Regression Models

Dan Kjeldstrøm Hansen

08 JAN 2017

Executive Summary:

In this analysis I will seek to answer the question:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

The short answer to this question, and based on the following, is: **"Neither" - given the provided data set.** We will see that this is a set of data that very easily could fool the untrained to conclude wrong. A true Data Scientist (or any scientist actually) should always question the results found, and as I will show, it is very important to have domain knowledge, or if one do not have that, to seek it elsewhere. We will also see an example of confounding variables. Finally we can not expect this small dataset to be a good general model base for all car types, as it is biased towards luxury/sporty European cars.

Comparing the mean values of mpg in relation to the type of transmissions seem to be indicating a positive correlation between transmission(am) and miles per gallon(mpg), but whether or not transmission really is a good predictor, is not evident or even intuitive. Many other variables would probably have a higher influence on mpg, even some of those present in the data set.

But the question is **not** about making a good model for estimating mpg, or if any other variables would be better predictors.

Let's begin with a small exploration of the raw data:

Exploratory Analysis

```
data(mtcars)
head(mtcars, 5)
```

```
##                   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  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
```

```
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

The complete data set consist of 32 different car models, with a few different variables measured for each car. (Original data set from *Motor Trend Magazine 1974* and also in *Biometrics, Vol. 37, No. 2. (Jun. 1981, pp. 391-411.)* along with a very in-dept analysis by *H.V.Henderson & P.F.Velleman* of the set and it's value as a predicting model compared to *Hocking (1976)* among other things which is out of this small assignments idea)

This rather small data set, is a typical observational study, and not a well-executed controlled experiment. And even if there seem to be a correlation between mpg and transmission, there is no evidence that the type of transmission has a causation effect on mpg. As a matter of fact it is not intuitive that just because you change the type of transmission you should get any benefit in mpg.

To move an object, you need to add some force in order to overcome all other forces like friction, wind resistance, gravity etc. etc. In the case with these cars, the force that makes them drive is the engine. And the more force you add, the more the car moves, but in order to add force you need to add energy, in this case gallons of fuel - the more force the lower the mpg. An engines "muscles" are more or less equal to it's horsepower, and the amount of energy you need to apply, would have to overcome the existing forces on the car. E.g. the higher weight of the car the more energy you need -> the more fuel you use -> the lower the mpg ...

Let's see how the two type of transmission compares in regards to mpg (all other even, and given the data at hand). I've added a red dashed line indicating the mean of mpg (not to be confused with the medians in the boxes) for both Automatic and Manual transmission. It is clearly visible that on average cars with manual transmission have a higher mpg than cars with automatic transmission. (approx. 7 mpg difference) - **see fig.1**

Let's fit a linear regression model on the mpg/transmission set, and do a Student's t-Test of mpg in relation to the two transmission types as well:

```
fit <- lm(mpg ~ am, data = mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

From the lm we can conclude: There seem to be and statistical significance between mpg and am, actually quite significant ($p = 0.000285$) but am can only vouch for approx. 33.9% of the variances in mpg. By the distribution of the residual around the mean, we can also see that they seem to follow the normal distribution. (symmetric around the mean)

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

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

From the Student's t-test we can conclude: Confirm that cars with automatic transmission has an average mpg value of 17.15 and cars with manual transmission has an average that is 7.25 mpg higher (24.39). And this difference is significant as we also have a very small p-value (0,1%) so we can reject the H0 hypothesis, and thus "conclude" that manual transmission is better than automatic in relation to mpg.

Are we satisfied with the result? Well not quite. With transmission as the only predictor and only vouching for 33,85% of the variance in mpg (adj. R-sqr: 0.3385) it is not a very good model for predicting mpg ...

Let us check for the correlations between mpg and each of the other predictive variables:

```r
round(cor(mtcars$mpg, mtcars), 3)
```

```
##      mpg    cyl   disp     hp  drat     wt  qsec    vs  am gear   carb
## [1,]   1 -0.852 -0.848 -0.776 0.681 -0.868 0.419 0.664 0.6 0.48 -0.551
```

The closer to +/- 1 the better (positive or negative) the correlation (or linear association). Transmission(am) has a correlation of 0.6 which is not very close but still not uncorrelated. The least correlated predictor seems to be the number of forward gears, perhaps not surprising.

Better predictors seem to be weight(wt), cylinders(cyl), displacement(disp) and horsepower(hp). All very intuitive as well. Heavier cars, and more powerful engines.

Let's take a look at the residuals (vertical distances from the regression line - or distance from the estimated y-value): **see fig.2** - the dotted red line indicates the regression line.

The smaller the residuals (the closer they are to the regression line) the better the model. These residuals are +/- 10 and this is a fairly large interval compared to the mpg interval between 10 and 44. Once again an indication that transmission is not a very good indicator for mpg.

You can see that the residuals are nicely spread on each side of the regression line thus indicating that the residuals are normally distributed. We could also conclude homoscedasticity (the variance of the variables is constant). So everything seems to be all-right then? Nah... something is still nagging.

Is it pure coincidence that in this data set it seems that manual transmission is better than automatic, or how can we explain this fact?

Let's see how the ordered weights are distributed, the dotted red line is the mean of weights. **see fig.3**

Surprise! Almost all heavy cars have automatic transmission. This could easily explain why it seems that automatic transmission is worse than manual, whereas in reality, the fact that heavier cars use more gasoline is the ***real reason for the lower mpg.*** - Weight seems to be a confounding variable to am.

It just so happens, that in this selection of cars, almost all heavy cars also have automatic transmission.

A good example of the saying: **Correlation does not imply causation.**

As this example is not about finding the best model, I stop here, concluding that am has no significant impact on mpg.

Further investigation could include: finding the best model based, on this data, re-expression of data, visualising the confounding between am and wt lm(wt~am,mtcars) and you'll find a very strong relationship between am and wt.

Appendices:

fig.1

```
CarCol <- c("blue", "green")
boxplot(mtcars$mpg ~ factor(mtcars$am, labels = c("Automatic", "Manual")), col = CarCol,
    xlab = "transmission", ylab = "mpg")
abline(h = mean(mtcars$mpg[mtcars$am == 0]), col = "red", lty = 2)
abline(h = mean(mtcars$mpg[mtcars$am == 1]), col = "red", lty = 2)
```

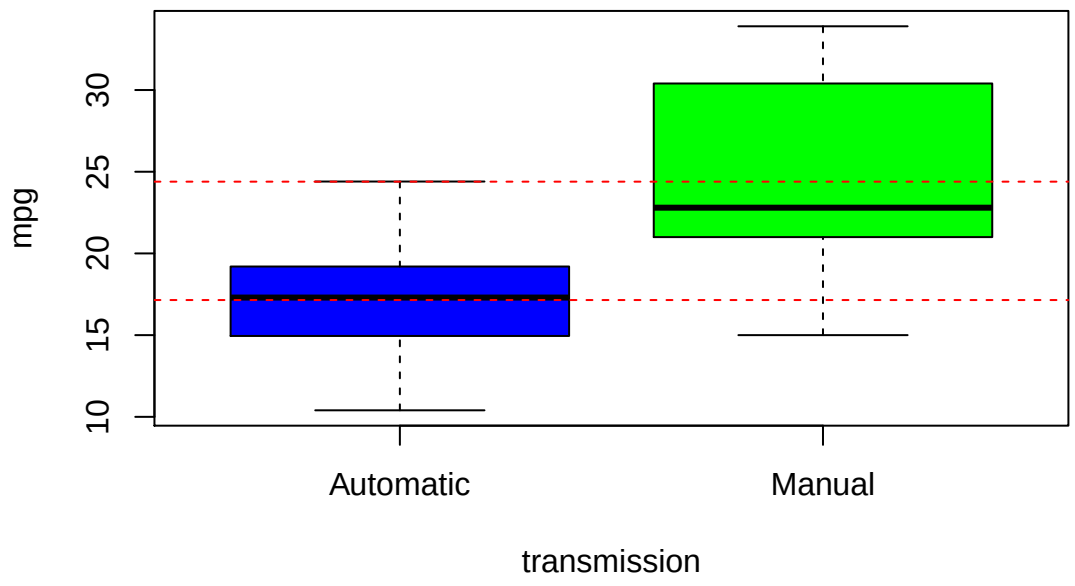


fig.2

```
CarColFact <- c("blue", "green")[as.factor(mtcars$am)]
plot(summary(lm(mpg ~ am, mtcars))$resid, col = CarColFact, pch = 19, xlab = "car# and am",
    ylab = "residuals", main = "Residuals of mpg / car and am")
abline(h = 0, col = "red", lty = 2)
legend("topleft", legend = c("Automatic", "Manual"), pch = 19, bty = "n", col = CarCol,
    cex = 0.75)
```

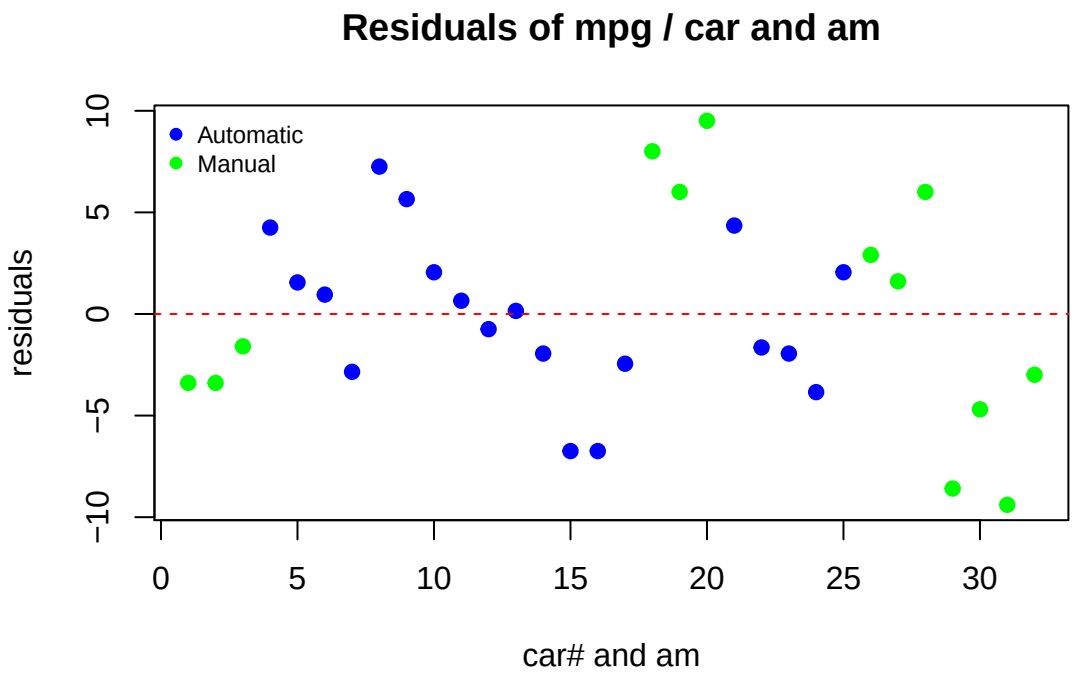
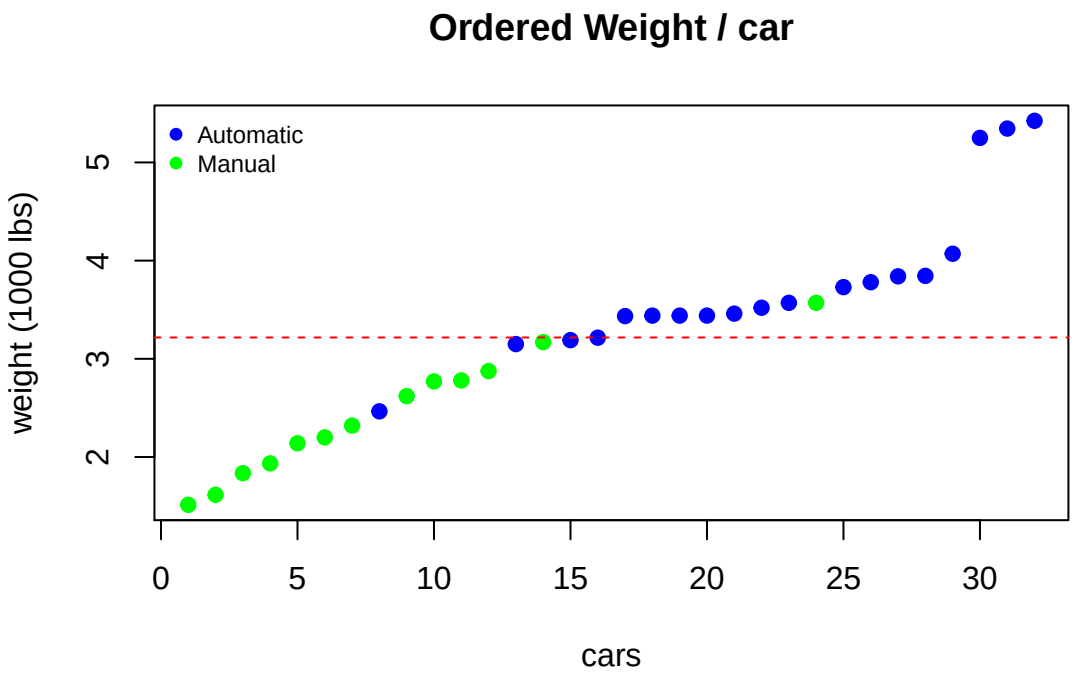


fig.3

```
plot(mtcars$wt[order(mtcars$wt)], col = c("blue", "green")[as.factor(mtcars$am[order(mtcars$wt)])],
    pch = 19, xlab = "cars", ylab = "weight (1000 lbs)", main = "Ordered Weight / car")
abline(h = mean(mtcars$wt), col = "red", lty = 2)
legend("topleft", legend = c("Automatic", "Manual"), pch = 19, bty = "n", col = CarCol,
    cex = 0.75)
```


Ordered Weight / car
Automatic
Manual
weight (1000 lbs)
2
3
4
5
0
5
10
15
20
25
30
cars