Impact of different variables on miles per gallon (MPG) in a data set of a collection of cars

Antonio Avella

September 2, 2018

Executive Summary

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

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

The dataset

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 format is a data frame with 32 observations on 11 (numeric) variables. A codebook for the dataset is given below:

- mpg Miles/(US) gallon
- cyl Number of cylinders
- disp Displacement (cu.in.)
- hp Gross horsepower
- drat Rear axle ratio
- wt Weight (1000 lbs)
- qsec 1/4 mile time
- vs Engine (0 = V-shaped, 1 = straight)
- am Transmission (0 = automatic, 1 = manual)
- gear Number of forward gears
- carb Number of carburetors

Exploratory analysis and data trasformations

We load the data set, perform data transformations by factoring the some variables and look the data:

```
data(mtcars)
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
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 "Automatic","Manual": 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 ...
```

We explored various relationships between variables of interest and the outcome. Initially, we plot the relationships between all the variables of the dataset (see plot 1 in the Appendix). From the plot 1 we notice that variables like cyl, disp, hp, drat, wt, vs and am seem to have some strong correlation with mpg. We will use linear models to quantify that in the next section. Additionally we plot a boxplot of the variable mpg when am is 'Automatic' or 'Manual' (see plot 2 in the Appendix). This plot shows that the mpg incresases when the transmission is 'Manual'.

Regression Analysis

In this section we will build linear regression models based on the different variables of interest and try to find out the best model fit. We will compare it with the base model which we have using ANOVA. After model selection, we will perform an analysis of residuals.

Based on plot 1 there are several variables seem to have high correlation with mpg. We will build an initial model with all the variables as predictors and perfom stepwise model selection to select significant predictors for the final model. This is taken by the step method, which runs lm multiple times to build multiple regression models and select the best variables from them, using both forward selection and backward elimination methods by the AIC algorithm:

```
mod_init <- lm(mpg ~ ., data = mtcars)
mod_best <- step(mod_init, direction = "both")
```

As we can see, the best model obtained from the above computations have cyl, wt, hp and am as relevant variables:

```
summary(mod_best)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

We can see that the adjusted R2 value is equal to 0.84 which is the maximum obtained considering all combinations of variables. Therefore we can conclude that more than 84% of the variability is explained by this model. Now, using ANOVA, we will compare the base model with only am as the predictor variable and the best model obtained above:

```
mod_base <- lm(mpg ~ am, data = mtcars)
anova(mod_best, mod_base)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     26 151.03
## 2     30 720.90 -4   -569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at this result, the p-value obtained is highly significant, and we reject the null hypothesis that the confounder variables cyl, hp and wt do not contribute to the accuracy of the model.

Analysis of the Residuals and Diagnostics

Now we explore the residual plots of our regression model and also compute some of the regression diagnostics for our model to find out some interesting leverage points (often called as outliers) in the data set:

From the final plots in the Appendix we can conclude the following:

- The Residuals vs Fitted plot shows random points on the plot that verifies the independence condition.
- In the Normal Q-Q plot the points mostly fall on the line indicating that the residuals are normally distributed.
- In the Scale-Location plot the points are in a constant band pattern, indicating constant variance.
- Finally, the Residuals vs Leverage plot shows some points of interest (outliers or leverage points) are in the top right corner.

Now we will compute some regression diagnostics of our model to find out these interesting leverage points. We compute top three points in each case of influence measures.

```
lev <- hatvalues(mod_best)
tail(sort(lev),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

```
inf <- dfbetas(mod_best)
tail(sort(inf[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

Looking at this result we see that they the same cars shown in the residual plots.

Statistical Inference

Finally, we will perform a t-test assuming that the transmission data has a normal distribution and we will see that the manual and automatic transmissions are significatively different:

```r
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Conclusions

From the summary(mod_best) we can conclude the following:

*Miles per gallon mpg will increase by 1.81 in cars with 'Manual' transmission compared to cars with 'Automatic' transmission (adjusted by hp, cyl, and wt). So, the conclusion for Motor Trend Magazine is: 'Manual' transmission is better for mpg.

The answer to the first question is: 'Manual' transmission is better than 'Automatic' from the point of view of miles per gallon

The answer to the second question is: Miles per gallon mpg will increase by 1.81 in cars with 'Manual' transmission compared to cars with 'Automatic' transmission (adjusted by hp, cyl, and wt).

Furthermore:

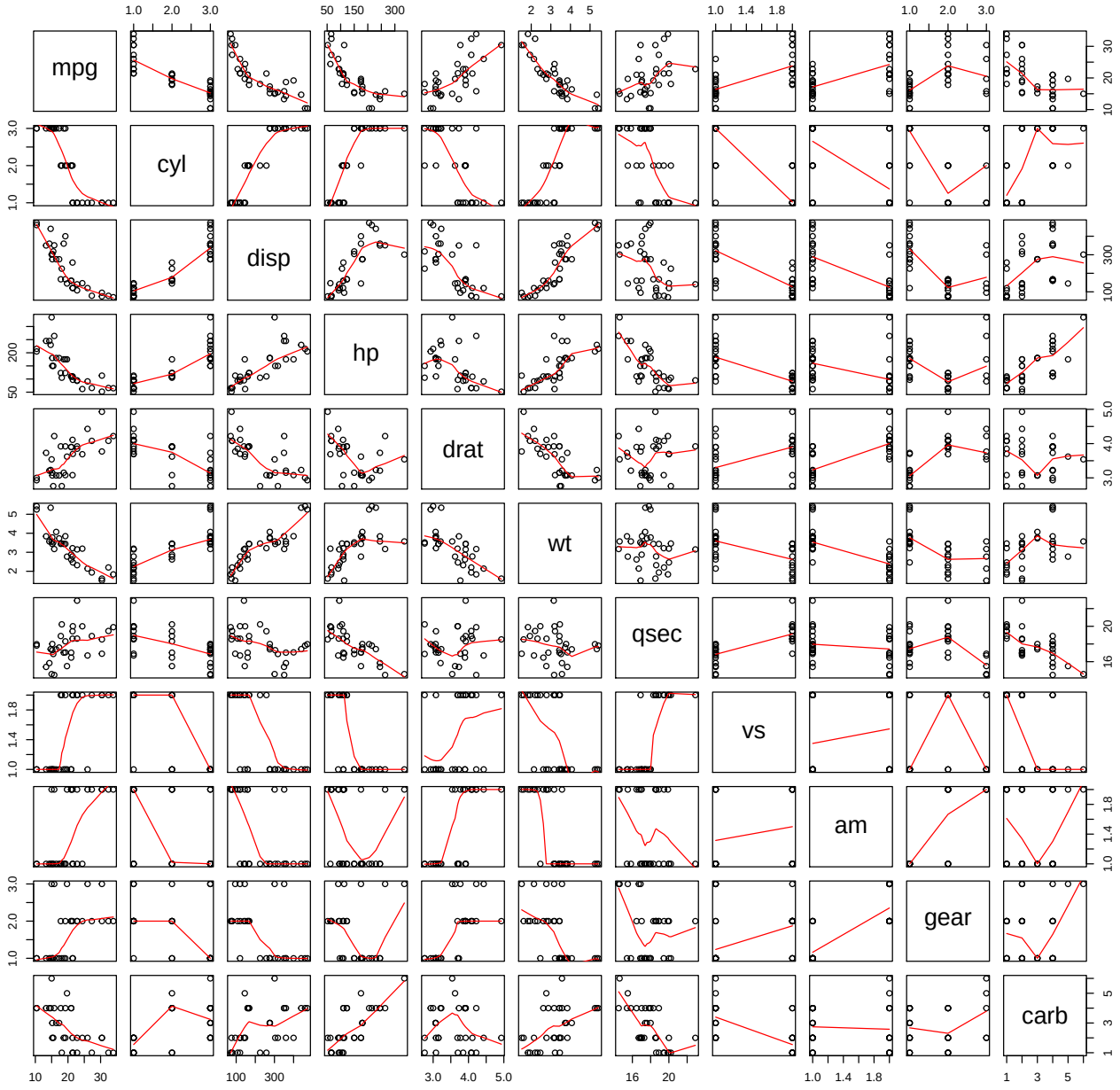
- Miles per gallon mpg will decrease by 2.5 for every 1000 lb of increase in wt (adjusted by hp, cyl, and am).
- Miles per gallon mpg decreases with increase of hp.
- Miles per gallon mpg will decrease by a factor of 3 and 2.2 if number of cylinders cyl increases from 4 to 6 and 8, respectively (adjusted by hp, wt, and am).

Appendix

As we shown in Exploratoy Analysis section we explored various relationships between variables of interest and the outcome. We plot the relationships between all the variables of the dataset:

```r
library(car)
pairs(~mpg+cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb,
      panel=panel.smooth,
      data=mtcars,
      main="Plot 1: Scatterplot Matrix")
```

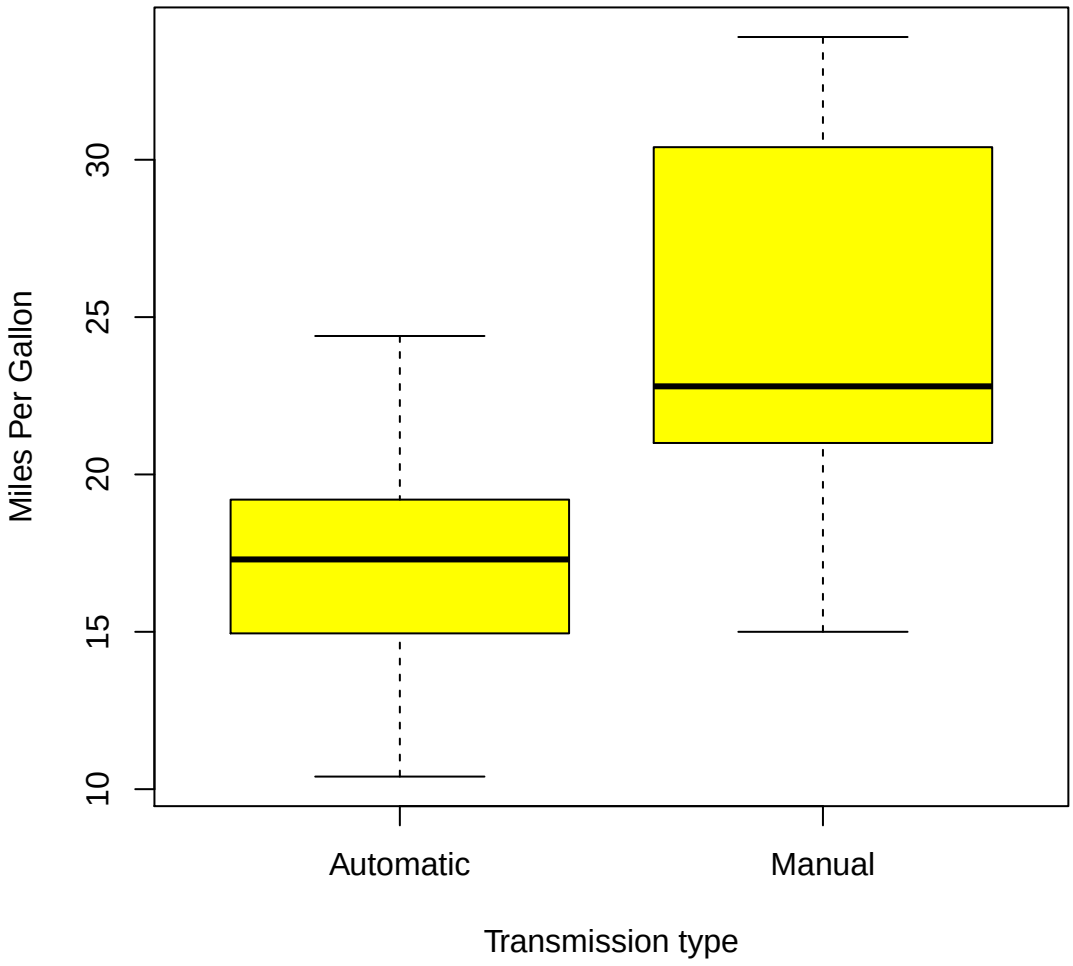
Plot 1: Scatterplot Matrix



Additionally we plot a boxplot of the variable mpg when am is 'Automatic' or 'Manual':

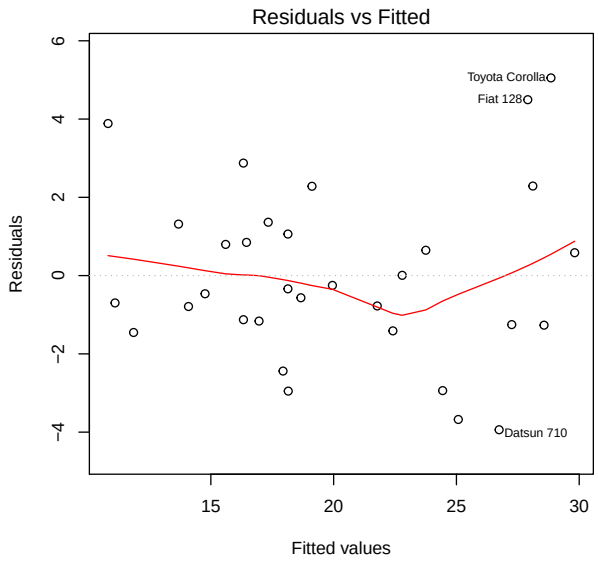
```
boxplot(mpg ~ am, data=mtcars, col="yellow", main="Plot 2: Miles per gallon by Transmission type",
   xlab="Transmission type", ylab="Miles Per Gallon")
```

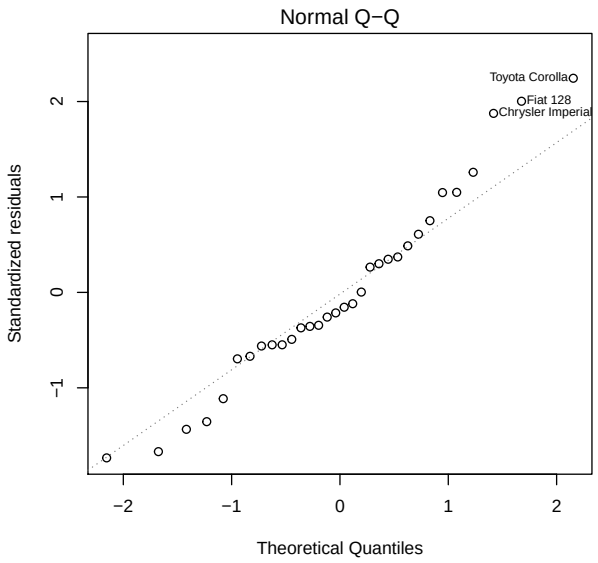
Plot 2: Miles per gallon by Transmission type

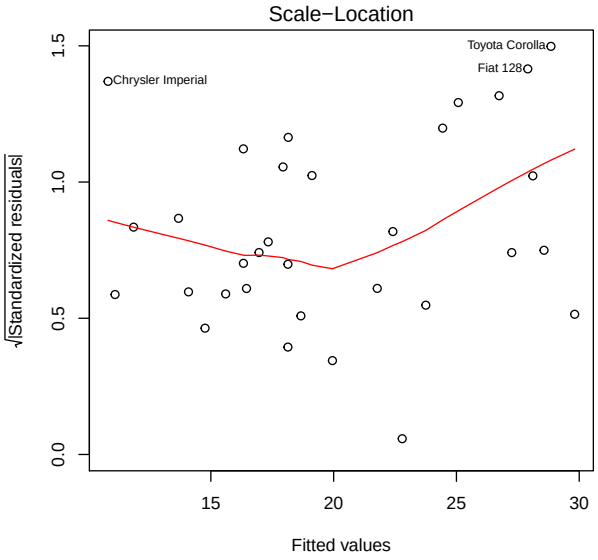


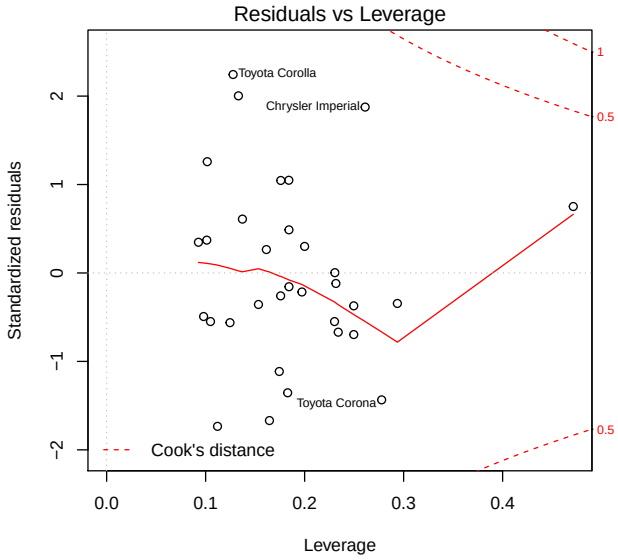
Now we plot the residual plots of our regression model necessary for Analysis of the Residuals and Daisgnostic section

```
par(mfrow=c(2,2))
plot(mod_best, which=1)
plot(mod_best, which=2)
plot(mod_best, which=3)
plot(mod_best, which=5)
```


Residuals vs Fitted
Toyota Corolla
Fiat 128
Datsun 710
Residuals
Fitted values


Normal Q−Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Standardized residuals
Theoretical Quantiles


Scale−Location
Toyota Corolla
Fiat 128
Chrysler Imperial
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
Standardized residuals
Leverage