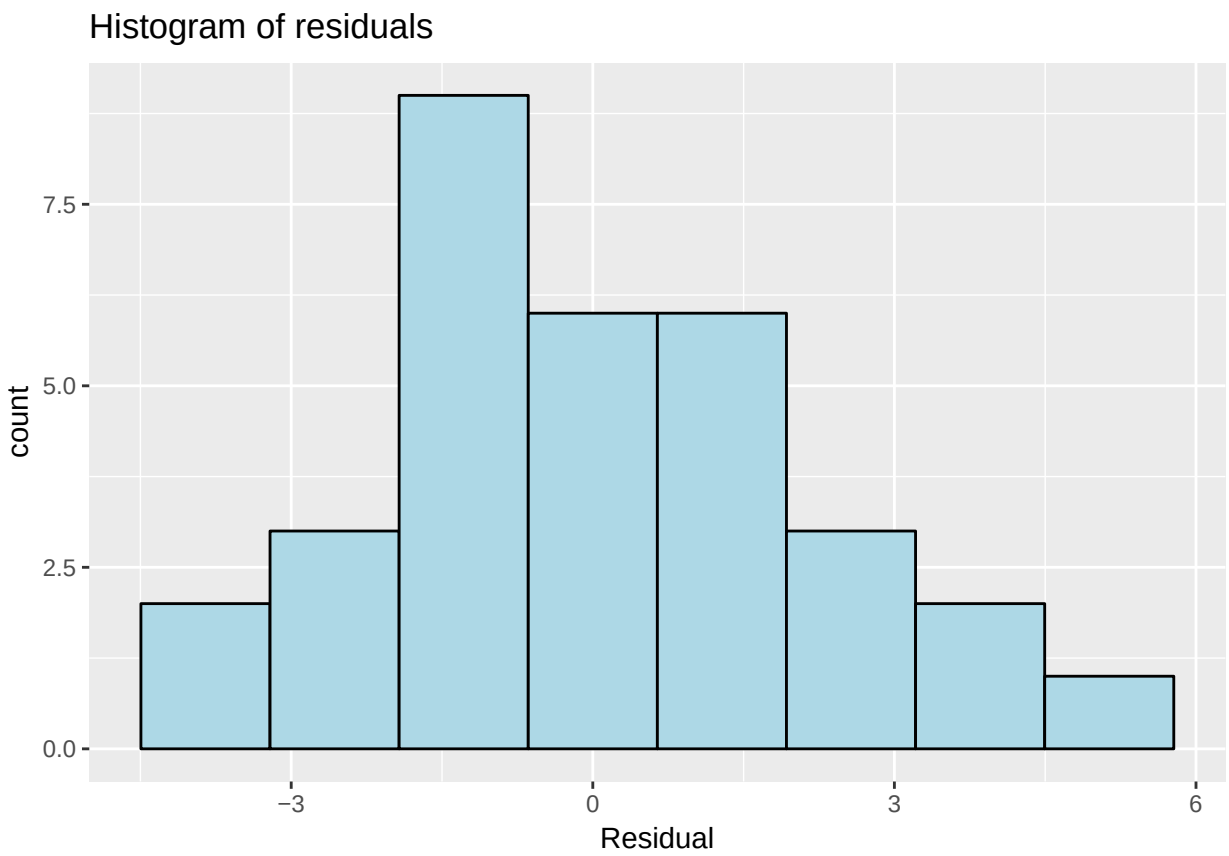
```
# plot residuals vs predicted
res <- data.frame(e=fit8$residuals, yh=predict(fit8))

ggplot(res, aes(x=yh,y=e)) +
    geom_point(color="lightblue") +
    geom_hline(yintercept=0) +
    ggtitle("Residual vs predicted") +
    xlab("Predicted") +
    ylab("Residual")
```



The residual vs predicted plot show no clear pattern, therefore it can be assumed that the residuals are indeed randomly distributed. In addition, the residuals should be normally distributed, which can be verified with residual histogram.

```
ggplot(res, aes(e)) +
    geom_histogram(fill="lightblue", color="black", bins=8) +
```

```
    ggtitle("Histogram of residuals") +
    xlab("Residual")
```



The residuals histogram displays nos particular indication that residuals are not normally distributed. It can be assumed, that selected model is not biased.