Motor Trend: Association between Transmission and MPG

Chiara Di Gravio

25 August 2016

Executive Summary

- The association between car transmission type and miles per gallon (MPG) was studied using multiple regression models. An unadjusted analysis with transmission type as the only predictor was fit.
- To account for possible confounding and to select the "best" model, we fit a full model (with all the variables in the dataset as predictors) and used a stepwise procedure to select the model that minimises the AIC. The resulting model includes transmission type (am), weight (wt) and 1/4 mile time (qsec).
- After adjustments, the "best" model showed that manual cars had (on average) 2.94 miles per gallon more than automatic cars (95% confidence interval: 0.05 - 5.83, pvalue = 0.046).
- Diagnostic plot were generated and leverage points were studied. The "best" model was again fitted after excluding high leverage points and the association between transmission type and MPG became not significnat (estimate: 2.20, 95% CI: -0.45 - 4.84, pvalue = 0.12).

Exploratory Data Analysis

The dataset comprised of 32 cars. For each car, 11 variables were collected.

```
# load packages
library(ggplot2)
library(gridExtra)
library(knitr)

# import dataset and transform variables to factors
data(mtcars)
mtcars$cyl <- factor(mtcars$cyl); mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am); levels(mtcars$am) <- c("Automatic", "Manual")
```

Of the 32 cars, 19 were automatic and 13 were manual. The mean MPG was 20.1 (SD: 6.03). Tables 1 and 2 summarises the results.

```
kable(as.data.frame(table(mtcars$am)), caption = "Frequency of Transmission Type",
      col.names = c("Type", "Frequency"))
```

Table 1: Frequency of Transmission Type

Type	Frequency
Automatic	19
Manual	13

```
kable(t(c(summary(mtcars$mpg), SD = sd(mtcars$mpg))), caption = "MPG Summary Statistics")
```

Table 2: MPG Summary Statistics

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
10.4	15.42	19.2	20.09	22.8	33.9	6.026948

From a short graphical analysis (see Appendix) manual cars seemed to have higher MPG. However, the variablity of MPG seemed

higher for manual than for automatic cars.

Regression Models

First, a regression model with only transmission type as predictor was implemented:

```
mod1 <- lm(mpg ~ am, data = mtcars)
kable(round(cbind(summary(mod1)$coeff, confint(mod1)),3), caption = "Unadjusted Analysis Results",
      col.names=c("Estimate", "Std. Error", "t value","p-value", "95% CI: LB","95% CI: UB"))
```

Table 3: Unadjusted Analysis Results

	Estimate	Std. Error	t value	p-value	95% CI: LB	95% CI: UB
(Intercept)	17.147	1.125	15.247	0	14.851	19.444
amManual	7.245	1.764	4.106	0	3.642	10.848

Then, to consider the presence of possible confounders and to select the best model based on the one AIC criteria, a full model (one with all the predictors) was fit, and a stepwise procedure was implemented to select the "best" model.

```
# full model
fullmod <- lm(mpg ~ ., data = mtcars)
# stepwise selection
bestmod <- step(fullmod, scope = mod1, direction = "both", trace = FALSE)
kable(round(cbind(summary(bestmod)$coeff, confint(bestmod)),3), caption = "Best Model Results",
      col.names=c("Estimate", "Std. Error", "t value","p-value", "95% CI: LB","95% CI: UB"))
```

Table 4: Best Model Results

	Estimate	Std. Error	t value	p-value	95% CI: LB	95% CI: UB
(Intercept)	9.618	6.960	1.382	0.178	-4.638	23.874
wt	-3.917	0.711	-5.507	0.000	-5.373	-2.460
qsec	1.226	0.289	4.247	0.000	0.635	1.817
amManual	2.936	1.411	2.081	0.047	0.046	5.826

Holding the weight and the 1/4 mile time variables constant, transmission type was significantly assiciated with MPG. The MPG in manual cars was (on average) 2.94 (95% CI: 0.05, 5.83) more when compared to automatic cars.

Model diagnostic (see appendix for graphs), carried out using residual plot and QQplot of the residual, showed that linear regression hypothesis are mostly satisfied with no evident pattern in the residual plot and residual being fairly normally distributed. Further investigation might be due to better understand whether the linear regression model was appropriate. The model diagnostic plots identied three points of interest:

```
tocheck <- c("Chrysler Imperial", "Fiat 128", "Toyota Corolla")
mtcars$leverage <- hat(model.matrix(bestmod))
kable(mtcars[rownames(mtcars) %in% tocheck,])
```

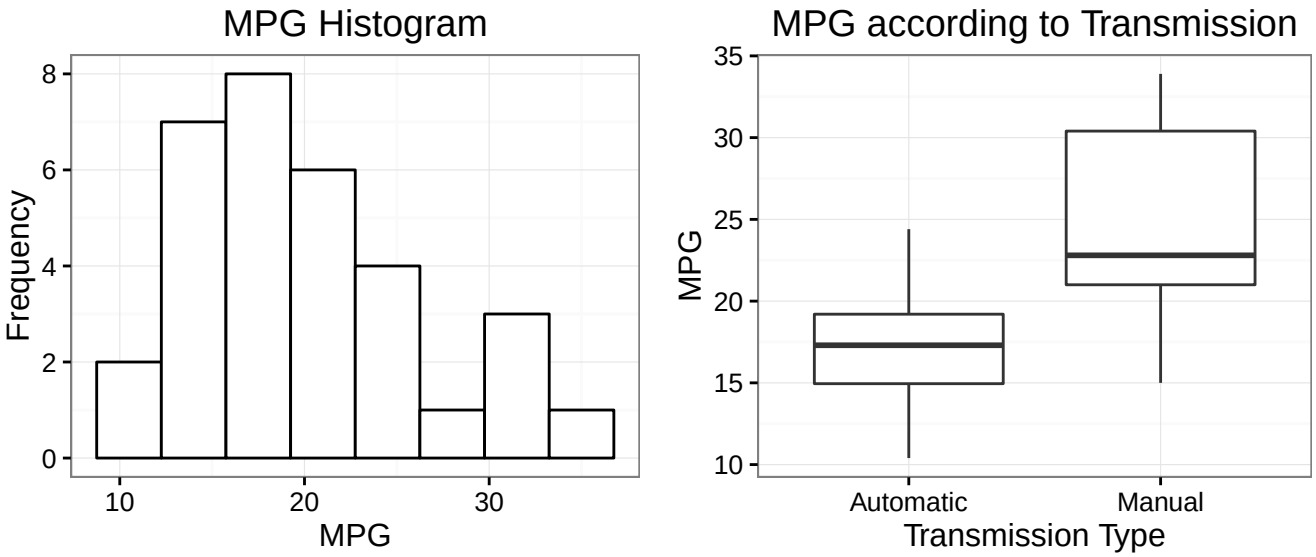
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	leverage
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	Automatic	3	4	0.2296338
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	Manual	4	1	0.1276313
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	Manual	4	1	0.1463489

Using as cut-off for high leverage point $2 \times p/n = 0.1875$ (p: number of variables, n: number of observations), only Chrysler Imperial could be considered a point of high leverage. After excluding the high leverage point, the stepwise procedure identified the same "best" model; however, the association between transmission type and MPG became not significant (see Appendix).

Appendix

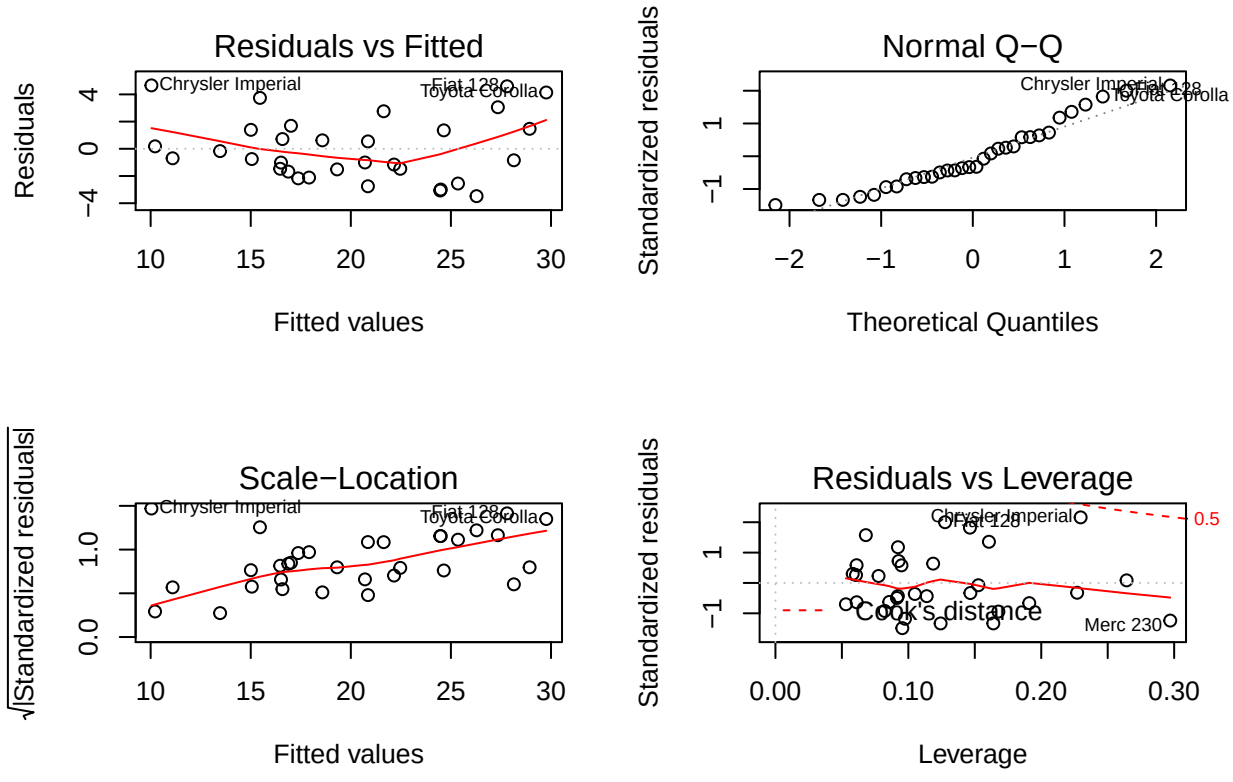
Exploratory Data Analysis: MPG Distribution and MPG Boxplots according to Transmission Type

```
p1 <- ggplot(mtcars, aes(x = mpg)) +
      geom_histogram(colour = "black", fill = "white", binwidth = 3.5) +
      theme_bw() + xlab("MPG") + ylab("Frequency") + ggtitle("MPG Histogram")
p2 <- ggplot(mtcars, aes(x = am, y = mpg)) + geom_boxplot() + theme_bw() +
      xlab("Transmission Type") + ylab("MPG") + ggtitle("MPG according to Transmission")
grid.arrange(p1, p2, ncol = 2)
```



Linear Regression: Diagnostic Plot

```
par(mfrow=c(2,2))
plot(bestmod)
```



Excluding leverage point

```
fullmod <- lm(mpg ~ ., data = mtcars[rownames(mtcars) != "Chrysler Imperial",])
bm <- step(fullmod, scope = mod1, direction = "both", trace = FALSE)
summary(bm)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars[rownames(mtcars) !=
##     "Chrysler Imperial", ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.453 -1.385 -0.791  1.381  4.633
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.6680     6.7009   2.040 0.051272 .
## wt           -4.6397     0.7308  -6.349 8.48e-07 ***
## qsec          1.1355     0.2712   4.188 0.000269 ***
## amManual      2.1978     1.3496   1.629 0.115029
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.286 on 27 degrees of freedom
## Multiple R-squared:  0.8713, Adjusted R-squared:  0.857
## F-statistic: 60.92 on 3 and 27 DF,  p-value: 3.819e-12
```