Course_Project1

doyougnu

September 19, 2015

Executive Summary

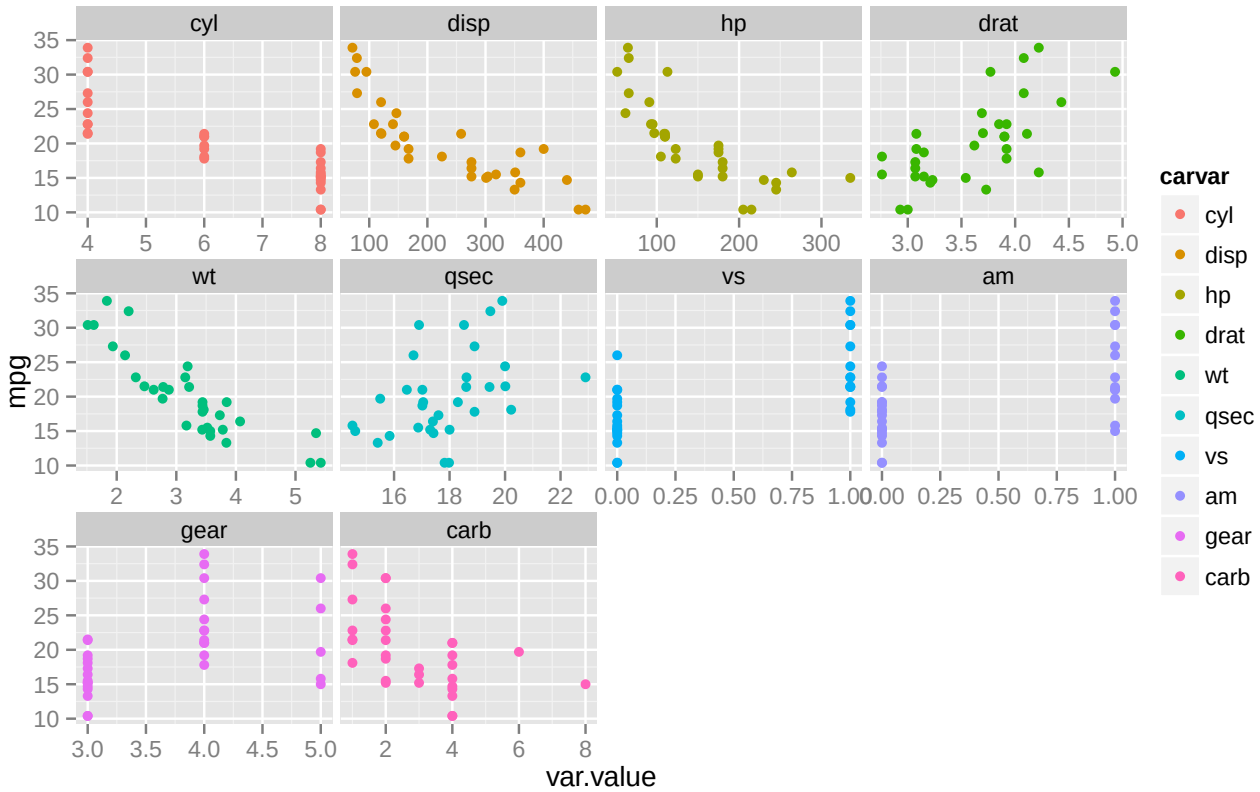
Data loading and first look

First lets load and look at the data to assess variation by variable, I will also do a quick plot to see a linear regression of each variable with mpg

```
data(mtcars)
mtcars <- mtcars #call mtcars because data load is lazy

#load libs
library(ggplot2); library(dplyr); library(tidyr)

#some quick munging
df <- mtcars %>% gather(carvar, var.value, cyl:carb)
ggplot(df)  +
    geom_point(aes(x = var.value, y = mpg, color = carvar)) +
    facet_wrap( ~ carvar, scales = "free_x")
```



Looks like most variables are well correlated to mpg. For a good model we need factors that explain the variation in MPG while being orthogonal to other explanatory factors. I will refer to domain specific

knowledge provided in the Henderson and Vellermans 1981 paper referenced in the mtcars dataset[1]. This is an important tactic because in that paper the authors provide a physical explanation of work of a car (work in the physics sense of force applied over a distance). I will be relying on their results heavily for this model.

Model Adjustments

In that paper the authors state: 1. GPM (Gallon per Mile) is a more useful response variable due to the non-linearity of several variables, Weight, Displacement etc.. and domain specific knowledge of the sources of work in a car 2. Given GPM, Weight is the single greatest predictor in a linear regression model(Henderson et al. 397) 3. Calculating the ratio of Horse Power to Weight provides a useful factor to add to the meodel because it provides a metric to asses how overpowered a car is in this dataset (defined by large HP/WT). Furthermore, it is uncorrelated to WEIGHT (r = 0.03) (Henderson et al. 398)

Model Construction

In this analysis we are interested in the difference between standard and automatic transmissions, given the considerations above we will want to look at the difference between automatic and manual transmissions **holding all other significant variables constant**. Thus we need to add the trans variable to the model described by Henderson et al

```
df <- mtcars; df$GPM <- mtcars$mpg ^ -1; df$HPWT <- mtcars$hp / mtcars$wt
fit <- lm(GPM ~ wt + HPWT + factor(am), data = df)
summary(fit)$coeff
```

```
##                   Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) -0.0050079683 6.752230e-03 -0.7416762 4.644595e-01
## wt           0.0150236408 1.799288e-03  8.3497717 4.392184e-09
## HPWT         0.0002329434 8.024195e-05  2.9030121 7.130316e-03
## factor(am)1  0.0008369016 3.625290e-03  0.2308509 8.191091e-01
```

As one can see from above the addition of a transmission variable is insignificant - which is unsurprising because most of the variation in the data is already explained by WEIGHT and HP/WT. However this is only true **for this dataset* looking at the cars listed in the dataset we are unlikely to find cars whose mpg would be significantly altered by manual and automatic transmitions. For a more detailed analysis on why see Henderson et al, page 396-397. Thus, for this dataset, neither automatic nor manual transmission is better for MPG and furthermore the difference between the two is negligible.

Conclusion

In this short analysis I set out to determine two things 1. Does transmission type matter in reference to MPG and 2. Is so, what is the quantitative difference between the two in terms of MPG. I found that 1. Given prior work on the subject, the variation in this dataset is explained by Weight of the vehicle and the Horse Power to Weight ratio. 2. The addition of transmission to the model was statistically insignificant which as noted in prior research on this dataset is most likely due to the absence of mid sized sedans in the dataset.

Links

1: http://www.mortality.org/INdb/2008/02/12/8/document.pdf