Is automatic or manual transmission better for MPG?

Bankbintje

17th august 2015

Executive Summary

The magazine Motor Trend is interested in exploring the relationship between a set of variables and miles per gallon (MPG). They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG?"
- "Quantify the MPG difference between automatic and manual transmissions"

The best (i.e. model with the highest adjusted R^2) multivariate regression model was based on *number of cylinders*, *horsepower*, *car weight*, and *transmission type*; explaining 84.0% of variability in the data. These variables were selected using the Akaike Information Criterion (AIC). After validating this model the conclusion is that *cars with manual transmission are better for MPG and have on average a 1.80 higher mpg than cars with automatic transmission.*

Data

Exploratory Data Analysis

This analysis is based on the mtcars dataset from the base r datasets-package. 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).

```
data=mtcars
```

The appendix *(table 1 & plots 1-2)* contains the results of basic data exploration. Main conclusions are:

- the data does not seem to contain missing values and/or clear outliers.
- there is a difference in fuel consumption (mpg) per Transmission Type. *(see appendix, plot 2)*
- the distribution of fuel consuption (mpg) appears *normal*, this allows for fitting a linear model. *(see appendix, plot 1)*

Preparing Data

The dataset is tidy and needs no cleaning. The variables am, cyl, vs, gear and carb will be converted to factors, the values 0 and 1 for transmssion type will be replaced by "Automatic" and "Manual".

```
mtcars.clean<-mtcars
mtcars.clean$am <- as.factor(mtcars$am)
mtcars.clean$cyl <- as.factor(mtcars$cyl)
mtcars.clean$vs <- as.factor(mtcars$vs)
mtcars.clean$gear <- as.factor(mtcars$gear)
mtcars.clean$carb <- as.factor(mtcars$carb)
levels(mtcars.clean$am) <-c("Automatic", "Manual")
```

Hypothesis testing

A simple boxplot suggests that there is a difference in fuel consumption per transmission type. *(see appendix, plot 2)* Running a Welch Two Sample t-test could confirm whether the difference is significant and whether we can reject our null-hypothesis that *no significant difference exists between cars with automatic transmission and cars with manual transmission*.

```
t.test(mtcars.clean[mtcars.clean$am == "Automatic",]$mpg,
       mtcars.clean[mtcars.clean$am == "Manual",]$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars.clean[mtcars.clean$am == "Automatic", ]$mpg and mtcars.clean[mtcars.clean$am == "Manual
## 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 of x mean of y
##  17.14737  24.39231
```

The low p-value (0.001374) indicates that the probability that this difference is accidental is very low; the difference is *significant* and the null-hypothesis can be rejected. For further quantifying the effect we need to apply linear regression.

Regression on a single variable

Perform a simple regression of fuel consumtion (mpg) on transmission type (am):

```
fit1<-lm(mpg ~ am, data=mtcars.clean)
```

The coefficents of this model are:

```
summary(fit1)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

The intercept of 17.15 is the mean mpg of cars with automatic transmission. The amManual estimate is the expected change in mpg from automatic transmission to manual transmission. The low p-value (0.000285) might tempt us to conclude that cars with manual transmission have - on average - a 7.25 higher mpg. However, the adjusted R^2 value of this model is quite low:

```
summary(fit1)$adj.r.squared
```

```
## [1] 0.3384589
```

This means that only 34% of variation can be explained by this regression model; this is not enough to quantify a possible effect.

Multivariate Regression

Model Selection

Creating a model using all variables does not single out any variable with a significant p-value.

```
fit.mv.0 <- lm(formula = mpg ~ ., data = mtcars.clean)
summary(fit.mv.0)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars.clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## amManual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

It does however mark weight and horsepower as significant candidates. Therefore, this will be the first multivariate model we will evaluate:`lm(formula = mpg ~ wt + hp + am, data = mtcars.clean)`

Another approach is to select a model on the Akaike Information Criterion (AIC) using a stepwise algorithm: `step(fit.mv.0,direction="both", k=2)` This produces the second multivariate model we will evaluate:`lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars.clean)`

Model Comparison

The conclusion will be based on the model having the highest Adjusted R^2 value.

Model 1 (wt + hp + am) The R^2 value of the model using weight, horsepower, and transmission type:

```
fit.mv.1 <- lm(formula = mpg ~ wt + hp + am, data = mtcars.clean)
summary(fit.mv.1)$adj.r.squared
```

```
## [1] 0.8227357
```

Model 2 (cyl + hp + wt + am) The R^2 value of the model using nbr of cylinders, horsepower, weight, and transmission type:

```
fit.mv.2 <- (lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars.clean))
summary(fit.mv.2)$adj.r.squared
```

```
## [1] 0.8400875
```

We will use model 2 that was selected using AIC because of its (slightly) higher Adjusted R^2, provided no anomalies pop up during model diagnosis.

Model Diagnosis

The diagnostic plots *(see appendix, plot 3)* confirm that

- no pattern exists in the residuals vs. fitted plot (i.e. independence)
- the Q-Q plot produces a line (i.e. residuals are distributed normally)
- no pattern exists in the scale-location plot (i.e. constant variance)
- the data contains no influential outliers in the residuals vs. leverage plot.

Appendix

Tables

Table 1: Summary of MPG per transmission type

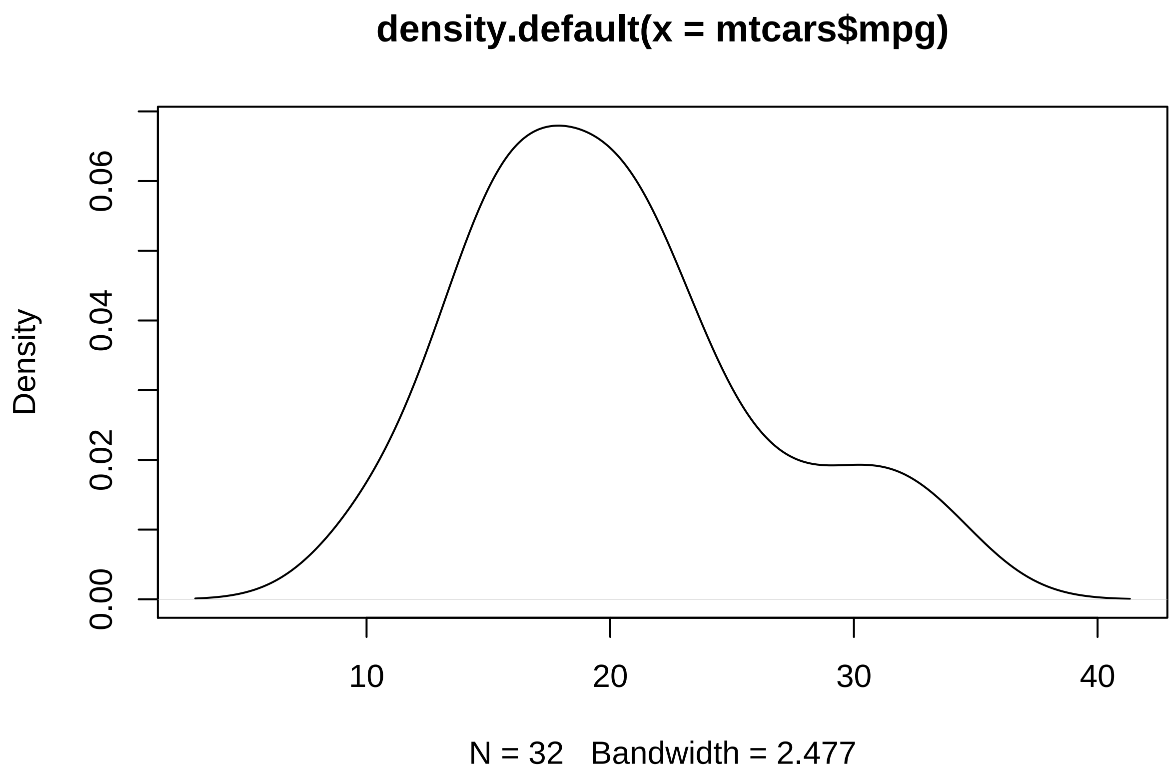
```
by(mtcars$mpg, mtcars$am, summary)
```

```
## mtcars$am: 0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
## --------------------------------------------------------
## mtcars$am: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

Plots

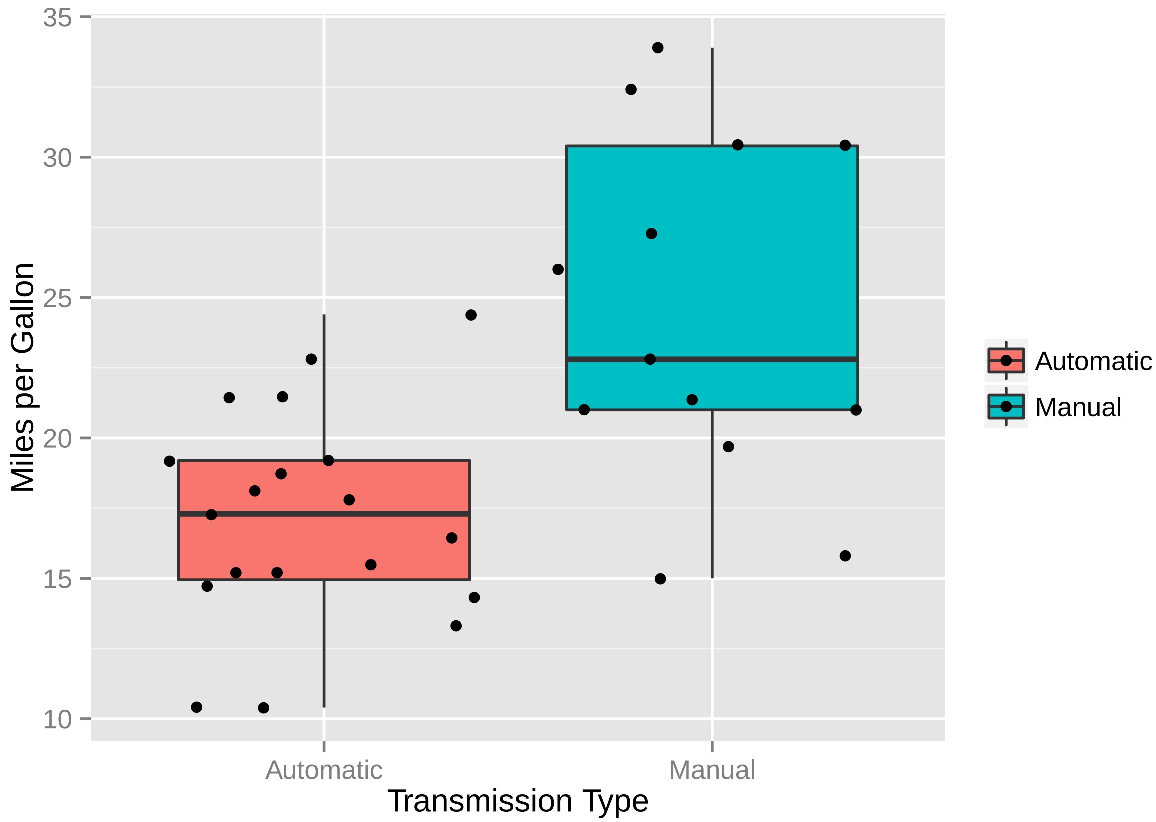
Plot 1: Density plot MPG

```r
plot(density(mtcars$mpg))
```



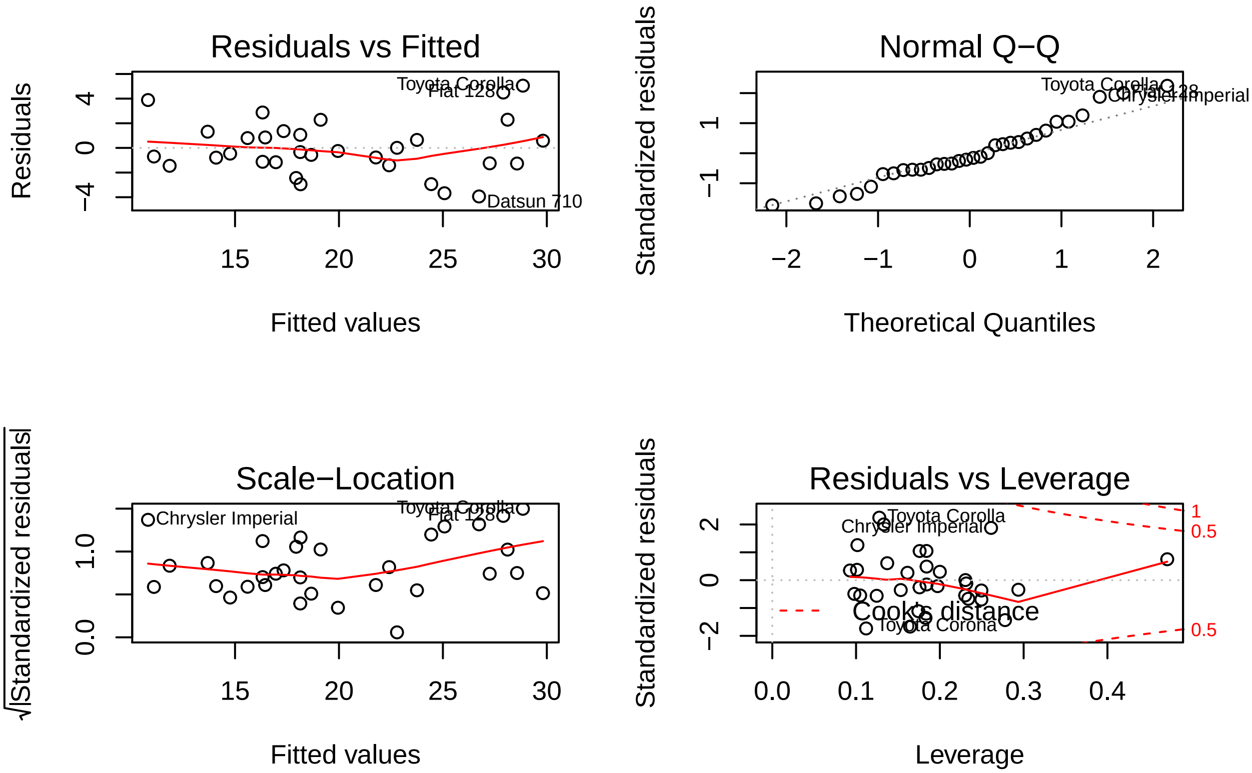
Plot 2: Boxplot MPG and Transmission Type

```r
library(ggplot2)
ggplot(mtcars.clean, aes(x=am, y=mpg, fill=am)) +
        geom_boxplot() +
        geom_jitter() +
        ylab("Miles per Gallon") +
        xlab("Transmission Type") +
        theme(legend.title=element_blank())
```



Plot 3: Diagnostics

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


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
Toyota Corona
1
0.5