Effects of vehicle transmission type on miles per gallon values

Istvan Andras Horvath

December 18th, 2018

Synopsis

We use linear regression models on the *Motor Trend Car Road Tests (mtcars)* dataset to identify whether or not there is any impact of a car's transmission type (automatic or manual) on how far it can travel with one US gallon of fuel.

First we overview the *mtcars* dataset, do some minor data transformations and explore the relevant part of the dataset, then we thoroughly analyse the possible relationship between miles per gallon values and transmission types using several regression models and a simple model selection algorithm, and conclude that in practice the transmission type has no statistically significant impact on MPG.

This analysis appears also on RPubs: http://rpubs.com/uxexax/452100

Input data

Overview

The analysis is based on the **Motor Trend Car Road Tests (mtcars)** dataset provided by the R package **datasets**. As its description says, *"the data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models)."*

The data is stored in a data frame, which has 32 observations on 11 numeric variables, which is summarized below:

Column #	Variable	Description
1	**mpg**	**Miles per US gallon**
2	*cyl*	*Number of cylinders*
3	*disp*	*Displacement (cu.in.)*
4	hp	Gross horsepower
5	*drat*	*Rear axle ratio*
6	wt	Weight (1000 lbs)
7	*qseq*	*Quarter mile time*
8	*vs*	*Engine (0 = V-shaped, 1 = straight*
9	**am**	**Transmission (0 = automatic, 1 = manual)**
10	gear	Number of forward gears
11	*carb*	*Number of carburetors*

The focus of the analyis is the relation between *miles per gallon (mpg)* and *transmission (am)*, marked with bold in the table. Other variables considered are in normal style, while variables not considered are in grey italic.

Transformations

Some of the variables in the dataset had been factorized:

```
mtcars$am <- factor(mtcars$am)
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

No other transformation had been done.

Exploratory analysis

The *mtcars* dataset has the following structure:

```
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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Non-factor variables considered in this analysis have the following summary:

```
summary(mtcars[c("mpg", "hp", "wt")])
```

```
##       mpg              hp              wt
##  Min.   :10.40   Min.   : 52.0   Min.   :1.513
##  1st Qu.:15.43   1st Qu.: 96.5   1st Qu.:2.581
##  Median :19.20   Median :123.0   Median :3.325
##  Mean   :20.09   Mean   :146.7   Mean   :3.217
##  3rd Qu.:22.80   3rd Qu.:180.0   3rd Qu.:3.610
##  Max.   :33.90   Max.   :335.0   Max.   :5.424
```

The only factor variable considered in this analysis has the following item counts:

```
table(mtcars$am)
```

```
##
##  0  1
## 19 13
```

Data is available in all observations for all considered variables (number of NAs is zero):

```
sapply(mtcars[c("mpg", "am", "hp", "wt")], function (X) sum(is.na(X)))
```

```
## mpg  am  hp  wt
##   0   0   0   0
```

Data analysis

We try three *linear* regression models to find out the how does *transmission (am)* impact *miles per gallon (mpg)*:

1. the *base model* takes only the *mpg* and *am* variables;
2. the *highest influence adjusted model* takes the base model and makes adjustments for *horse power* and *weight*, which turned out to be the largest infuencers of the base model;
3. the *transmission removed model* does not contain the transmission as the regressor, only the horse power and weight.

Note: Significance levels are pre-set at *0.05* for every test in this analysis.

The base model

First we analyse the relation between *miles per gallon* and *transmission* without any adjustment. Boxplot indicates that manual transmission is better for miles per gallon than automatic transmission:

```
explorafigs <- list()
explorafigs$base <-
  ggplot(mtcars, aes(x = am, y = mpg)) +
  theme_light() +
  labs(x = "Transmission (mpg)", y = "Miles per gallon (am)") +
  coord_cartesian(ylim = c(10, 35)) +
  geom_boxplot(color = I("steelblue")); explorafigs$base
```

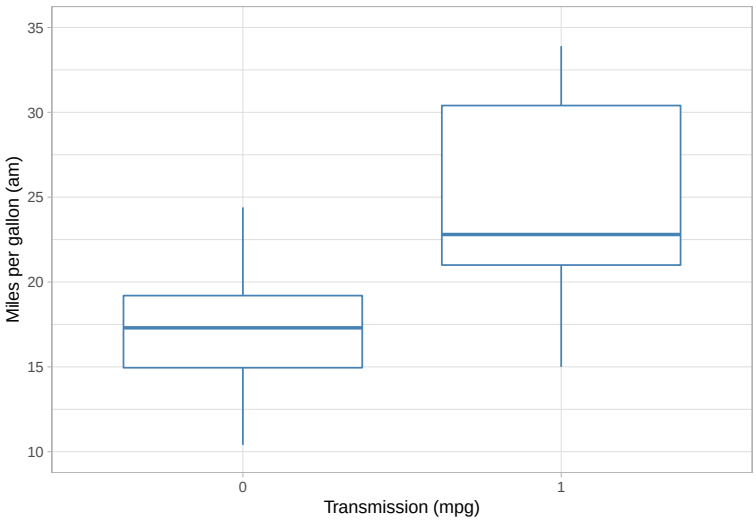

Figure 1: Fuel usage against transmission types in the mtcars dataset

This assumption is supported by the simple linear model *mpg* ~ *am*. We get the following estimations for the model coefficients:

```
M1 <- lm(mpg ~ am, data = mtcars)
coef(summary(M1)) %>% kable() %>% table.style
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
am1	7.244939	1.764422	4.106127	0.000285

And the following confidence intervals for the estimations above:

```
M1.civ <- confint(M1)
M1.civ %>% kable() %>% table.style
```

	2.5 %	97.5 %
(Intercept)	14.85062	19.44411
am1	3.64151	10.84837

The results tell us that the mean of MPG is 17.1473684 for automatic transmission (am = 0), but for manual transmission (am = 1) the mean MPG increases by 7.2449393 (± 3.6034297 using a 95% confidence interval) to 24.3923077 miles per gallon. Based on the P-value of the coefficient of *am1* the alternative hypothesis *transmission affects MPG* is accepted in favor of the null hypothesis *MPG is the same for automatic and manual transmissions* at significance level 0.05. **Based on this model, transmission has a statistically significant impact on MPG.**

Residuals seem properly scattered, although some grouping is visible.

```
residufigs <- list()
residufigs$base <-
  ggplot(mapping = aes(x = 1:nrow(mtcars), y = resid(M1))) +
  theme_light() +
  labs(x = "Model fit (mpg ~ am)", y = "MPG residual") +
  coord_cartesian(ylim = c(-10,10)) +
  geom_point(color = I("steelblue")) +
  geom_hline(yintercept = sum(resid(M1)), color = I("steelblue")); residufigs$base
```

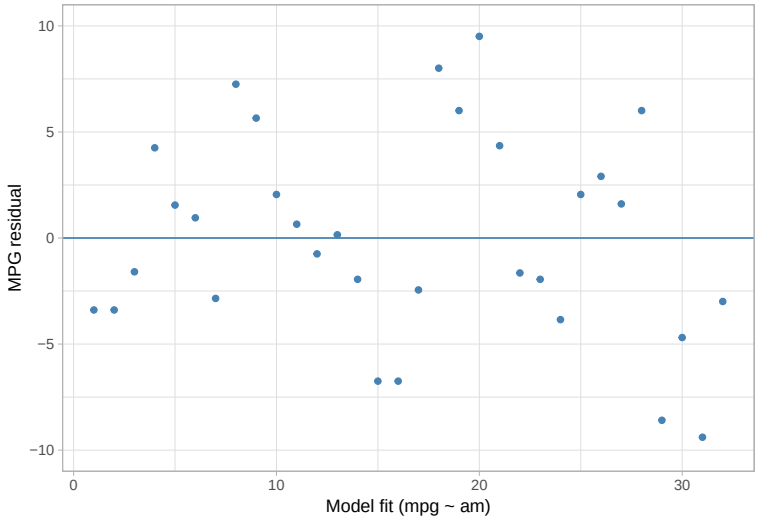

Figure 2: MPG residuals with their sum in the base model

Highest influence adjusted model

The base model indicates significant dependency of MPG on transmission type, in favor of manual transmission. However, is there any other variable which can have significant effect on the relationship between *mpg* and *am*? It appears to be a natural step to adjust the base model for the *number of forward gears*, as it sounds related to transmission and so should have an effect on *mpg* ~ *am*.

Instead of relying on gut feelings, we implemented a simple model selection method to get a model adjusted for variables with the highest influence on the MPG-transmission relationship. Starting with a base model, the algorithm iteratively extends it with new variables from the dataset, one at a time, whose addition has the lowest P-value. The algorithm does this until there is no new variable which has statistically significant impact on the previous model, that is its addition bears a P-value greater than the pre-set significance level

(or all the variables were used up). The algorithm uses ANOVA for testing. The algorithm is specified in the Annex section.

```
M2 <- model.selection(mtcars, "mpg ~ am")
```

```
## Selected model: mpg ~ am + hp + wt
```

In this particular case, we start with the base model ($mpg \sim am$), and get a model adjusted for horse power and weight in 1000 lbs: $mpg \sim am + hp + wt$. ANOVA shows strong evidences for adding these variables to the model one after the other:

```
anova(M2) %>% kable() %>% table.style
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
am	1	405.15059	405.150588	62.92168	0.000000
hp	1	475.45731	475.457311	73.84062	0.000000
wt	1	65.14822	65.148217	10.11781	0.003574
Residuals	28	180.29107	6.438967	NA	NA

However, taking a look at the coefficients of the model, it turns out that the expected effect of transmission type is smaller in this context than in the base model, and it is statistically less significant:

```
coef(summary(M2)) %>%
  kable() %>% table.style %>% row_spec(2, bold = TRUE)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34.0028751	2.6426593	12.866916	0.0000000
am1	**2.0837101**	**1.3764202**	**1.513862**	**0.1412682**
hp	-0.0374787	0.0096054	-3.901830	0.0005464
wt	-2.8785754	0.9049705	-3.180850	0.0035740

The confidence interval of the transmission coefficient estimation is narrower than in case of the base model, and its lower end is negative, which means chances are high that having manual transmission in a car instead of an automatic one has no positive effect on MPG at all:

```
confint(M2) %>%
  kable() %>% table.style %>% row_spec(2, bold = TRUE)
```

	2.5 %	97.5 %
(Intercept)	28.5896329	39.4161174
am1	**-0.7357587**	**4.9031790**
hp	-0.0571545	-0.0178029
wt	-4.7323235	-1.0248273

Residuals of this model are in general smaller than in the base model:

```
residufigs$highest <-
  ggplot(mapping = aes(x = 1:nrow(mtcars), y = resid(M2))) +
  theme_light() +
  labs(x = "Model fit (mpg ~ am + hp + wt)", y = "MPG residual") +
  coord_cartesian(ylim = c(-10,10)) +
  geom_point(color = I("steelblue")) +
  geom_hline(yintercept = sum(resid(M2)), color = "steelblue")

grid.arrange(grobs = residufigs, ncol = 2)
```

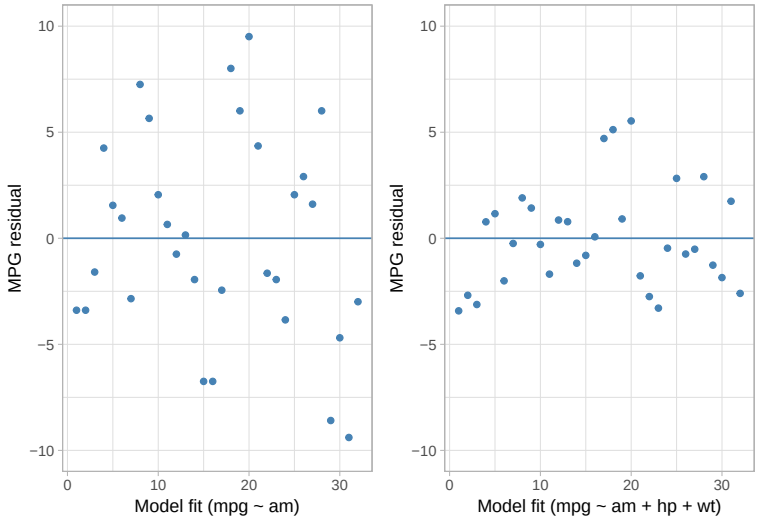

Figure 3: MPG residuals with their sum in the base and highest influence adjusted models

Transmission removed model

The *highest influence adjusted model* estimated the effect of transmission type on miles per gallon values to be non-significant in the presence of the two most powerful influencer of that relationship, horse power (*hp*) and weight in 1000 lbs (*wt*). In fact it seems that the only statistically significant regressors of MPG are horse power and weight. If we run the same `model.selection` algorithm used for the highest influence adjusted model, but now starting with an empty initial model for the MPG (i.e. no initial regressor specified), we get the model $mpg \sim hp + wt$ with the following ANOVA:

```
M3 <- model.selection(mtcars, "mpg ~ ")
```

```
## Selected model: mpg ~ wt + hp
```

```
anova(M3)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## wt         1 847.73  847.73 126.041 4.488e-12 ***
## hp         1  83.27   83.27  12.381  0.001451 **
## Residuals 29 195.05    6.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As a recap, `model.selection` iteratively extends a linear regression model with new variables one-by-one, until the change caused by the addition is statistically not significant anymore. Testing is done with ANOVA.

Comparing the residuals of the three models, they are smaller in this model than in the base, but looks quite similar like the residuals of the highest influence adjusted model, maybe a bit more tightly ordered around their expected value:

```
residufigs$notrans <-
  ggplot(mapping = aes(x = 1:nrow(mtcars), y = resid(M3))) +
  theme_light() +
  labs(x = "Model fit (mpg ~ hp + wt)", y = "MPG residual") +
  coord_cartesian(ylim = c(-10,10)) +
```

```
  geom_point(color = I("steelblue")) +
  geom_hline(yintercept = sum(resid(M3)), color = I("steelblue"))

grid.arrange(grobs = residufigs, ncol = 3)
```

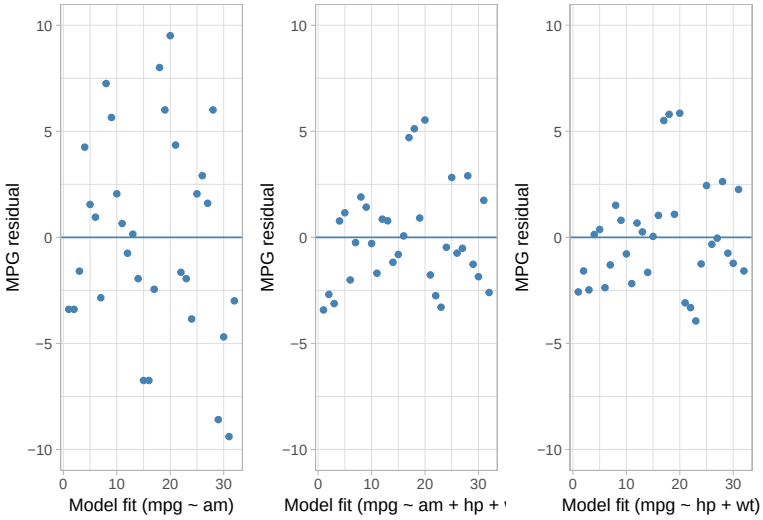

Figure 4: MPG residuals in the three models

Summary

We started with a simple linear regression model, which had only one regressor for MPG, the transmission type. This model indicates that there is a statistically significant difference between expected miles per gallon values for automatic and manual transmission, in favor of the latter; in other words, transmission type has a statistically significant impact on MPG, at least in this model.

Adjusting the base model for horse power and weight, however, made the expected effect non-significant. The used linear regression model $mpg \sim am + hp + wt$ was the outcome of a simple model selection algorithm specified in the Annex.

Finally, we have seen that transmission type seems to not have any significant effect on MPG; the only statistically significant regressors are horse power and weight.

Based on this the verdict of the analysis is that from MPG point of view it does not matter if a car has automatic or manual transmission system.

Annex

Model selection algorithm

```
model.selection <- function(data, initial.model, signif.threshold = 0.05)
{
  initial.variables <- strsplit(initial.model, " *[~+*] *")[[1]]
  potential.variables <- setdiff(colnames(data), initial.variables)

  if (length(initial.variables) > 1) # initial.model == "y ~ x ..."
  {
```

```
    model <- initial.model
    p.index <- 2
    operator <- "+"
  }
  else # initial.model == "y ~ "
  {
    model <- initial.variables[[1]]
    p.index <- 1
    operator <- "~"
  }

  while (length(potential.variables != 0))
  {
    p.values <-
      sapply(potential.variables,
           function (X)
             anova(lm(paste(model, operator, X), mtcars))$`Pr(>F)`[p.index])

    if (sum(p.values <= signif.threshold, na.rm = TRUE) == 0)
    {
      break
    }

    selected.variable <- potential.variables[which.min(p.values)]
    potential.variables <- potential.variables[-which.min(p.values)]
    model <- paste(model, operator, selected.variable)
    p.index <- p.index + 1
    operator <- "+"
  }
  message(paste("Selected model:", model))
  return (lm(model, data))
}
```