Motor Trends Analysis

wdewit

October 8, 2016

Executive Summary

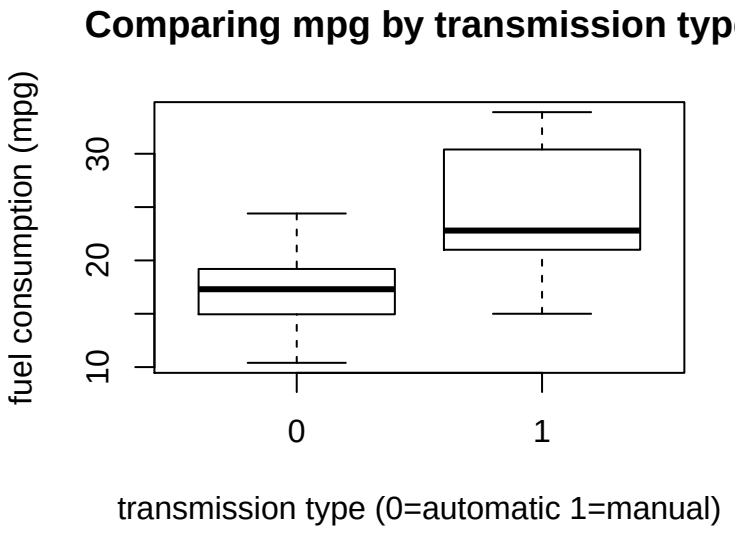
In this analysis we investigated the relationship between fuel consumption and the transmission type. We had information on both variables as well as a set of 9 other car characteristics at our disposal for 32 car brands. Keeping horsepower and weight constant we found no significant difference in fuel consumption between a manual and automatic transmission.

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 noticed 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. From a boxplot we notice that the median miles per gallon (mpg) is quite higher for a manual transmission than for an automatic transmission and it is also more variable.

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



Formal modelling

We now want to confirm whether there is a statistically significant difference. We therefore fit a simple linear regression model between both. We see that the coefficient for manual transmission is with a p-value

of 0.000285 (table 1 - Appendix) indeed signifant at alpha=5%, and positive so that we can confirm that miles per gallon is higher with a manual transmission than with an automatic transmission. This seems counterintuive. Therefore, we'll check whether any of the other variables included in the mtcars dataset might confound this relationship. We start in fitting the full model, including all variables in the dataset, 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% (Table 2 - Appendix), 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 still low p-values (< 0.10) 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. The F-test for jointly dropping the 7 predictors is 0.8155 with p-value of 0.6089 (Table 3 - Appendix) and therefore we conclude we can use the simplified model with 3 predictors. The F-test comparing the simplified model to the model only including transmission type attains 41.979, which is highly significant at alpha=0.05. So, horsepower and weight should be included in the model.

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

```
##
## Call:
## lm(formula = mpg ~ am + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am1          2.083710   1.376420   1.514 0.141268
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## wt          -2.878575   0.904971  -3.181 0.003574 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

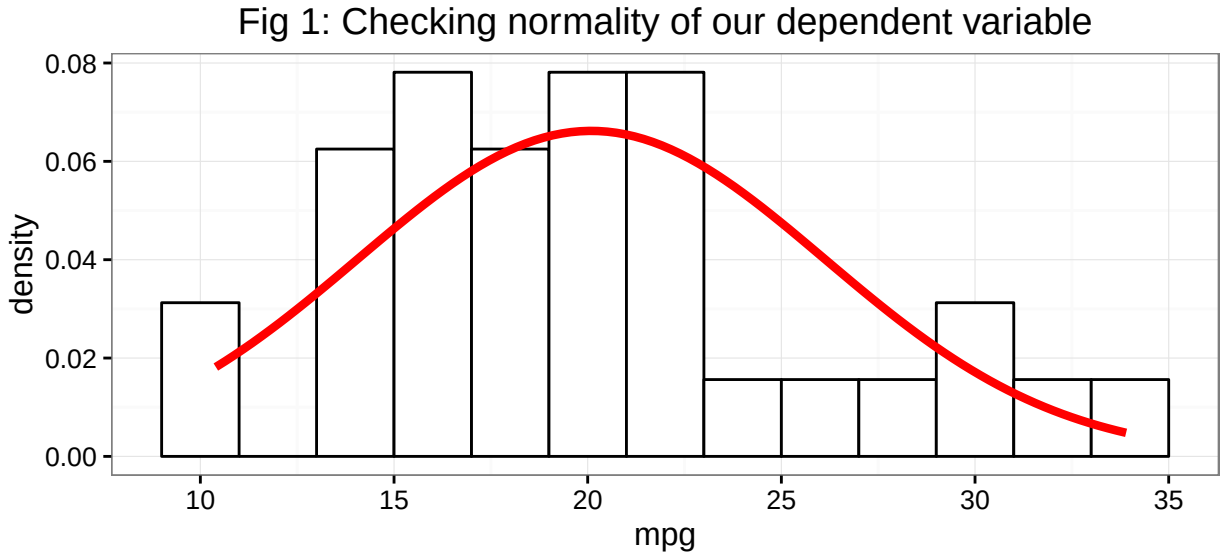
In our final model we see that when comparing cars of similar horsepower and weight, there is nolonger a significant difference between fuel consumption of manual versus automatic transmission cars. The counterintuitive relationship between mpg and transmission type found earlier was entirely due to the fact that the cars in our dataset with manual transmission had in general lower weights and horsepower and lower weight and horsepower lead to higher miles per gallon.

Finally, we have a look at the residual plots to confirm the validity of our model (Fig 2 - Appendix). Although from the plot of the residuals versus the fitted values it seems the model might benefit from adding a quadratic relationship of horsepower and/or weight this won't change our conclusions wrt relation between mpg and transmission type. Neither does there seem to be a concern of influential points (dfbetas for obs 17, 18 and 20 below 1). Overall we confirm the validity of our model and conclude there is no evidence of differences in mpg between a manual and automatic transmission, holding weight and hp constant.

Appendix:

Figures of exploratory analysis

```
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()
```



Figures and charts of formal modelling

```
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
```

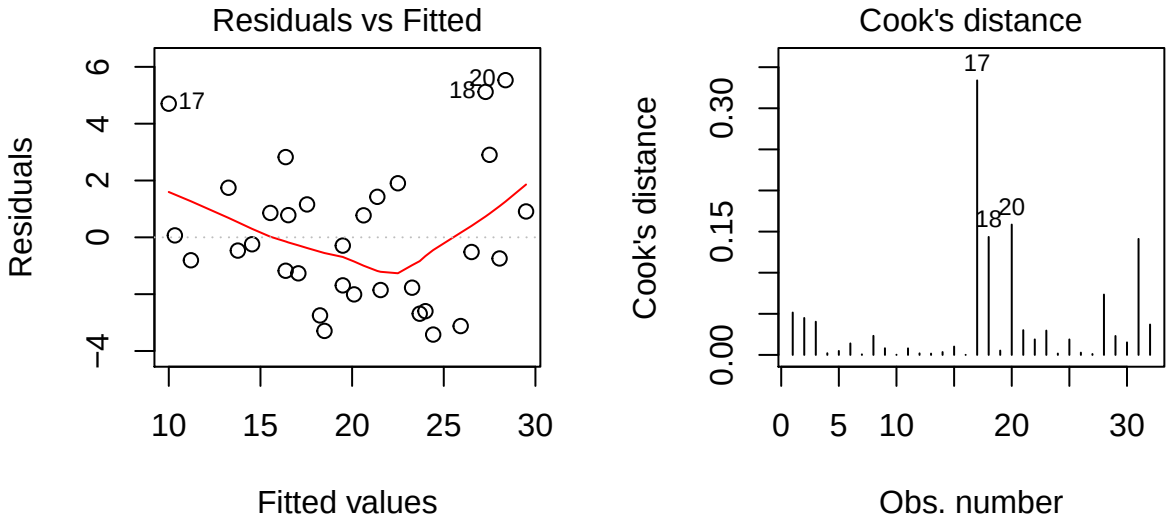
```
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am + hp + wt
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     19 130.05
## 2     28 180.29 -9    -50.24 0.8155 0.6089
```

```
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
```

```
par(cex.main=0.8,mfrow=c(1,2))
plot(fit_red,which=1)
plot(fit_red,which=4)
title(main="Fig 2: Residual plots")
```

Fig 2: Residual plots




```
dfbetas(fit_red)[c(17,18,20),]
```

```
##    (Intercept)       am1         hp         wt
## 17 -0.92990022 0.5400209 -0.3350091  0.9726881
## 18 -0.01883095 0.4322459 -0.4329640  0.2100033
## 20  0.28199793 0.2231945 -0.2400377 -0.1101555
```