Automatic or manual transmission?

Joanne Breitfelder

11 Aug 2016

In 1974, *Motor Trend* magazine published the main caracteristics of 32 automobiles, including fuel consumption and 10 aspects of automobile design and performance. Today, we are interested in exploring the relationship between these different variables and miles per gallon (hereafter MPG). We are particularly interested in quantifying the MPG difference between automatic and manual transmissions and answer the following question : which is better for MPG between automatic and manual transmission?

Exploratory data analysis

```r
data <- mtcars
data$am <- factor(data$am, labels=c("automatic", "manual"))
kable(head(data, 3))
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	manual	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	manual	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	manual	4	1

By looking at the graphics shown in the **appendix 1**, we can already infer that manual transmission is better for MPG. Let's test this hipothesis with a one-sided T-test :

```r
t.test(filter(data, am=="manual")$mpg,
       filter(data, am=="automatic")$mpg,
       alternative="greater", var.equal=FALSE)$p.value
```

```
## [1] 0.0006868192
```

The p-value is well below 0.05, which indicates that we can adopt the alternative hyothesis : *the true difference in means is greater than 0.* However, this relationship assumes that the other variables are left constant, which is not true. We therefore need to include them in our study.

Regression models

Univariate Linear Regression

Let's first adjust a simple linear model between `am` and `mpg`, with the command `lm(mpg~am, data)`. This regression gives a R-squared value of **0.3597989** and the following coefficients :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
ammanual	7.244939	1.764422	4.106127	0.000285

The coefficients suggest that manual transmission cars have in average 7.245 ± 1.764 MPGs more than automatic transmission cars, which confirms the result of our t-test. However, the R^2 indicates that this model explains only 35% of the data variance. As said before, some information is missing and we need to include more variables.

Multivariate regression analysis

Let's now include all the variables in our model, with the command `lm(mpg~., data)`. We now get a R-squared of **0.8690158** and the following coefficients (rounded for clarity):

	(Intercept)	cyl	disp	hp	drat	wt	qsec	vs	ammanual	gear	carb
Estimate	12.30	-0.11	0.01	-0.02	0.79	-3.72	0.82	0.32	2.52	0.66	-0.20
Std. Error	18.72	1.05	0.02	0.02	1.64	1.89	0.73	2.10	2.06	1.49	0.83
t value	0.66	-0.11	0.75	-0.99	0.48	-1.96	1.12	0.15	1.23	0.44	-0.24
Pr(>|t|)	0.52	0.92	0.46	0.33	0.64	0.06	0.27	0.88	0.23	0.67	0.81

This new model explains 86% of the variance, which is much improved compared to before. However, none of the coefficients is associated to a p-value lower than 0.005, and almost all standard errors are bigger than the estimates! The coefficients are therefore not significative and we can not conclude.

To go further in the variables selection, let's run a stepwise algorithm:

```r
stepmodel <- step(lm(mpg~., data), trace=FALSE)
kable(coef(summary(stepmodel)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
ammanual	2.935837	1.4109045	2.080819	0.0467155

As we can see, the best model includes the *transmission mode*, the *cars weight* and the *quarter mile time*. These three variables alone explain about 84% of the MPG variance ($R^2 = 0.8496636$) and all the coefficient's p-values are significant : a very convincing result!

Regression diagnostics

Based on the diagnostic plots shown in **Appendix 2**, we can say that the residuals seem to show homogeneity (points are randomly distributed in the *Scale-Location* plot), normality (points are close to the line in the *Q-Q* plot), and independence (no pattern in the *Residuals vs. Fitted* plot).

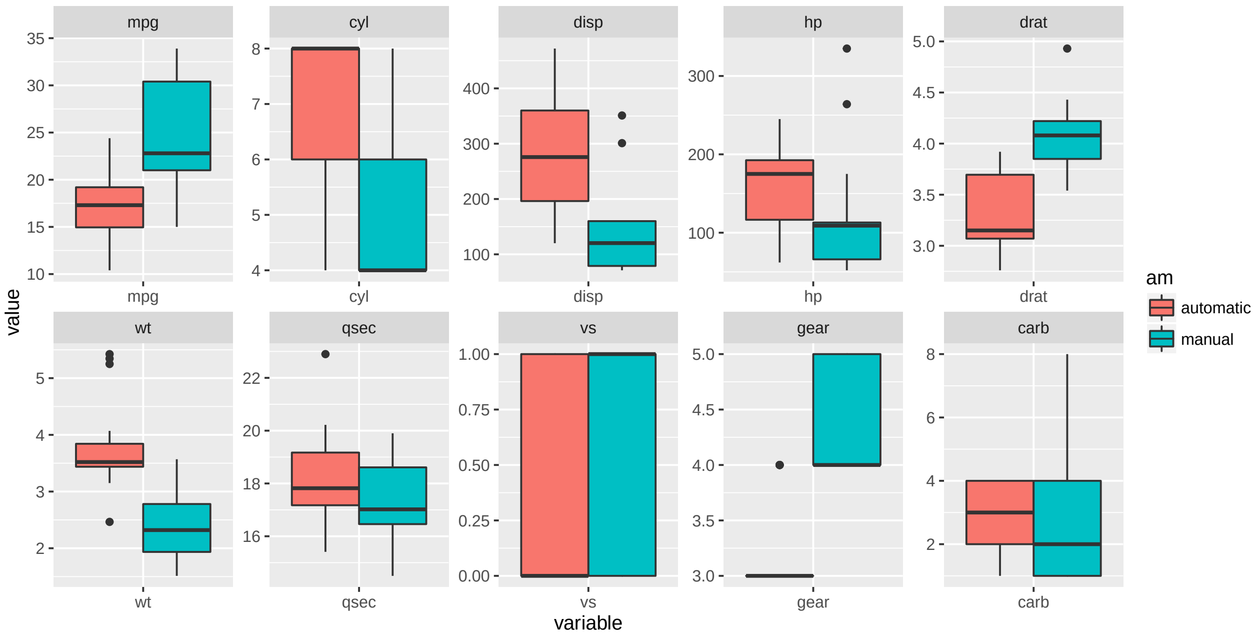
Results interpretation

Thanks to the last model, we can finally answer the initial question. In average, manual transmission cars have 2.93 ± 1.41 MPGs more than automatic transmission cars, and this value is significant at a 95% confidence level (p-value of 0.047). We can note that the difference between the two transmission modes is much lower than the one we obtained with the simple linear regression, which shows the importance of including all relevant variables in the fitting process.

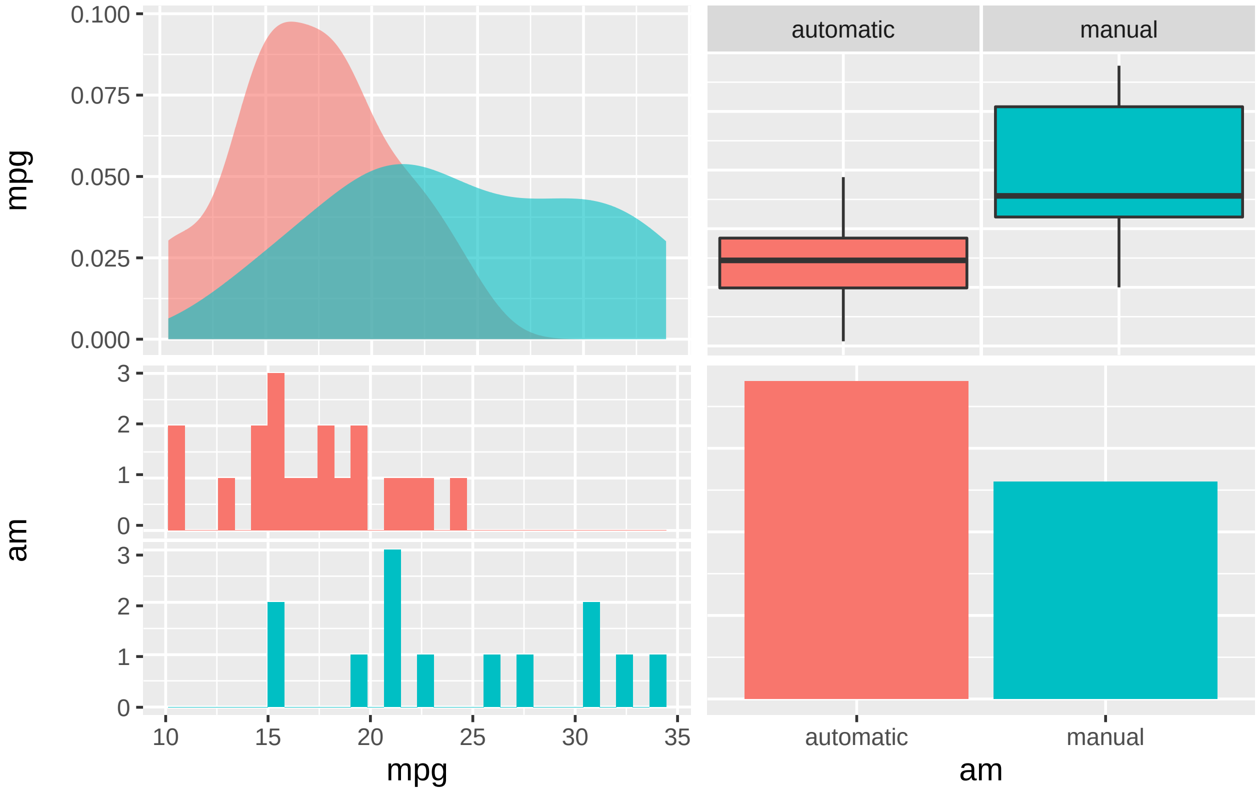
Appendix

Appendix 1: Exploratory data analysis graphics

Variation of each variable depending on the transmission mode:

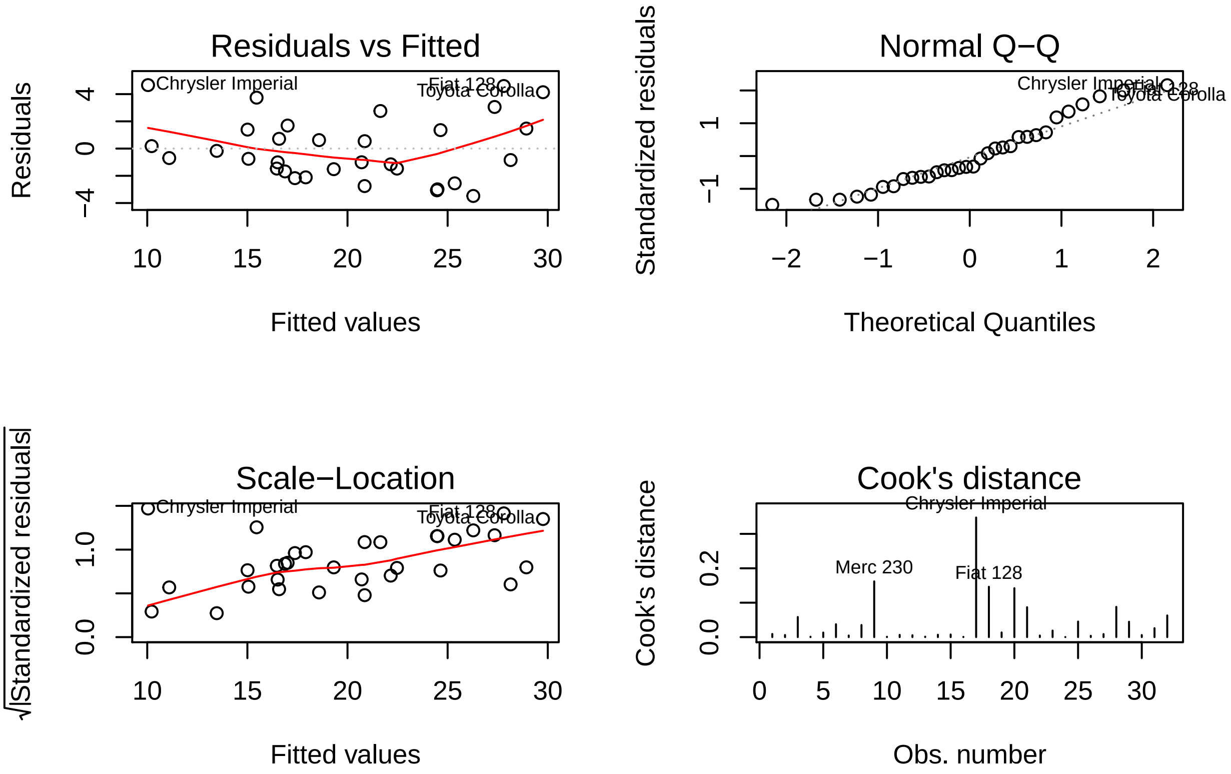


Focus on the MPG:



Appendix 2: Regression diagnostics

```
par(mfrow=c(2, 2))
plot(stepmodel, which=1:4)
```



Appendix 3: R code used in this study

```
library(knitr); library(reshape2); library(ggplot2)
library(dplyr); library(GGally)

data <- mtcars
data$am <- factor(data$am, labels=c("automatic", "manual"))

# Exploratory graphics:
ggplot(melt(data, id.vars="am"), aes(x=variable, y=value)) +
        geom_boxplot(aes(fill=am)) + facet_wrap(~variable, scale="free", ncol=5)

ggpairs(data, aes(color=am), columns=c("mpg", "am"))

# T-test:
t.test(filter(data, am=="manual")$mpg, filter(data, am=="automatic")$mpg,
       alternative="greater", var.equal=FALSE)$p.value

# Simple linear regression:
kable(coef(summary(lm(mpg~am, data))))
```

```r
summary(lm(mpg~am, data))$r.squared

# Linear regression with all variables:
kable(t(round(coef(summary(lm(mpg~., data))), 2)))
summary(lm(mpg~., data))$r.squared

# "Step" linear regression and summary of the best model:
kable(coef(summary(step(lm(mpg~., data), trace=FALSE))))
summary(step(lm(mpg~., data), trace=FALSE))$r.squared

# Residuals diagnostic
par(mfrow=c(2, 2))
plot(step(lm(mpg~., data), trace=FALSE), which=1:4)
```

Appendix 4: Environment

```r
sessionInfo()
```

```
## R version 3.2.3 (2015-12-10)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.10.5 (Yosemite)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] GGally_1.0.1   dplyr_0.4.3    ggplot2_2.1.0  reshape2_1.4.1
## [5] knitr_1.13
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.5      magrittr_1.5     munsell_0.4.3    colorspace_1.2-6
##  [5] R6_2.1.2         highr_0.6        stringr_1.0.0    plyr_1.8.4
##  [9] tools_3.2.3      parallel_3.2.3   grid_3.2.3       gtable_0.2.0
## [13] DBI_0.3.1        htmltools_0.3    lazyeval_0.1.10  yaml_2.1.13
## [17] digest_0.6.9     assertthat_0.1   formatR_1.4      evaluate_0.9
## [21] rmarkdown_0.9.5  labeling_0.3     stringi_1.0-1    scales_0.3.0
## [25] reshape_0.8.5
```