Gas Mileage as a function of Transmission type

Chris Emerson

Wednesday, September 16, 2015

Executive Summary

Looking at a data set of a collection of cars, let us explore the relationship between a set of variables and miles per gallon (MPG) (outcome). Particularly of interest is the following two questions:

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

Please see my conclusion for full details but in brief, manual trasmissions are have significantly better mpg. See Appendix A for details regarding the dataset.

Exploratory Data Analyses

I normalized the data set on columns that I defined as continuous; I did so for the data as a whole as well as subset of automatic transmissions and subset of manual transmissions then combined the 3 results. See Appendix A and B for my definition of continuous.

Correlation

Normalizing all the data has no affect on correlation calculation of total or by subset. Normalzing each subset, Manual and Automatic, does have an affect when corelating all the data but not the subset.

```
##    mpg disp  hp    drat wt    qsec description
## c  1   -0.85 -0.78 0.68 -0.87 0.42 "All data normalized together"
## cc 1   -0.61 -0.65 0.36 -0.63 0.57 "All data normalized by subset"
## nm 1   -0.83 -0.8  0.47 -0.91 0.8  "Manual normalized"
## na 1   -0.79 -0.83 0.47 -0.77 0.66 "Automatic normalized"
```

Variance Inflation factor

```
##     disp    hp  drat    wt  qsec
## c  3.018 2.281 1.524 2.648 1.787
## cc 2.495 2.214 1.165 1.934 1.897
```

drat and qsec vary drastically between the two normalization strategies.

Model Selection

fit multiple models to dtermine optimal model selection

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + disp + wt + hp
## Model 3: mpg ~ factor(am) + disp + wt + hp + qsec
## Model 4: mpg ~ factor(am) + disp + wt + hp + qsec + drat
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     27 179.91  3    540.99 30.0363 1.885e-08 ***
## 3     26 153.44  1     26.47  4.4089   0.04601 *
## 4     25 150.09  1      3.34  0.5571   0.46240
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With near zero p-value, model 2 seems to be the best choice of regressors, but the addition of qsec gives us two strong multivariable infleunces. So model 3.

Interpret The Coefficients

```
##             Estimate Std. Error t value Pr(>|t|)
## disp           0.011      0.011   1.060    0.299
## wt            -4.084      1.194  -3.420    0.002
## hp            -0.021      0.015  -1.460    0.156
## qsec           1.007      0.475   2.118    0.044
## factor(am)0   17.147      0.557  30.768    0.000
## factor(am)1   24.392      0.674  36.203    0.000
```

```
##                    2.5 %       97.5 %
## disp        -0.01055781  0.033033109
## wt          -6.53883919 -1.629824922
## hp          -0.05098537  0.008644273
## qsec         0.02963058  1.984163085
## factor(am)0 16.00178580 18.292951044
## factor(am)1 23.00736583 25.777249551
```

Wt seems to be the only factor of weight, if you'll pardon the pun. But qsec is at least significant compared to the other factors. Transmission type looks even more significant but I think that could be explained using appendix C, graphing.

Residuals

```
## [1] 5.901454
```

Residiual variance is roughly the span on the confidence interval for weight.

Conclusions

As shown by the graphs (see Appendix C), automatics run heavier on average than manual transmission vehicles though they can be significantly faster at the quarter mile those few are outliers. The result is distinct separation in the confidence intervals for automatics and manual transmission with manuals the clear winner for mpg. It is possible my decisions for normalizing and which variables should be ignored skewed the results. Based on the results I produced, great gas mileage is mostly a factor of weight with accleration a distant second although one could easiy explain how qsec could be inversely related to weight.

Appendix A Data Description

> 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).

- [, 1] mpg Miles/(US) gallon
- [, 2] cyl Number of cylinders
- [, 3] disp Displacement (cu.in.)
- [, 4] hp Gross horsepower
- [, 5] drat Rear axle ratio
- [, 6] wt Weight (lb/1000)
- [, 7] qsec 1/4 mile time
- [, 8] vs V/S
- [, 9] am Transmission (0 = automatic, 1 = manual)
- [,10] gear Number of forward gears
- [,11] carb Number of carburetors

Appendix B, Rational for dropping columns.

Good regressors have three main qualities.

1. They vary enough along the x axis (distribution).
2. When they vary along the x axis, the outcome variable varies along the y axis in linear or curvilinear or some identifiable pattern.
3. Using the linear pattern as an example, the points stay fairly close to that line.

Cylinder count, gear count, carboretor count, and engine configuration are removed becuase they are poor regressors.

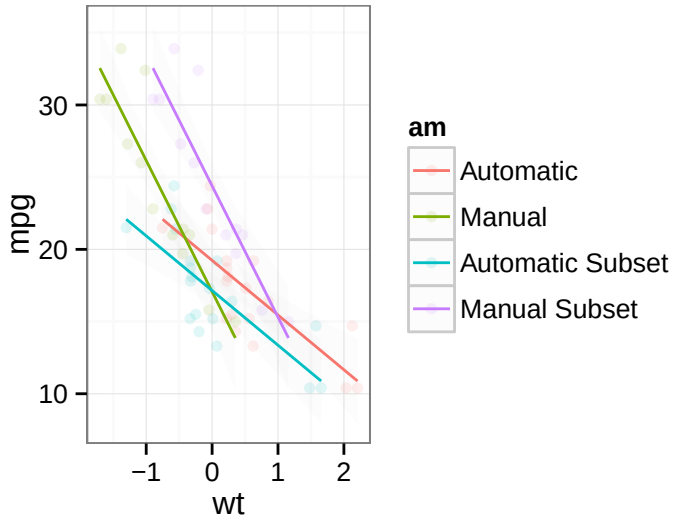
```
cols <- colnames(mtcars)
enumerated <- c('cyl','gear','carb','vs','am')
continuous <- cols[ !cols %in%  enumerated]
```

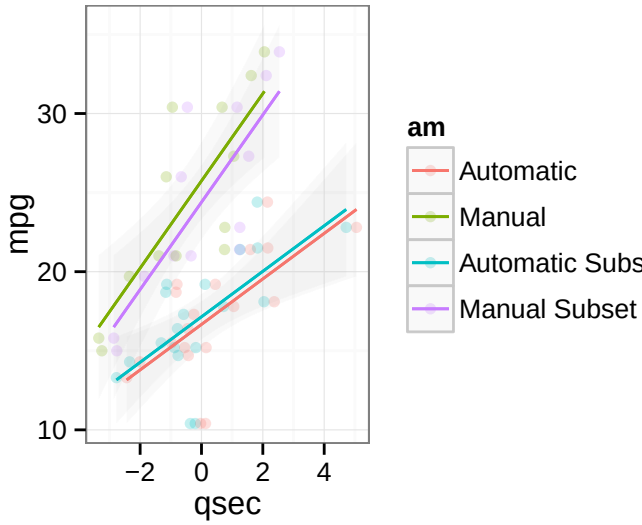
Appendix C Exploring the data

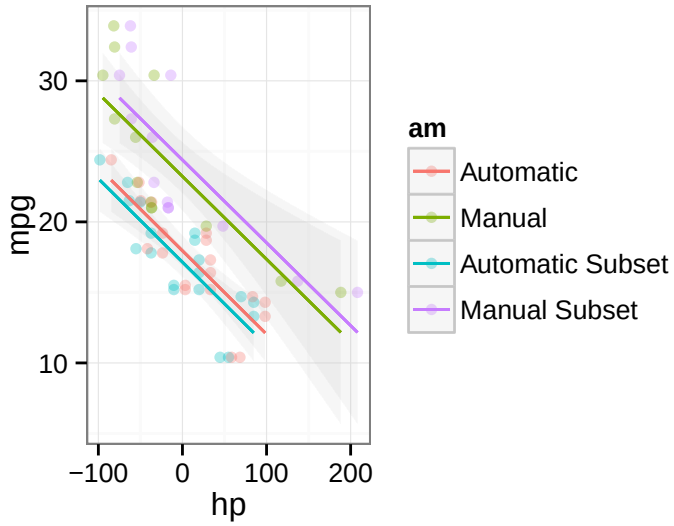
Here is are scatter plots with linear regression line for normalized data sets. Automatic and Manual were normalized together where Automamitic subset and Manual subset were normalized separately and added into the first data set. The difference in Manual and Automatics has little do with transmission and is more of a function of design as evidenced by the subset lines overlapping with little variation in slope of the line.

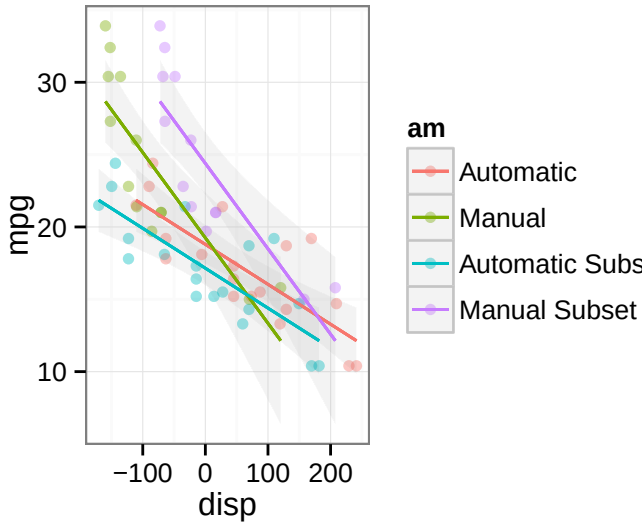
```
normalize_mtcars <- function(ret,cols) {
  for (val in cols) {
    if (val == "mpg"){
      next
    }
    t <- ret[[val]]
#    t <- (t - mean(t))/sd(t)
    t <- (t - mean(t))
    ret[[val]] <- t
  }
  ret
}
```

```
n <- normalize_mtcars (mtcars,continuous)
n0 <- normalize_mtcars (mtcars[mtcars$am==0,],continuous)
n1 <- normalize_mtcars (mtcars[mtcars$am==1,],continuous)
```



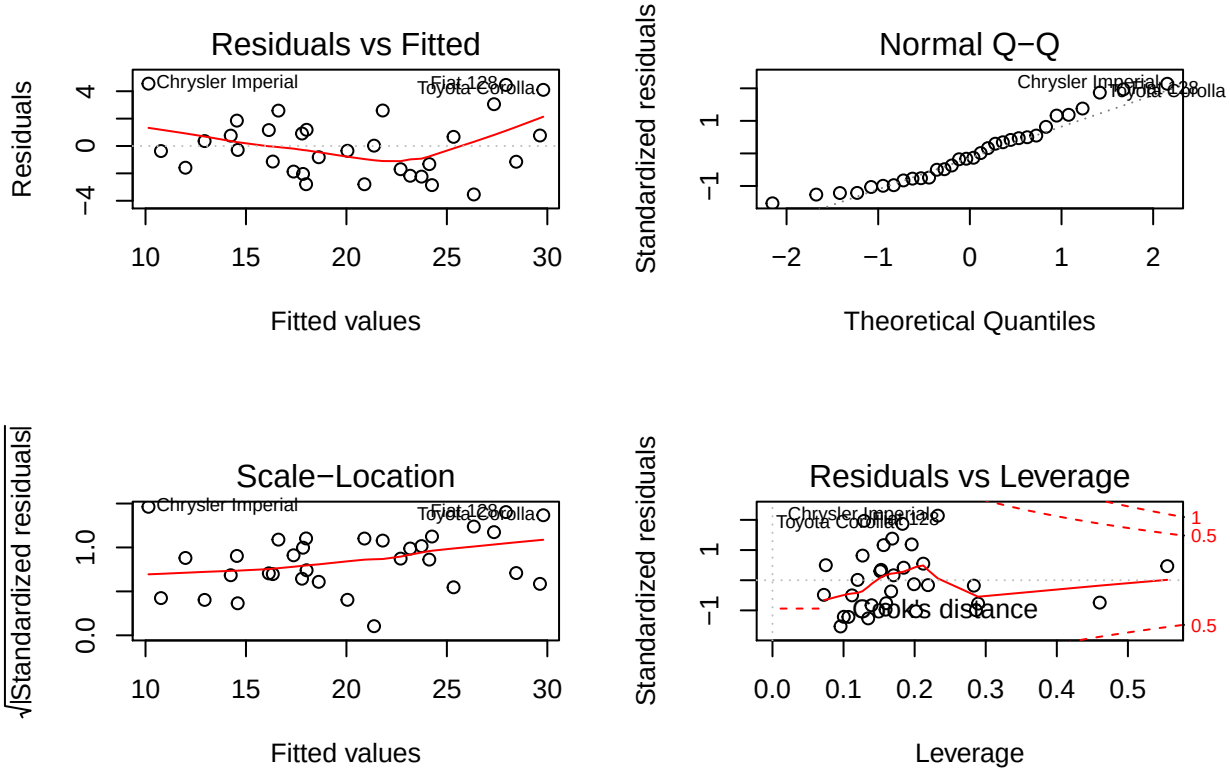






Appendix D Residual Analysis

No abnormalities are observed in the residual plot.


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage