Is transmission a key to better car mileage?

Mogens Yde-Andersen

Executive Summary

In the years after 1st Oil Crisis many americans asked themselves: "Is an automatic or manual transmission better for the Miles Per Gallon performance of my next car? A "flat" comparison between two groups of 1973-4 car models with automation gear or with manual gear shows that "manuals" go x,x miles per gallon more than "automatics." But that is far from the whole story.

Transmission as THE explaining factor for better MPG performance is wrong. 4 different data model suggests entirely different outcome regressor connections, which makes it impossible to say something definitive about transmission as factor for the mileage outcome.

One model suggests that manual gear cars yields better mileage over automatic egar cars. Another data model ignores transmission, and two 2 data models contains transmission as a factor and cannot reject that manual gear cars are better for mileage than automatic cars, but they can not reject a draw neither.

And it all depence on an acceptance of a shift fra automatic transmission to manual transmission as a so called "one unit increase in transmission as a discrete variable""

Automatic gear cars versus manual gear cars

We have access to the mtcars car data set with 11 variables including MPG and transmission type. By taking the mean of MPG for all automatic gear cars ("automatics") and for all manual gear cars ("manuals"), manuals (men of 24,4 MPG) outperforms automatics (mean of 17.1) by a difference of 7,2 MPG. This is done with a one-dim. t-test comparison. See the difference in figure 1 in appendix (app), sa well as the calculated data i app A.

Is transmission really the true key question?

No. Many factors play a part of the collected performance of the MPG outcome. But how important is transmission then?

By fitting all variables into a all encompassing linear regression-data model (the "All in"" model), the data set itself shows each variables importance as a causal factor for the outcome MPG. The calculation is in app B shows what one unit increase for a given variabel will give in changed outcome MPG in the Estimat column. The "All in" data model and equation: Outcome MPG = 12,3 -(0,11*cylinders) +(0.01*displacement) -(0.02*horsepower) +(0.79*rearaxleratio) -(3.72*weight) +(0.821/4miletime) +(0.32*V/S) +(2.52*transmission) +(0.65*gear) -(0.20*carburetor

The "All in"" data model tells us that shifting from "automatic" to "manual" will increase mileage by 2,52 MPG with a standard error of 2.06. An increase with 95% certainty will fall within an interval of [-0.86,5.90]in increased MPG. A 1000 lbs. increase in weight impacts outcome MPG negatively by 3.71 MPG. Accordingly it is possible to see in app B, how much one unit change in the variable will change outcome MPG. So transmission is far from being the main explaining factor behind better mileage.

Data models for prediction of mileage

The "All in" data model contains as many as 10 variables without considering correlation/redundance between the variables and risk of unnecessary variance inflation

A objective tree analysis can crystallize the variables that differentiate the data points the most with regard to the outcome MPG. See fig. 2 i app. It reveals weight, cylinders and horse power as the three most influential factors. I call it the "Top 3 differentiator" data model. See summary in app C. Top 3 differentiator data model and equation: Outcome MPG = 38.75 -(0.94*cylinders) -(0.02*horsepower) -(3.17*weight)

The "Build up"-model is built up (manually) from bottom by putting the one variables at a time in a "empty" model to see, which "1-variabel"-model yields the best (lowest) model P-value. That is weight. Then weight is combined with one other variabel, one at a time into a "2-variables"-model. The lowest P-value yielding combination is weight + cylinders.

This winning combination is then combined with one new variable into a "3-variables"-model with each of the remaining variables, each at a time. The winning "3 variables"-model combination is weight + cylindres + horsepower. hi T ...and so on.

The winning "5 variables"-model is weight + cylinders + horse power + transmission and 1/4 mile time. Se summary in app D. The "Build up" data model and equation: Outcome MPG = 19.35 -(3.15*weight) -(0.15*cylinders) -(0.02*horsepower) +(2.73*transmission) + (0.74*"1/4miletime)

The final model is the "Leave out"-model" that pulls out the most unwanted variables due to their high P-value contribution - one at a time. You start with the "All in"-model. You scan the variables' P-values. Cylinders has the highest P-value and is removed. The model is refitted. Next variable to remove is V/S due to highest P-value. And so on. I stopped, when the model contains 5 variables due to a relatively low P-value score and highest adjusted residuals second to the power score for the model and by having not too few and neither too many variables. Se summary in app E and code for the "Leave out"" model in app F. The leave out data model and equation: Outcome MPG = 14,36 + (0.01*displacement) -(0.02*horsepower) -(4.08*weight) +(1.01*1/4 miletime) +(3.47*transmission)

Note that transmission does not appear to be among the top 3 big differentiators, but transmission does show up on a 4^th or 5th place in the two ladder models.

What the data models show

	X1	X2	X3	X4	X5
l1	Model	Variables	Transm. impact on MPG	Std.Error	95% conf. int.
l2	All in	All 10 variables	2,52 MPG	2.06 MPG	[-0.86,5.90] MPG
l3	Top 3 diff.	Weight Cyl HP	NA	NA	NA
l4	Build up	Weight Cyl HP trans. 1/4 mile time	2.73 MPG	1.72 MPG	[-0.11,5.57] MPG
l5	Leave out	Weight Disp. HP Transm. 1/4 mile time	3.48 MPG	1.49 MPG	[1.03,5.91] MPG

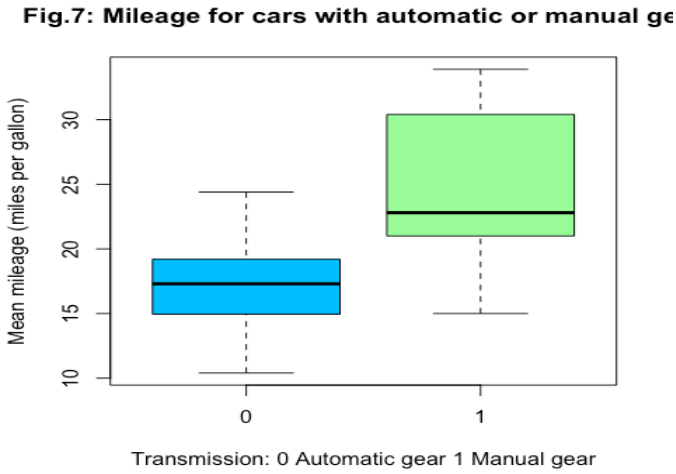
The "All in"" data model and the "Build up"" data model fail to reject the claim about the increase in MPG outcome due to a shift to "manuals" from "automatics." It can not be ruled out either that it is not the case due to the appearance of "0" in the confidence interval. The Top 3 differentiator data model does not say anything about a connection between transmission and outcome MPG. The "Leave out" data model fail to reject that "manuals" instead of "Automatics" lead to a positive change in outcome MPG. It is supported by the purely positive confidence interval.

The models come with one major misrepresentation of facts. In the model automatic gear cars has the 0 at the transmission value, where as manual gear cars takes the value 1. All 4 models build on the assumption that an increase in a given variable impacts the outcome mileage in MPG. But can you say that going from transmission type automatic gear to manual gear is an increase? If you are willing to neglect this qualitative error, go on. Otherwise stop reading here.

The robustness of the models also depends how the residuals behave. Correlation can induce variance inflation into models. See app H-I. The "All in" model does have that. Mainly on the variable displacement. The "Top 3 differentor model does not seem to have any worthy to mention. The "Build up" model may have some on the cylindre variable and the "Leave out" model on the displacement variable. Se code in app. The so called anova function indicates, whether the data models have too few, too many or a suitable number variables. See app G for the code to run for checking the residuals and do the diagnostics. "All in has somewhat of. Namely the displacement variable. The Top 3 differentiator"" data model may have too few variables. The "1/4 mile time" variable semms to be one variable too many in the "Build up" data model. The "Leave out" data model should perhaps only contain of weight, 1/2 mile time, transmission and horse power. And finally, the regression models are linear without interaction or polynomia. A better data model and a better representation of the truth may be out there.

Appendix

Appendix Figure 1: "Mileage for cars with automatic or manual gear"



Appendix A "T-test of mileage of manual gear cars vs. automatic gear cars"

```
##
##  Welch Two Sample t-test
##
## data:  mtcars_man$mpg and mtcars_auto$mpg
## 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:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

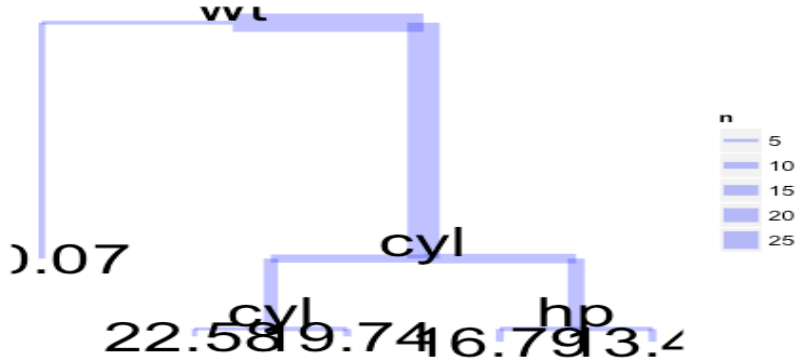
Appendix B Summary of the "All in" data model

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs +
##     as.factor(am) + gear + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.30337   18.71788   0.657   0.5181
## cyl            -0.11144    1.04502  -0.107   0.9161
## disp            0.01334    0.01786   0.747   0.4635
## hp             -0.02148    0.02177  -0.987   0.3350
## drat            0.78711    1.63537   0.481   0.6353
## wt             -3.71530    1.89441  -1.961   0.0633 .
## qsec            0.82104    0.73084   1.123   0.2739
## vs              0.31776    2.10451   0.151   0.8814
## as.factor(am)1  2.52023    2.05665   1.225   0.2340
## gear            0.65541    1.49326   0.439   0.6652
## carb           -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```
## [1] -0.8626582  5.9031182
```

Appendix Figure 2 Tree dendrogram diagram of the mtcars data set

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```



Appendix C Summary of the "Top 3 differantiator" data model

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9290 -1.5598 -0.5311  1.1850  5.8986
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
## wt          -3.16697    0.74058  -4.276 0.000199 ***
## cyl         -0.94162    0.55092  -1.709 0.098480 .
## hp          -0.01804    0.01188  -1.519 0.140015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.512 on 28 degrees of freedom
## Multiple R-squared:  0.8431, Adjusted R-squared:  0.8263
## F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11
```

Appendix D Summary of the "Build up" data model

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + hp + as.factor(am) + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5346 -1.5307 -0.1674  1.2298  4.6108
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    19.34710   13.40951   1.443  0.16102
## wt             -3.15167    1.00274  -3.143  0.00415 **
## cyl            -0.14876    0.73915  -0.201  0.84206
## hp             -0.01692    0.01486  -1.138  0.26529
## as.factor(am)1  2.72942    1.72428   1.583  0.12553
## qsec            0.73822    0.57363   1.287  0.20946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.479 on 26 degrees of freedom
```

```
## Multiple R-squared:  0.8581, Adjusted R-squared:  0.8308
## F-statistic: 31.44 on 5 and 26 DF,  p-value: 3.103e-10
```

Appendix E Summary of the "Leave out" data model

```
##
## Call:
## lm(formula = mpg ~ disp + hp + wt + qsec + as.factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5399 -1.7398 -0.3196  1.1676  4.5534
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    14.36190    9.74079   1.474  0.15238
## disp            0.01124    0.01060   1.060  0.29897
## hp             -0.02117    0.01450  -1.460  0.15639
## wt             -4.08433    1.19410  -3.420  0.00208 **
## qsec            1.00690    0.47543   2.118  0.04391 *
## as.factor(am)1  3.47045    1.48578   2.336  0.02749 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.429 on 26 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8375
## F-statistic: 32.96 on 5 and 26 DF,  p-value: 1.844e-10
```

Appendix F Code for the "Leave out" iterated down to three variables

```
data(mtcars)
all <- lm(mpg ~ ., data = mtcars);summary(all);ar_full <- summary(all)$adj.r.squared
# Leaveout one variable at a time
lm_k9 <- update(all, ~. -cyl );summary(lm_k9);ar_k9 <- summary(lm_k9)$adj.r.squared
lm_k8 <- update(lm_k9, ~. -vs);summary(lm_k8);ar_k8 <- summary(lm_k8)$adj.r.squared
lm_k7 <- update(lm_k8, ~. -carb);summary(lm_k7);ar_k7 <- summary(lm_k7)$adj.r.squared
lm_k6 <- update(lm_k7, ~. -gear);summary(lm_k6);ar_k6 <- summary(lm_k6)$adj.r.squared
lm_k5 <- update(lm_k6, ~. -drat);summary(lm_k5);ar_k5 <- summary(lm_k5)$adj.r.squared
lm_k4 <- update(lm_k5, ~. -disp);summary(lm_k4);ar_k4 <- summary(lm_k4)$adj.r.squared
lm_k3 <- update(lm_k4, ~. -hp);summary(lm_k3);ar_k3 <- summary(lm_k3)$adj.r.squared
```

Appendix G Correlation between all variables in mtcars data set

```
data(mtcars);cor(mtcars)
```

Appendix H Variance inflation in the 4 data models

```
## Warning: package 'car' was built under R version 3.1.3
```

Appendix I Code for residual Variance Estimation in 3 of the data models (anova function and detailed residual plots)