

# FUEL CONSUMPTION: MANUAL VS AUTOMATIC CARS

## EXECUTIVE SUMMARY

This report aims at investigating the following questions:

1. Is an automatic or manual transmission better for MPG (miles per gallons)
2. Quantify the MPG difference between automatic and manual transmissions

The analysis is carried out on the dataset `mtcars`. Due to the small number of models and the age of the dataset, the results cannot be generalised to the current population of cars, but can give an insight within the set of cars considered.

The sole distinction between manual and automatic cars is a poor predictor of the fuel consumption and other factors more influential have to be taken into account in a more precise prediction. The other reason we need to consider other factors is that in the database, manual cars tend to be imported models generally lighter and smaller, hence with lower fuel consumption.

To better answer the question we therefore propose a model to predict fuel consumption which includes, on top of the a/m distinction, the weight of the car and the number of cylinders. In this context, there is not a statistically significant difference in fuel consumption between manual cars and automatic ones.

## REPORT

The database contains **32** car models of which **19** have automatic transmission and **13** have manual. The variable “am” is **0** for automatic transmission and **1** for manual.

A simple t-test confirm that the two transmissions are not from the same population as the interval does not contain 0 and the p-value **0.0014** is sufficiently low (see *Appendix 2*), therefore there is a statistically significant difference in the “mpg” of the two groups. For the sake of this projects we assume “mpg” to have a normal distribution, although it looks a little skewed.

We consider a linear model with “am” as a predictor and “mpg” as a response (see *Appendix 3*). The model performs poorly: R-squared **0.36** is very low (1 being best fit, 0 being no fit at all) and this, together with a Residual Standard Error (RSE) of **4.822**, suggests that the model is not accurate.

We look therefore at other variables that might be confounders. Let’s call primary the variables that refer to specifications of the car or of the engine (“cyl”, “disp”, “drat”, “wt”, “vs”, “am”, “gear”, “carb”) and secondary the ones that are performance on the car (“mpg”, “hp”, “qsec”) and therefore consequences of the primary ones. We are interesting in the outcome of a secondary variable (“mpg”) with primary variables as predictors. Based on a brief literature review, the variables that can be relevant to our analysis are weight, the number of cylinders and the displacement. The rear axle ratio (“drat”) is also mentioned as having an impact on fuel consumption. The weight and the number of cylinders can tell a lot about a car because they are related to the size and power of the engine and therefore the size and the type of car. Displacement (the amount of fuel burnt per stroke of the engine) gives similar information about the engine but in this case it is correlated with cylinders (see *Appendix 4*) and we therefore use interaction rather than confounding.

We will proceed at setting up 5 nested models, verifying the best fit and then drop redundant variables through an ANOVA test.

A look at the “wt” plot of *Appendix 1* we can imagine two groups of cars, one of light and manual cars and the other of heavy and automatic cars. The plot in *Appendix 5* suggests that there are indeed two groups with different reaction to the increase to weight. Rather than automatic/manuals having different “mpg” trend based on the weight, we suggest that there is a quadratic relationship between “wt” and “mpg”, that is

the slope changes with the weight (steeper on the left side of the graph, flatter on the right). This would explain why mpg in manual cars, that are on the left of the graph, have a steeper slope than automatic ones. Let's have *model1* with the linear predictor "wt" and *model2* where we add "wt"<sup>2</sup>. Indeed *model2* better explains the relationship between "wt" and "mpg", as the R-squared goes from **0.75** to **0.82** with a reduced RSE from **3.05** to **2.65** and improved p-value (see *appendix 6*). In *model 3* we add the variable "cyl". In *model4* we introduce the variable "am", to try to answer the initial question and its interaction with "wt" (see *Appendix 5*). Based on literature review, displacement can have a big impact on fuel consumption, hence in *model5*, we introduce the interaction between "cyl" and "displ", plus the variable "drat". We can now run an anova test to verify which of these variables can be dropped without loss of relevant information.

```
a <- anova(fit_wt, fit_wt2, fit_wt2_cyl, fit_wt2_cyl_am, fit_wt2_cyl_am_displ_drat)
```

The Anova test reveals that the impact of manual/automatic transmission on "mpg", with a p-value of 0.15 is not statistically significant (the variable "am" can be dropped) and that the most relevant variables in the dataset in predicting fuel consumption are the weight and the number of cylinders. We propose *model3* as the most accurate in predicting the fuel consumption based on primary variables, as below:

$$mpg_i = 47.86 + -9.27wt_i + 0.81wt_i^2 + -1.15cyl_i$$

The model explains **0.9%** of the variation in fuel consumption with a residual standard error of **2.4 miles per gallon**. All of the coefficients are significant at 0.05 significant level. As final check we look at the plots of the residuals (see *Appendix 8*) to verify the following assumptions:

1. The variables are independent : the Residuals vs Fitted shows no pattern and confirm independency;
2. Normality of the residuals: the Normal Q-Q plot shows the standardised residuals laying on the line and confirm normality;
3. Constant variance: the Scale-Location plot shows the points randomly distributed and confirm the variance is constant;
4. In the Residuals vs Leverage plot all the points fall within the 0.5 band and confirm that there are no outliers.

## APPENDIX

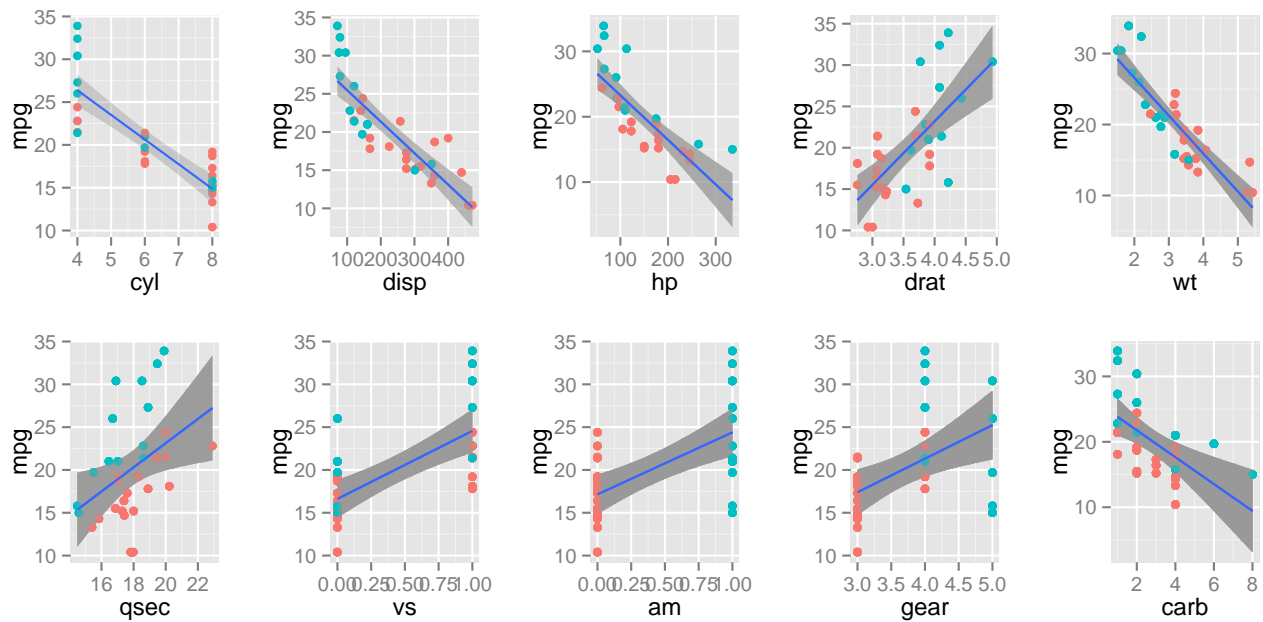
### Appendix 1

#### Exploratory graphs

```
##Print multiple exploratory graphs with ggplot using grid.arrange
names_var <- names(select(mtcars, -mpg))
out <- NULL
g <- ggplot(mtcars, aes(y = mpg))
for(i in 1:length(names_var)){

  g <- g + aes_string(x = names_var[i]) +
    geom_point(aes(colour = factor(am))) +
    geom_smooth(method = "lm", aes(group = 1)) +
    guides(colour = FALSE)

  out[[i]] <- g # creates a list with all the plots
}
grid.arrange(out[[1]], out[[2]], out[[3]], out[[4]],
              out[[5]], out[[6]], out[[7]], out[[8]],
              out[[9]], out[[10]], nrow = 2)
```



## Appendix 2

```
t.test(mtcars$mpg ~ mtcars$am)
```

## Appendix 3

```
fit_am <- lm(mpg ~ am, mtcars)
summary(fit_am)
```

## Appendix 4

```
attach(mtcars)
cor(cyl, disp)
t.test(disp[cyl == 4], disp[cyl == 6])
t.test(disp[cyl == 6], disp[cyl == 8])
summary(lm(disp ~ cyl))
detach(mtcars)
```

## Appendix 5

```
fit_interaction <- lm(mpg ~ am + wt + am*wt, mtcars)
summary(fit_interaction)$coef
plot(mtcars$wt, mtcars$mpg, col = as.factor(mtcars$am))
abline(fit_interaction$coef[1], fit_interaction$coef[3])
abline(fit_interaction$coef[1] + fit_interaction$coef[2],
       fit_interaction$coef[3] + fit_interaction$coef[4], col = "red")
```

## Appendix 6

```
summary(fit_wt)
summary(fit_wt2)
```

## Appendix 7

```
attach(mtcars)
cor(wt, am)
t.test(wt[am == 0], wt[am == 1])
summary(lm(wt ~ am))
detach(mtcars)
```

## Appendix 8

```
par(mfrow=c(2,2))
plot(fit_wt2_cyl)
```

