Analysis of MPG by transmission type for Motor Trend - Regression Models Course

Armando Guereca

October 24, 2015

Executive summary

Motor Trend, a magazine about the automobile industry has requested to analyse their data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome).

Our focus is to answer this two questions:

- *Is an automatic or manual transmission better for MPG?*
- *Quantified MPG difference between automatic and manual transmissions*

As demostrated in our analysis we concluded that manual transmission is **better** for mpg than automatic transmission by **1.81 mpg**.

Exploratory data analysis

We start by loading and inspecting our source dataset. Refer to **Appendix A** for details on its format.

```
# This are all libraries that we are going to use
library(knitr); library(ggplot2);

data <- mtcars
# Defined factor columns
data$cyl <- factor(data$cyl)
data$am <- factor(data$am)
levels(data$am) <-c ("automatic", "manual")
data$gear <- factor(data$gear)
data$carb <- factor(data$carb)
```

Some statistics of our subsetted data:

```
# Summarized subset of the columns identified as relevant for this analysis
xt<-summary(data[,c('mpg', 'cyl','am', 'gear','carb')])
kable(xt, format="markdown")   # Output omited for brevity
```

We have then two types of transmision (automatic, manual), an initial visualization of MPG of all vehicles by their transmission type (**Appendix B**) makes evident that manual transmission is better than automatic by at least 5 MPG, although is has higher varibility; If number of cylinders is also considered (**Appendix C**), this variability seems to happen mainly on 4 cylinders manual transmission.

Is also worth noting that car weight is strongly related with the number of cylinders (**Appendix D**), autos of less than 3 tons have 4 cylindes, from 2.5 to 3.5 tons have 6 cylinders and over 3 tons have 8 cylinders. All this factors seem to strongly influence the MPG economy of the automoviles.

Model fitting

Based on conclusions drawned at our exploratory stage would seem naive to define a model based only on MPG VS Transmission type, our initial hypotesis should also consider number of cylinders and weight.

Since there are only 11 dimensions on our dataset, our approach to validate our baseline model and identify the right combination of factors will be to define a model with all variables and *step* through them identifying influencers to prevent overfitting.

```
baseline <- lm(mpg ~ cyl + wt + am, data=data) # Baseline hypotesised model
full <- lm(mpg~., data) # Linear model across all dimentions
fit <- step(full, direction="backward") # Identify the most influential confounder
s.
```

```
summary(fit)$coefficients # Our best model
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## ammanual     1.80921138 1.39630450  1.295714 2.064597e-01
```

This ‘best fitted’ model seems to confirm our exploratory intuitions and improve it by identifing HorsePower as an influential metric, after quantifying this improvement via analisys of variance and adjusted R squared (**Appendix E**) we confirm that this model best fits our analysis needs. Refer to **Appendix F** for the analysis of residuals of our selected model, from it we conclude that no influencing outliers exist out our dataset that might skew our conclusions.

Conclusion

From the coeficients in our best fit model we conclude that manual transmissions are better than automatic by **1.81 MPG**.

This is the diffence between the *Intercept* of the model (which is really de automatic transmission) and the *ammanual* coeficient.

Appendix

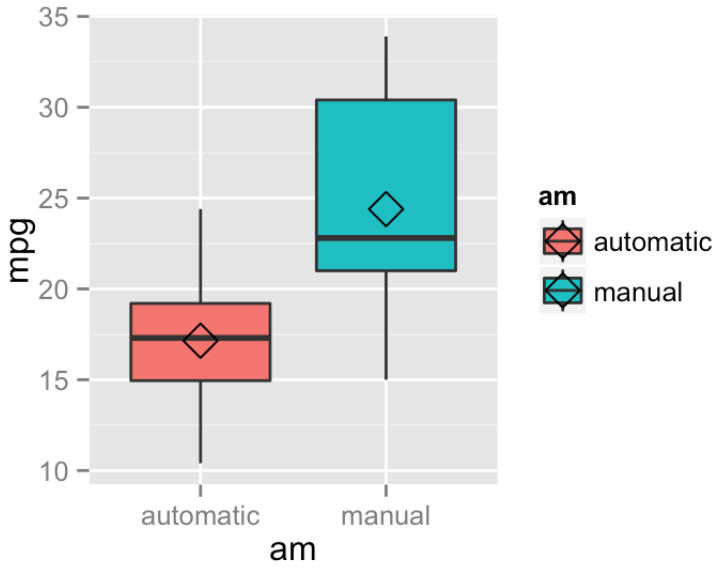
A) Format of our data source:

Documentation of our source data is available on R by typing ***help(mtcars)***. **mtcars** is a data frame with 32 observations on 11 variables.

- [, 1] mpg Miles/(US) gallon
- [, 2] cyl Number of cylinders
- [, 3] disp Displacement (cu.in.)
- [, 4] hp Gross horsepower
- [, 5] drat Rear axle ratio
- [, 6] wt Weight (lb/1000)
- [, 7] qsec 1/4 mile time
- [, 8] vs V/S
- [, 9] am Transmission (0 = automatic, 1 = manual)
- [,10] gear Number of forward gears
- [,11] carb Number of carburetors

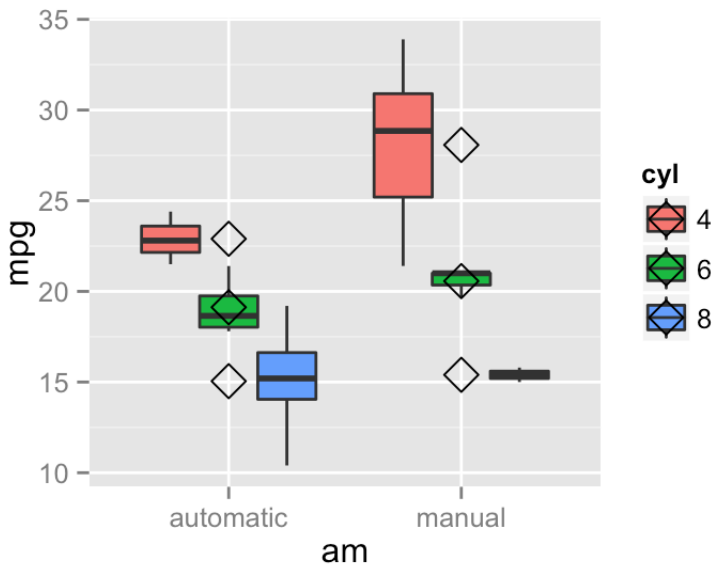
B) MPG by transmission type

```
ggplot(data=data, aes(x=am, y=mpg, fill=am), main="MPG vs Trans Type") + geom_boxp
lot() + stat_summary(fun.y=mean, geom="point", shape=5, size=4)
```



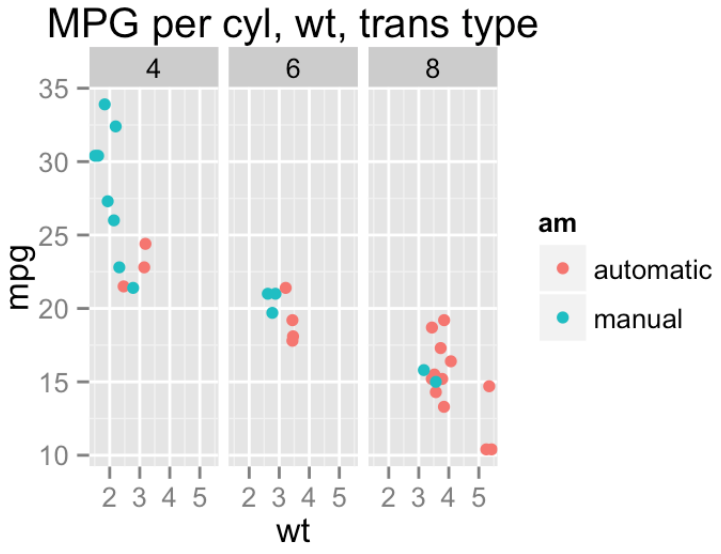
C) MPG by cylinders and transmission type

```
ggplot(data=data, aes(x=am, y=mpg, fill=cyl), main="MPG vs Trans Type") + geom_box
plot() + stat_summary(fun.y=mean, geom="point", shape=5, size=4)
```



D) MPG per cylinders, weight, and transmission type

```
qplot(x=wt, y=mpg, data=data, colour=am, facets=. ~ cyl, main="MPG per cyl, wt, tr
ans type")
```



E) Analysis of Variance: Baseline vs Best fitted models

```
anova(baseline, fit) # Analysis of Variance: Baseline vs Best fitted models
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + wt + am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     27 182.97
## 2     26 151.03  1    31.943 5.4991 0.02693 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Adjusted R squared comparission: Baseline vs Best fitted models
c(summary(baseline)$adj.r.squared, summary(fit)$adj.r.squared)
```

```
## [1] 0.8134405 0.8400875
```

F) Analysis of residuals

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