Modeling Car Fuel Economy

Marko Intihar

12/7/2020

Executive Summary

Analysis Outline

In our analysis we will be using "**Motor Trend Car Road Tests**" (**MTCARS**) data set. Our main goal is to answer given questions:

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

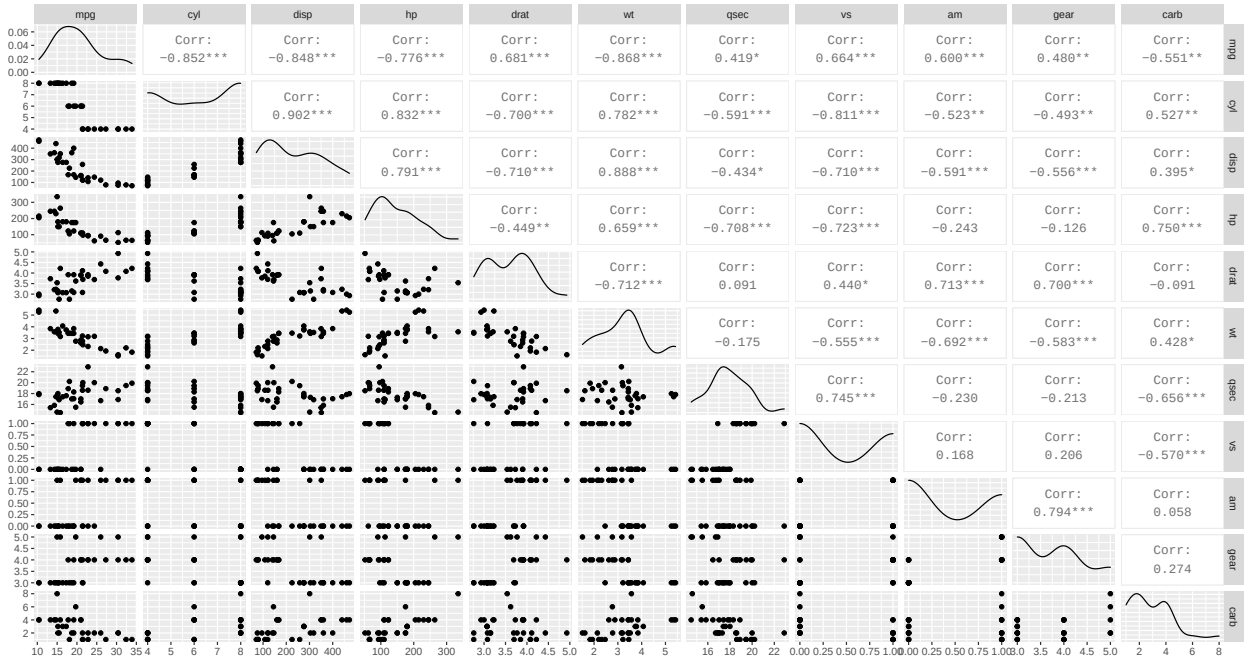
Main idea of the analysis is:

- to build a linear regression model
- outcome we try to predict with model is variable **mpg** (car's fuel economy - Miles/(US) gallon)
- proposed model will include regressor variable **am** (type of transmission)
- we will also include some other regressors variables, if this is a logical step proposed by model selection procedure
- first we start with data exploration

Exploratory Data Analysis

Let's first create scatter plot matrix, which shows us pairwise plots and correlation estimation for given variables:

```
ggpairs(data = mtcars)
```



From the figure above we can see how outcome variable **mpg** is related to other variables. If we compare **mpg** to type of transmission **am**, we can see there are some differences in fuel consumption regarding "manual" or "automatic" transmission. The linear correlation between two variables is 0.6 (estimated on sample) indicated a medium positive linear correlation between outcome (**mpg**) and variable of interest (**am**). The scatter plot matrix also show other linear correlation factors between outcome variable and other potential regressor variables. Based on the figure we will code the following variables as factors (based on their vlaues):

- **am** - type of transmission
- **cyl** - number of cylinders
- **vs** - type of engine
- **gear** - number of forward gears
- **carb** - number of carburetors

Remaining data set variables are coded as numerical variables.

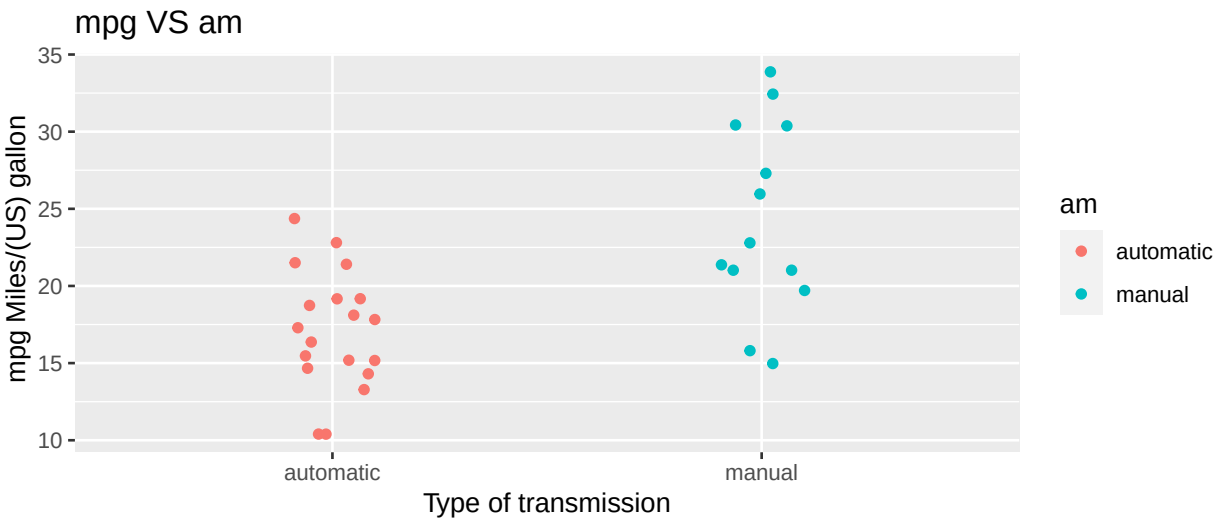
Now let's recode factor variables:

```
df <- mtcars %>%
  mutate(am = case_when(am == 0 ~ "automatic",
                        am == 1 ~ "manual"),
         am = as.factor(am),
         vs = case_when(vs == 0 ~ "V-shaped",
                        vs == 1 ~ "straight"),
         vs = as.factor(vs),
         cyl = as.factor(cyl),
         gear = as.factor(gear),
         carb = as.factor(carb))
```

To dig a little more deeper lets create a scatter plot drawing **mpg** VS *am*:

```
df %>%
  ggplot(aes(x = am, y = mpg, color = am)) +
```

```
  geom_jitter(width = 0.1) +
  xlab("Type of transmission") +
  ylab("mpg Miles/(US) gallon") +
  ggtitle("mpg VS am")
```



Modeling

We will follow given modeling strategy:

- build linear model where outcome is **mpg** and all other MTCARS variables are regressor variables - **benchmark model**
- using **bencmark model** calculate variance inflation factor (VIF) for each regressor
- do a nested model search using VIF as indicator what is included first and ANOVA to tell us which model is the most significant

Let's build a benchmark model and calculate VIF (and also visualize regressors based on VIF):

```
modb <- lm(mpg ~ ., mtcars) # fit benchmark model
VIF <- car::vif(modb) # variance inflation factors
VIF <- data.frame(var = names(VIF),
                  VIF = VIF) %>%
  mutate(priotiry = case_when(var == "am" ~ 1,
                              T ~ 0)) %>%
  arrange(desc(priotiry), VIF) # sort variables
```

Now lets fir models for a model nested search, we strat with one regressor **am**, then for each next model we add one additional regressor:

```
for(m in 1:nrow(VIF)){
  mod <- lm(mpg ~ .,
            data = df[, c("mpg", VIF %>% head(m) %>% pull(var))])
  assign(paste0("mod",m), mod)
}
```

Now apply ANOVA for selecting final model:

```
anova(mod1, mod2, mod3, mod4, mod5, mod6, mod7, mod8, mod9, mod10)
```

```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ am + drat
## Model  3: mpg ~ am + drat + vs
## Model  4: mpg ~ am + drat + vs + gear
## Model  5: mpg ~ am + drat + vs + gear + qsec
## Model  6: mpg ~ am + drat + vs + gear + qsec + carb
## Model  7: mpg ~ am + drat + vs + gear + qsec + carb + hp
## Model  8: mpg ~ am + drat + vs + gear + qsec + carb + hp + wt
## Model  9: mpg ~ am + drat + vs + gear + qsec + carb + hp + wt + cyl
## Model 10: mpg ~ am + drat + vs + gear + qsec + carb + hp + wt + cyl + disp
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1      30 720.90
## 2      29 573.64  1   147.256 18.3455 0.0006536 ***
## 3      28 339.99  1   233.651 29.1087 7.435e-05 ***
## 4      26 326.17  2    13.818  0.8607 0.4427160
## 5      25 284.36  1    41.808  5.2085 0.0374896 *
## 6      20 176.84  5   107.525  2.6791 0.0635498 .
## 7      19 154.43  1    22.412  2.7922 0.1154530
## 8      18 144.18  1    10.250  1.2769 0.2762177
## 9      16 130.37  2    13.807  0.8600 0.4429914
## 10     15 120.40  1     9.967  1.2417 0.2826734
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```