# Does Transmission Type Have an Impact on Miles per Gallon?

## Executive Summary

Using the mtcars database, I examine the relationship between miles per gallon (MPG) and transmission type (manual vs automatic) for a group of 32 car models released in 1973/74 to address two related questions:

- Is an automatic or manual transmission better for MPG?
- What is the MPG difference between automatic and manual transmissions?

For the cars in the sample, I find that manual cars drive an expected 7.24 MPG more than automatic cars. However, I conclude, after adjusting the model for car weight, that this difference is driven by the fact that, in the sample, almost all manual cars are lighter than almost all automatic cars, with lighter cars having better MPG. For this, and others reasons detailed in the report, I argue that we can't use the sample to infer whether transmission type matters for MPG.

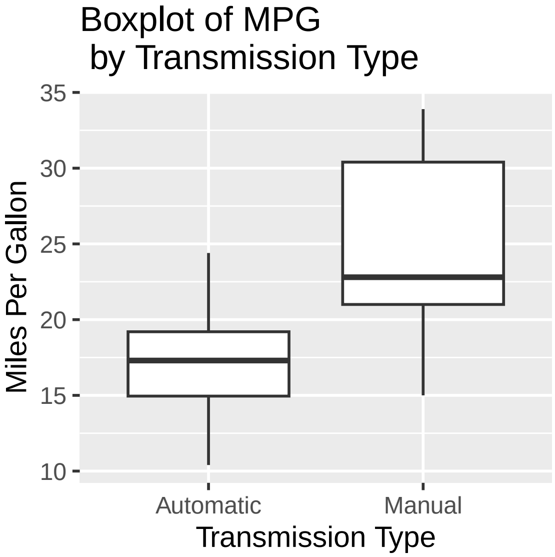
The R code that loads and formats the data, and produces graphs and regression results is in the appendix.

## Data Formatting and Exploratory Analysis

Per the help file (?mtcars), the dataset is from the 1974 Motor Trend magazine, containing 11 variables concerning 32 cars models released in 1973-74.

The only data formatting performed consisted of transforming some of the variables into factors.

Given our guiding questions, it is natural to start with a box plot of MPG values across transmission types:



The box plot raises the hypothesis that vehicles with manual transmissions drive more miles per gallon than vehicles with an automatic transmission. We will examine this hypothesis in more detail in the next section.

## Model Selection, Regression and Analysis

I use a linear model to fit the data. The baseline model is $Y_i = \sum_{k=1}^{p} X_{ki}\beta_k + \epsilon_i$ where $i$ is any value from 1 to 32 (the number of car models), $Y_i$ is the i-th observation of the dependent variable MPG, and p is the number of explanatory variables we include in the model. $X_{1i} = 1$ for all i, which means that its coefficient is the intercept. By assumption, the term error $\epsilon_i$ is iid from a normal distribution with mean 0 (this assumption is, strictly speaking, wrong, since MPG cannot be 0 or negative; however with the range for mpg being distant from 0 (10.4 to 33.4), the assumption seems harmless in practice).

There is a substantial number of models one could fit: we have 10 possible explanatory variables for MPG, some of which are factors (allowing for variable interaction). So, we face the fundamental trade-off with model selection: adding unnecessary variables increases the standard error of the regression variables; omitting relevant variables may lead to biased coefficients. Ideally, this trade-off might be addressed through subject matter knowledge, which is not the case for me: I'm fairly ignorant about car mechanics and engineering.

My strategy for model selection is thus to start with the simplest possible model, which quantifies and tests the relation presented in the box plot: a model with only transmission (variable am) as the predictor of mpg.

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

The intercept tells us that, when the transmission is automatic (am = 0), the expected value of MPG among cars in the sample is 17.15. The slope on am1 indicates that, in the sample, cars with a manual transmission are expected to drive 7.24 more MPG than cars with automatic transmissions (so, 24.39mpg). Given the p-values, we reject the null hypotheses that the coefficients are equal to 0 with very large confidence levels (close to 100%). That is, the coefficients are many standard errors away from 0.

We can now think about adjustments to the model. The concern is that the relation we see in this first model between mpg and transmission type may not actually be the result of transmission type affecting mpg, but of some other variable correlated with transmission being the one that actually affects mpg.

With my limited knowledge about cars, I hypothesized that a car's weight (wt) might influence mpg and fitted a new model with transmission and weight as predictors. If weight is significantly correlated with transmission, the hypothesis would imply that the coefficients of the first model are biased.

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
## am1         -0.02361522  1.5456453 -0.01527855 9.879146e-01
## wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
```

The results are very interesting and confirm the bias of the first model: the effect of transmission type, measured by the coefficient on am1, "disappears:" the coefficient is very close to 0 and, given its p-value, we can no longer reject the null hypothesis that, with all else constant, cars with manual transmissions have the same expected mpg as cars with automatic ones. This can be seen clearly in the figure in the appendix: there is negative relation between a car's weight and its miles per gallon (not surprising; even with my limited car knowledge I'd expect heavier cars to be able to drive fewer miles per gallon of gas consumed). As it happens, in our sample, almost all light cars (weight less than 3000 pounds) are manual and almost all heavy cars (weight above 3000 pounds) are automatic. Therefore, the relation we saw in the first model between transmission and MPG was actually driven by weight: the light cars with better MPG values are mostly manual, and the heavy cars with worse mpgs are mostly automatic. We have almost no cars in the same weight with different transmissions, making the role of transmission impossible to parse.

As a crude measure of whether additional variables might have improved the fit, I ran an ANOVA analysis on the two models above plus a model including all variables in the dataset:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 55.1371 2.129e-06 ***
## 3     15 120.40 14    157.92  1.4053    0.2604
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The test concludes that model 2 (which adds weight to model 1) is necessary for a better fit; it also concludes, given the p-value on model 3 (the fit with all variables) that adding all additional variables is not necessary.

## Residuals and Diagnostics

Please check the appendix for a panel plot with different graphs of residuals for model 2, which is model that best fits the data. The normality of the residuals (top right panel) is “questioned” by the cars in the tail quantiles that deviate from the diagonal line. The leverage graph (bottom right) seems to indicate that the points with higher leverage don’t seem to exert a large amount of influence. Moreover, leverage is limited: no point has leverage larger than 0.25 on a 0 to 1 scale.

## Inference for the Population

There is no indication that the sample of cars under analysis was randomly drawn from the population of cars in existence in 1973-4. For that reason, we can’t really draw inferences for the population. Moreover, even if we could draw such inferences, they would be limited to the period under analysis as technology has changed significantly in the 50 years since the database was assembled. In other words, even if the sample were representative of cars in 1973, it certainly wouldn’t be for cars in 2024.

## Appendix

### Data Loading and Transformation

```r
##load data
require(datasets)
data(mtcars)
## load tidyverse for use across report
library(tidyverse)
## convert cyl (cylinders), vs (engine vshaped or straight), am (transmission), gear (number of gears)
df <- mtcars
cols <- c("cyl", "vs", "am", "gear", "carb")
df <- df %>% mutate(across(all_of(cols), as.factor))
```

### Boxplot of Miles per Gallon by Transmission Type

```
library(ggplot2)
## initiate graph
g <- ggplot(df, mapping = aes(am, mpg))
g <- g + geom_boxplot()
#relabel the values, change the names of the axes and add a title
g <- g + scale_x_discrete(labels = c("0" = "Automatic", "1" = "Manual"))
g <- g + xlab("Transmission Type") + ylab("Miles Per Gallon")
g <- g+ ggtitle("Boxplot of MPG by Transmission Type")
g
```

### Model Fitting and ANOVA test

Model 1: only transmission as a predictor

```
fit1 <- lm(mpg~am, df)
summary(fit1)$coef
```

Model 2: transmission and weight as a predictor

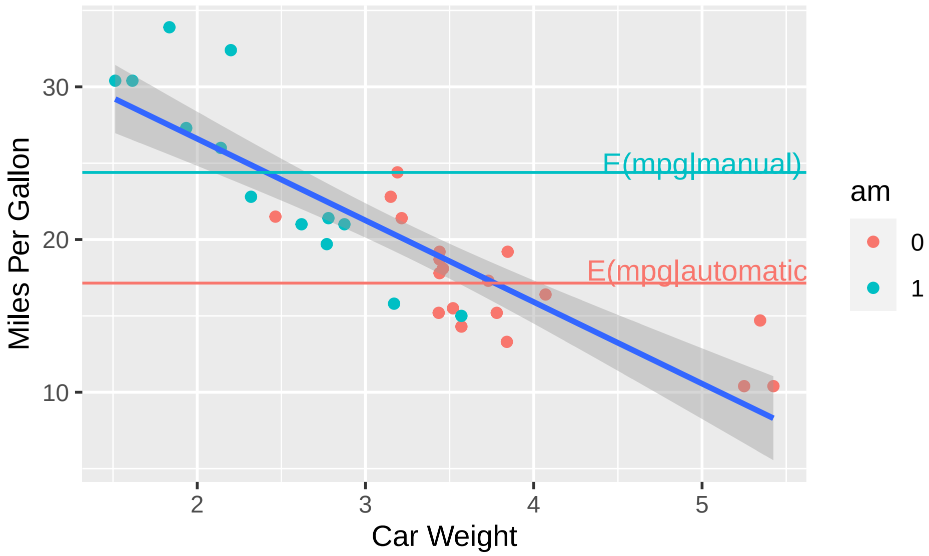
```
fit2 <- lm(mpg~am+wt, df)
summary(fit2)$coef
```

ANOVA on models 1, 2 and a model with all variables

```
fit3 <- lm(mpg~., df)
anova(fit1, fit2, fit3)
```

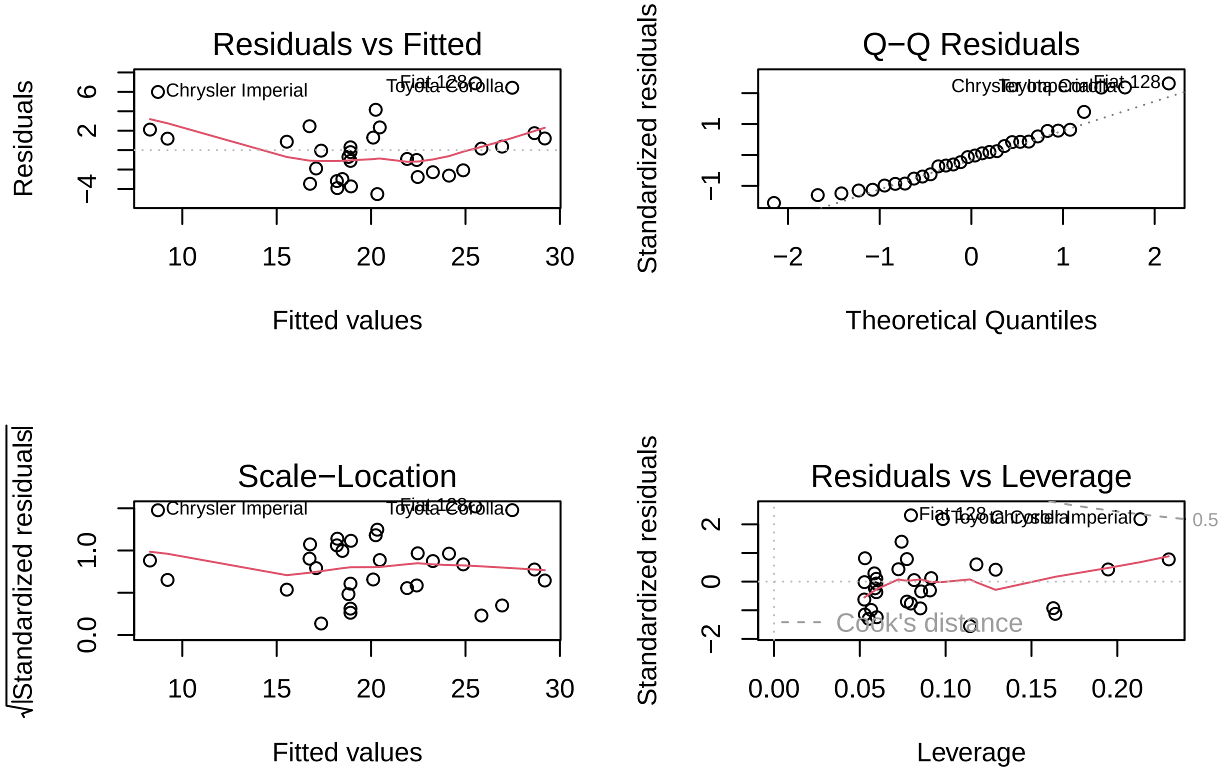
### Relation between Weight and Mpg

```
#initiate plot between wt (x axis) and mpg(y axis)
g1 <- ggplot(df, aes(wt, mpg))
#add points illustrating the relation, but differentiating by color
# depending on transmission type; also add linear fitted line
g1 <- g1 + geom_point(aes(colour = am))+stat_smooth(method = "lm")
# horizontal lines with average MPG for cars with automatic and manual transmissions
g1 <- g1 + geom_hline(yintercept = mean(df$mpg[df$am==0]), color = "#F8766D") +
annotate("text", x=5, y = 18, label = "E(mpg|automatic)", color = "#F8766D")
g1 <- g1 + geom_hline(yintercept = mean(df$mpg[df$am==1]), color = "#00BFC4") +
annotate("text", x=5, y = 25, label = "E(mpg|manual)", color = "#00BFC4")
g1 <- g1 + xlab("Car Weight") + ylab("Miles Per Gallon")
g1
```



### Residuals and Diagnostics

```
par(mfrow = c(2,2))
plot(fit2)
```


Residuals vs Fitted
Residuals
−4
2
6
Chrysler Imperial
Fiat 128
Toyota Corolla
10
15
20
25
30
Fitted values
Q−Q Residuals
Standardized residuals
−1
1
Chrysler Imperial
Toyota Corolla
Fiat 128
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
1.0
Chrysler Imperial
Fiat 128
Toyota Corolla
10
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
−2
0
2
Fiat 128
Toyota Corolla
Chrysler Imperial
0.5
Cook's distance
0.00
0.05
0.10
0.15
0.20
Leverage