Does Transmission effect MPG?

Rudy Veenhoff

20 oktober 2016

Executive Summary

Is this report we investigate whether transmission type (automatic or manual) has any influence in the number of miles a car can drive per gallon using the data set mtcars. We find that MPG can be sufficiently predicted by the weight of the car and the number of cylinders it has. In view of this, transmission type does not seem to significantly impact MPG.

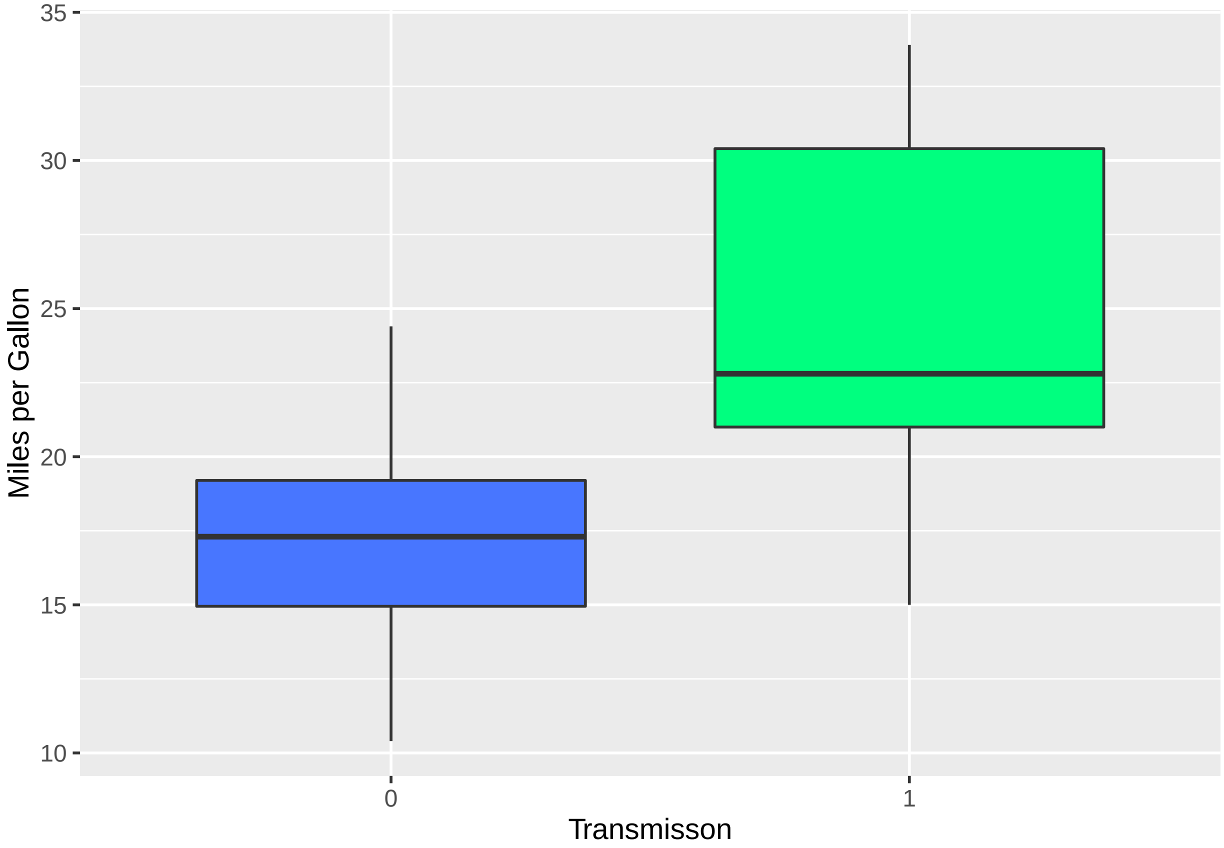
Exploratory analysis

```
head(mtcars)
```

```
##                    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
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

From the help file we learn that the 'am' column stands for transmission (0 = automatic, 1 = manual).

```
library(ggplot2)
mtcars$am <- as.factor(mtcars$am) ## just to get a nicer plot
g <- ggplot(data=mtcars,aes(x=am,y=mpg))
g <- g + geom_boxplot(fill=c("royalblue1","springgreen"))
g <- g + xlab("Transmisson") +ylab("Miles per Gallon")
g
```



Cars with an automatic transmission seem to have lower mpg compared to cars with manual transmission when viewed in a vacuum. However, this might be too simple an explanation. There might be other variables confounding the relation between mpg and transmission. A variable that might confound might be the weight. It seems reasonable to assume that heavier cars have a higher fuel consumption, hence they travel less miles per gallon.

Model Selection

We wish to find a linear model for the outcome mpg. Which regressors should we choose? Let's find out how correlated the variables are. We chose to show our correlation matrix in the appendix as to to keep our report clean and less of a data dump (See Appendix, section Correlation Matrix).

Almost all variables seem to be moderately to strongly correlated with mpg, however, they also seem highly correlated with each other. The ANOVA method seems appropriate in this situation.

Starting from our base model $\text{mpg} = \beta_0 + \beta_1 \text{am}$, we will create a sequence of nested models by linearly adding all the regressors. Note that the order in which we add correlated regressors is of importance. If a regressor turns out to be significant; its correlated regressors will likely be insignificant. We also check whether the model residuals are approximataly normal with the Shapiro-Wilk test.

See Appendix, section ANOVA for the results.

With a p-value of 0.05, we decide to include the regressors cyl and wt. Our model is:

$$\text{mpg} = \beta_0 + \beta_1 \text{am} + \beta_2 \text{cyl} + \beta_3 \text{wt}$$

How well is the fit?

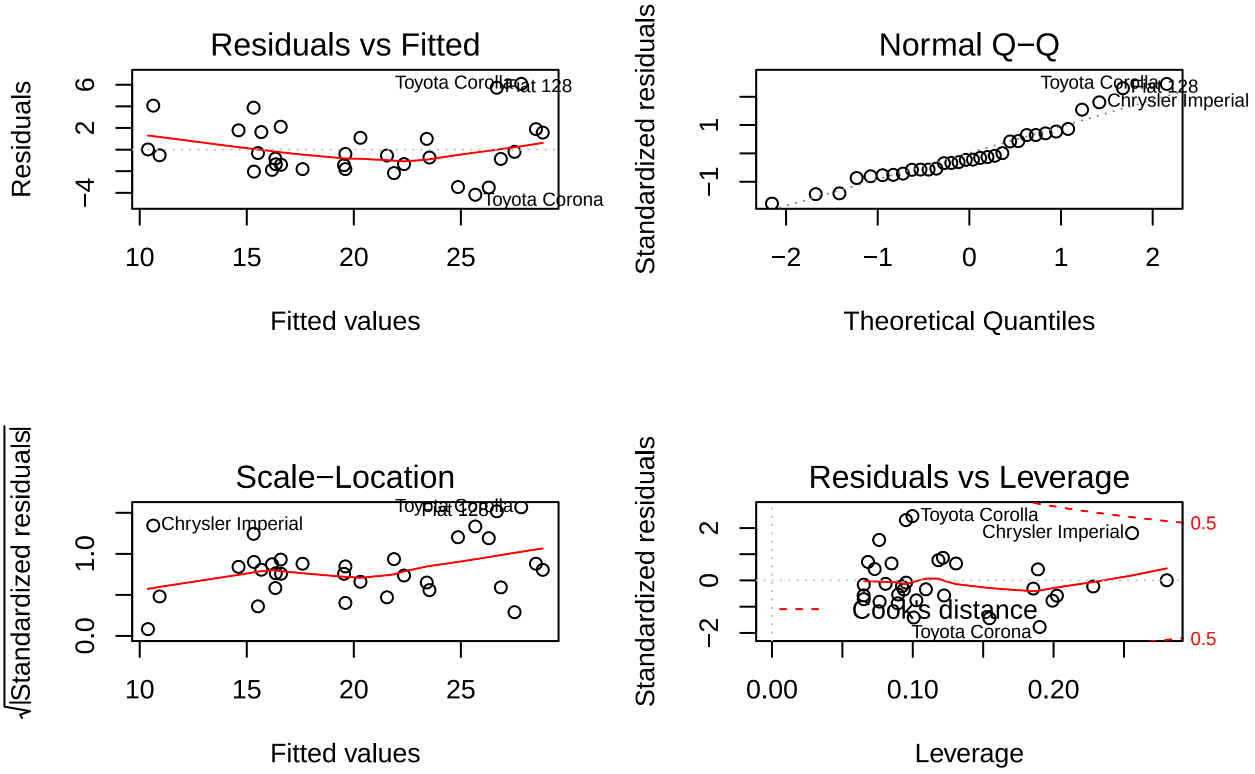
We will now check to see how well our model fits the data by looking at R^2 and the residual plots.

```
mdl<- lm(mpg~am+cyl+wt,data=mtcars)
summary(mdl)$r.squared
```

```
## [1] 0.8303383
```

About 83% of the variance is explained by our model. A decent number considering we have a relatively small amount of predictors.

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



Our residuals seem nicely random.

```
summary(mdl)$coef
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 39.4179334  2.6414573 14.9227979 7.424998e-15
## am1          0.1764932  1.3044515  0.1353007 8.933421e-01
## cyl         -1.5102457  0.4222792 -3.5764148 1.291605e-03
## wt          -3.1251422  0.9108827 -3.4308942 1.885894e-03
```

Here's we notice something is wrong. Our estimated transmission coefficient turns out to confidently not reject the null hypothesis that it's different from 0. To make things worse; it seems to be centered around 0. This makes us believe that mpg isn't dependent on transmission at all. The other regressors are highly significant though.

Automatic or Manual?

From the data we've fitted a linear model for the outcome mpg with the predictors cyl, wt, and am. Which regressors to include was decided by the ANOVA tests. For the estimated transmission coefficient we found value of 0.17 and standard error of 1.3. The test with null hypothesis $\beta_1 = 0$ fails the reject convingly (p > 0.89). We conclude that for this model the transmission type is negligible for predicting the amount miles per gallon. Our quantification for the difference between mpg for automatic and manual transmission cars would be 0.

Appendix

Correlation matrix

```r
mtcars$am<-as.numeric(mtcars$am) ## Changing it back again
cor(mtcars)
```

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

ANOVA

```r
fit1 <- lm(mpg~am,data=mtcars)
fit2 <- update(fit1,mpg~am+cyl,data=mtcars)
fit3 <- update(fit2,mpg~am+cyl+disp,data=mtcars)
fit4 <- update(fit3,mpg~am+cyl+disp+drat,data=mtcars)
fit5 <- update(fit4,mpg~am+cyl+disp+drat+wt,data=mtcars)
fit6 <- update(fit5,mpg~am+cyl+disp+drat+wt+qsec,data=mtcars)
fit7 <- update(fit6,mpg~am+cyl+disp+drat+wt+qsec+vs,data=mtcars)
fit8 <- update(fit7,mpg~am+cyl+disp+drat+wt+qsec+vs+gear,data=mtcars)
fit9 <- update(fit8,mpg~am+cyl+disp+drat+wt+qsec+vs+gear+carb,data=mtcars)
anova(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8,fit9)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + drat
## Model 5: mpg ~ am + cyl + disp + drat + wt
## Model 6: mpg ~ am + cyl + disp + drat + wt + qsec
## Model 7: mpg ~ am + cyl + disp + drat + wt + qsec + vs
## Model 8: mpg ~ am + cyl + disp + drat + wt + qsec + vs + gear
## Model 9: mpg ~ am + cyl + disp + drat + wt + qsec + vs + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 64.0801 5.842e-08 ***
## 3     28 252.08  1     19.28  2.7485  0.111541
## 4     27 252.03  1      0.05  0.0076  0.931103
## 5     26 188.40  1     63.63  9.0700  0.006419 **
## 6     25 160.46  1     27.94  3.9832  0.058482 .
## 7     24 160.17  1      0.29  0.0409  0.841618
## 8     23 159.95  1      0.22  0.0309  0.862047
## 9     22 154.33  1      5.62  0.8008  0.380528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
shapiro.test(fit2$residuals);shapiro.test(fit5$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit2$residuals
## W = 0.97711, p-value = 0.7123
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit5$residuals
## W = 0.94745, p-value = 0.1219
```

The test gives no reason to reject normality.