Motor Trends Analysis

wdewit

October 8, 2016

Executive Summary

In this analysis we investigated the relationship between fuel consumption and the transmission type. We had information on both variables as well as a set of 9 other car characteristics at our disposal for 32 car brands.

Exploratory Data Analysis

Since fuel consumption (mpg) is a continous variable we will estimate a linear regression model. One of the assumptions of a linear model is that the dependent variable is normally distributed.

Comparing the density of mpg with the normal density we see mpg is somewhat skewed to the right (fig 1 - Appendix). However the deviation from normality is not too strong and probably caused by our rather small sample size. Given that least squares estimation is quite robust against violation of the normality assumption, we continue fitting a linear regression model by means of least squares.

Next we'll have a first glance whether there does seem to be a relationship between transmission type and fuel consumption. The average fuel consumption with an automatic transmission is 17.14 mpg, which is nearly one third less than the average fuel consumption with a manual transmission. From a boxplot (fig 2 - Appendix) we also notice that the fuel consumption with a manual transmission is far more variable than with an automatic transmission, especially skewed towards higher values of mpg.

Formal modelling

We now want to confirm whether there is a statistically significant difference between fuel consumption of automatic versus manual transmission types. We therefore fit a simple linear regression model between both.

Fitting a linear model of mpg on transmission type, we see that the coefficient for manual transmission is indeed signifant at alpha=5%, and positive so that we can confirm that fuel consumption with manual transmission is higher than with an automatic transmission. Next we'll check whether any of the other variables included in the mtcars dataset might confound this relationship. Therefore we'll fit the full model and see whether any of the other predictors also have a significant impact on fuel consumption. The F-test - testing whether the increase in the residual sum of squares when reducing the number of predictors is sufficiently small - is highly significant at alpha=5%, concluding there are important other predictors than transmission type solely in play. Nevertheless when looking at the p-values of the individual parameter estimates none of them are significant at alpha=5%, indicating a likely problem of multicollinearity among the predictors. Since horsepower and weight have the lowest p-value we will test whether we can jointly drop the other predictors: cyl, disp, drat, qsec, vs, gear and carb in fitting a third model including solely transmission type, horsepower and weight and comparing this to the full model including all variables.

!!! Draai dit hierboven en zie of je joint test ok is en dan checken of je 2 extra variabelen relatie veranderen. Erna nog residual plots (in appendix) en als geen outliers dan nog effect in absolute getallen inschatten en een executive summary schrijven !!!

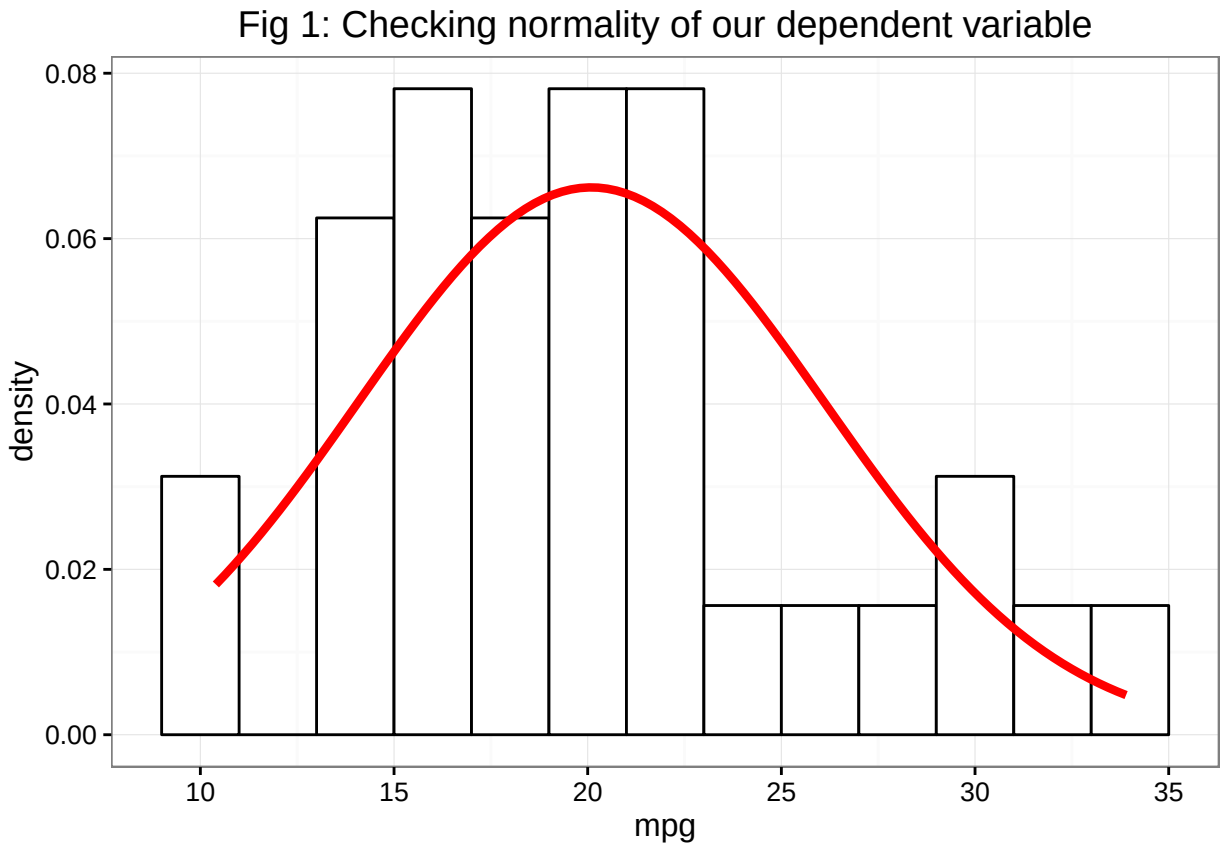
Appendix:

Figures and charts of exploratory analysis

```
tapply(mtcars$mpg,mtcars$am,mean)
```

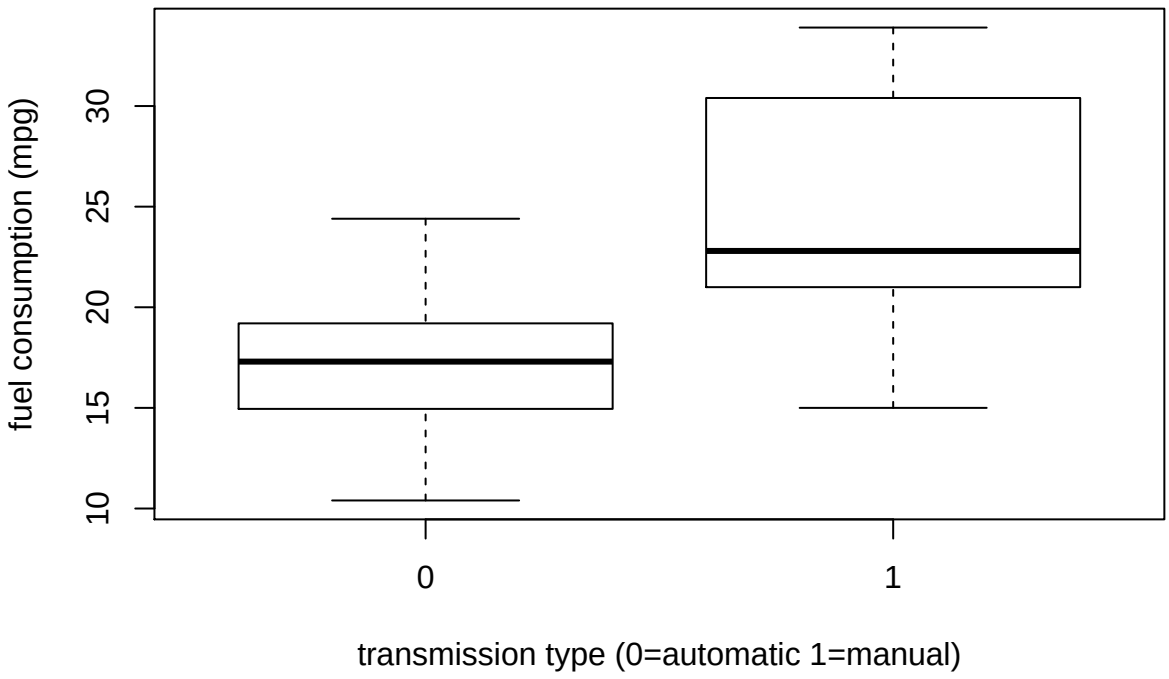
```
##        0        1
## 17.14737 24.39231
```

```
library(ggplot2)
g <- ggplot(data=mtcars,aes(x = mpg))+ labs(title="Fig 1: Checking normality of our dependent variable")
g <- g + geom_histogram(aes(y = ..density..), fill = "white", binwidth=2, colour = "black")
g + stat_function(fun=dnorm,colour="red",
                  size=1.5, args=list(mean=mean(mtcars$mpg), sd=sd(mtcars$mpg)))+theme_bw()
```



```r
boxplot(mtcars$mpg~mtcars$am,xlab="transmission type (0=automatic 1=manual)", ylab="fuel consumption (mp
```

Fig 2: Comparing distribution of mpg by transmission type



Figures and charts of formal modelling

```r
fit_main <- lm(mpg~am,data=mtcars)
summary(fit_main)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
```

```
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
fit_full <- lm(mpg~.,data=mtcars)
anova(fit_full,fit_main)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1     19 130.05
## 2     30 720.90 -11   -590.85 7.8473 5.682e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit_full)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2015 -1.2319  0.1033  1.1953  4.3085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.09262   17.13627   0.881   0.3895
## cyl6        -1.19940    2.38736  -0.502   0.6212
## cyl8         3.05492    4.82987   0.633   0.5346
## disp         0.01257    0.01774   0.708   0.4873
## hp          -0.05712    0.03175  -1.799   0.0879 .
## drat         0.73577    1.98461   0.371   0.7149
## wt          -3.54512    1.90895  -1.857   0.0789 .
## qsec         0.76801    0.75222   1.021   0.3201
## vs1          2.48849    2.54015   0.980   0.3396
## am1          3.34736    2.28948   1.462   0.1601
## gear4       -0.99922    2.94658  -0.339   0.7382
## gear5        1.06455    3.02730   0.352   0.7290
## carb         0.78703    1.03599   0.760   0.4568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.616 on 19 degrees of freedom
## Multiple R-squared:  0.8845, Adjusted R-squared:  0.8116
## F-statistic: 12.13 on 12 and 19 DF,  p-value: 1.764e-06
```

```
fit_red <- lm(mpg~am+hp+wt,data=mtcars)
anova(fit_full,fit_red)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am + hp + wt
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     19 130.05
## 2     28 180.29 -9    -50.24 0.8155 0.6089
```

```
anova(fit_red,fit_main)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + hp + wt
## Model 2: mpg ~ am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     28 180.29
## 2     30 720.90 -2   -540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```