The effect of transmission on MPG

Willy Bakker

Source on GitHub (http://github.com/FrieseWoudloper/Coursera_Regression_Models/blob/master/The_effect_of_transmission_on_MPG.RMD)

Executive Summary

The editors of Motor Trend, a magazine about the automobile industry, are interested in exploring the relationship between a set of variables and miles per gallon (MPG). They are particularly interested in the following two questions:

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

Motor Trend has a data set of a collection of cars comprising MPG and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). Simple linear analysis showed an effect of transmission type on MPG. The presence of a manual transmisson led to an expected increase of 7.2 miles per gallon. However, when we adjusted the effect of transmission for weight, number of cylinders and horse power, the effect proved to be non-significant.

The data set

The data set has 32 observations on 11 variables. Miles/gallon (mpg), displacement (disp), gross horsepower (hp), weight (wt), rear axle ratio (drat) and 1/4 mile time (qsec) are quantitative variables. Number of cylinders (cyl), number of forward gears (gears) and number of carburetors (carb) are ordinal variables. Type of engine (vs) and transmission (am) are nominal variables. Ordinal and nominal variables are treated als factor variables in the data analysis.

Exploratory data analysis

The boxplot in appendix 1 reveals that the mean MPG of automatic transmission cars in the data frame is lower than that of manual transmission cars. The correlation matrix in appendix 2 gives an idea of the degree and type of relationships between the variables in the data set. Note that the matrix consists of Pearson's correlation coefficients, which may not be the best choice for ordinal and nominal variables. Weight, number of cylinders, displacement and gross horse power are highly correlated with MPG. The absolute correlation coefficient is (almost) 0.80 or higher. The plots in appendix 3 visualize the relationships between the outcome variable MPG and the highest correlated predictor variables, differentiated by transmission. It appears that for displacement and weight there may be an interaction with transmission.

The unadjusted effect of transmission on MPG

The effect of transmission on MPG is estimated using a simple linear regression model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where ϵ_i are iid $N(0, \sigma^2)$ with transmission as the predictor and MPG as the outcome. No other covariates are included in the model. According to this simple linear regression model, there is an increase of 7.2 miles per gallon if the car has a manual transmission. MPG differs significantly between automatic and manual transmission (p-value = 0.0003). However, these are unadjusted estimates. They may be biased! The effect of transmission on MPG explains about 34 % of the variation in MPG (adjusted R-squared). So, the greater part of the variation in MPG remains unexplained.

The adjusted effect of transmission on MPG

The adjusted effect of transmission on MPG is also estimated using a multivariable regression model:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i = \sum_{k=1}^{p} X_{ik}\beta_j + \epsilon_i$$

where ϵ_i are iid $N(0, \sigma^2)$ with transmission as X_{1i} and MPG as Y_i.
The model is constructed by adding the four predictors with the highest correlation coefficient step-by-step to the simple linear model. These are weight, displacement, number of cylinders and horse power. Analysis of variance (ANOVA) is used to determine which predictors are kept in the model. Transmission is kept in every step, even if the coefficient is not significantly different from zero, because of the main research questions. Not only are the main effects added to the model and tested for, but also the interaction between transmission and weight, because the scatter plot in appendix 3 showed there might be a significant interaction effect between MPG and weight.

See appendix 4 for the results of the ANOVA. Displacement seems to be a confounder for weight (also see the scatter plot in appendix 3). Adding displacement, does not significantly improve the model fit. Likewise for adding the interaction of transmission and weight. In the resulting model the outcome MPG is best predicted by the main effects of transmission, weight, number of cylinders and horse power. About 84 % of the variation in MPG is explained by this multivariable model. According to the multivariable model there is an expected increase of 1.8 miles per gallon if the car has a manual transmission, assuming all other predictors (weight, number of cylinders and horse power) are kept constant. However, this adjusted effect of transmission on MPG is not significant (p-value = 0.2065).

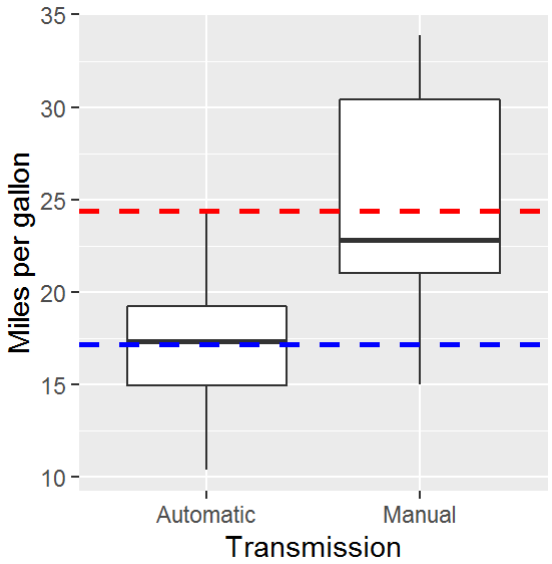
Testing the assumption of normality

ANOVA is sensitive to its assumption that model residuals are approximately normal. If they are not, we could get small p-values for that reason. The Shapiro-Wilk test is used to test for normality. Normality is its null hypothesis. The Shapiro-Wilk p-value of 0.4479 fails to reject normality, supporting confidence in our analysis of variance.

Residual plot

Residuals should be uncorrelated with the fit, and should be independent and (almost) identically distributed with mean zero. The residual plot in appendix 5 shows these assumptions hold. The mean of the residuals (blue dashed line) is very close to zero. The residuals seem uncorrelated to the fit and independent and identically distributed.

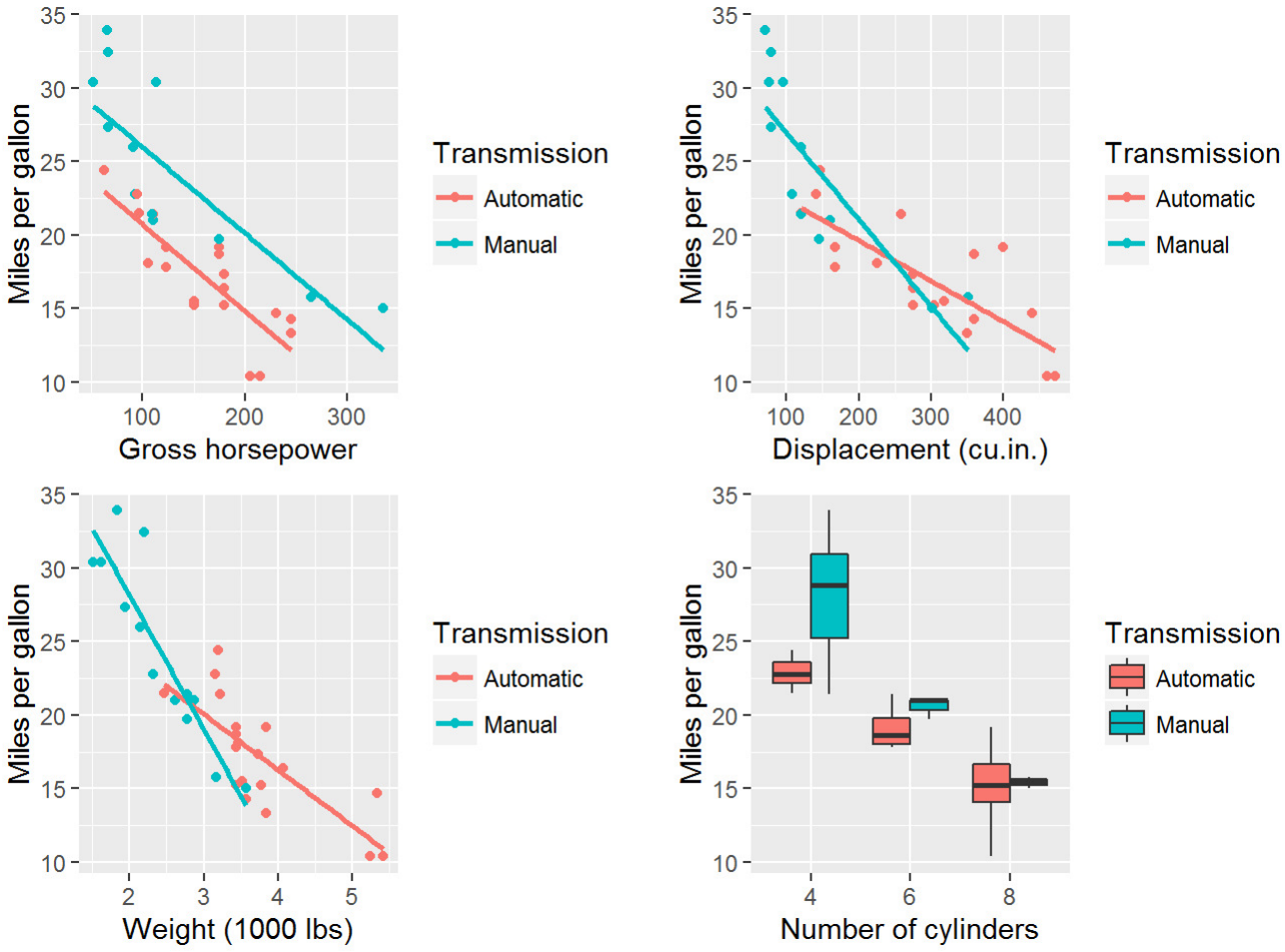
Appendix 1: MPG by transmission



Appendix 2: Correlation matrix

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.85	-0.85	-0.78	0.680	-0.87	0.420	0.66	0.600	0.48	-0.550
cyl	-0.85	1.00	0.90	0.83	-0.700	0.78	-0.590	-0.81	-0.520	-0.49	0.530
disp	-0.85	0.90	1.00	0.79	-0.710	0.89	-0.430	-0.71	-0.590	-0.56	0.390
hp	-0.78	0.83	0.79	1.00	-0.450	0.66	-0.710	-0.72	-0.240	-0.13	0.750
drat	0.68	-0.70	-0.71	-0.45	1.000	-0.71	0.091	0.44	0.710	0.70	-0.091
wt	-0.87	0.78	0.89	0.66	-0.710	1.00	-0.170	-0.55	-0.690	-0.58	0.430
qsec	0.42	-0.59	-0.43	-0.71	0.091	-0.17	1.000	0.74	-0.230	-0.21	-0.660
vs	0.66	-0.81	-0.71	-0.72	0.440	-0.55	0.740	1.00	0.170	0.21	-0.570
am	0.60	-0.52	-0.59	-0.24	0.710	-0.69	-0.230	0.17	1.000	0.79	0.058
gear	0.48	-0.49	-0.56	-0.13	0.700	-0.58	-0.210	0.21	0.790	1.00	0.270
carb	-0.55	0.53	0.39	0.75	-0.091	0.43	-0.660	-0.57	0.058	0.27	1.000

Appendix 3: Scatter plots



Appendix 4: ANOVA

```
fit1 <- lm(mpg ~ factor(am), data = mtcars)
fit2 <- lm(mpg ~ factor(am) + wt, data = mtcars)
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 46.115 1.867e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit3 <- lm(mpg ~ factor(am) + wt + disp, data = mtcars)
anova(fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) + wt
## Model 2: mpg ~ factor(am) + wt + disp
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     29 278.32
## 2     28 246.56  1    31.763 3.6072 0.06788 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit4 <- lm(mpg ~ factor(am) + wt + factor(cyl), data = mtcars)
anova(fit2, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) + wt
## Model 2: mpg ~ factor(am) + wt + factor(cyl)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     29 278.32
## 2     27 182.97  2    95.351 7.0353 0.003473 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit5 <- lm(mpg ~ factor(am) + wt + factor(cyl) + hp, data = mtcars)
anova(fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) + wt + factor(cyl)
## Model 2: mpg ~ factor(am) + wt + factor(cyl) + hp
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     27 182.97
## 2     26 151.03  1    31.943 5.4991 0.02693 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit6 <- lm(mpg ~ factor(am) + wt*factor(am)+ factor(cyl) + hp, data = mtcars)
anova(fit5, fit6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) + wt + factor(cyl) + hp
## Model 2: mpg ~ factor(am) + wt * factor(am) + factor(cyl) + hp
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     26 151.03
## 2     25 130.47  1    20.554 3.9384 0.05828 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 5: Residuals versus fitted