Discriminating Fuel Efficiency Between Automatic and Manual Cars

T. Ruzmetov

November 5, 2016

Introduction

In this project, we are interested in exploring how set of given variables affect MPG (outcome) of Motor Vehicles. Particularly, central questions to be investigated are:

- Which transmission type is better for MPG, automatic or manual?
- How to quantify the MPG difference between automatic and manual transmissions?

Data

The data is 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).

Variables

- mpg –> miles / gallon
- cyl –> number of cylinders
- disp –> Displacement (in cubic inches) - swept volume of all the pistons inside the cylinders
- hp –> gross horsepower
- drat –> Rear axle ratio - comparison of the number of gear teeth on the ring gear of the rear axle and the pinion gear on the driveshaft. Higher axle ratio offers more towing power and quicker acceleration.
- wt –> weight (1000 lbs)
- qsec –> 1/4 mile time - time it takes for vehicles to accelerate 1/4 mile
- vs –> V-engine or a straight engine
- am –> Transmission (0 = automatic, 1 = manual)
- gear –> Number of forward gears
- carb –> Number of carburetors

```
library(knitr); library(ggplot2); library(lattice)
library(plyr); library(Rmisc); library(reshape2)
data(mtcars); head(mtcars,3)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

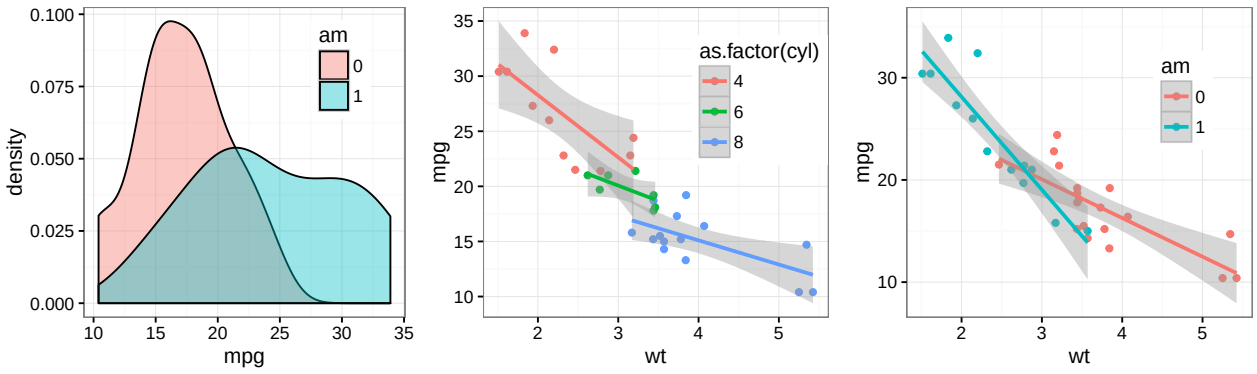
```
mtcars$am <- as.factor(mtcars$am)
mtcars$vs <- as.factor(mtcars$vs)
```

Exploratory Analysis

```
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 : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ 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: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
p1 <- ggplot(mtcars, aes(x=mpg, fill=am)) +
      geom_density(alpha = 0.4)+
      theme_bw()+
      theme(legend.position=c(0.8,0.8))

p2 <- ggplot(mtcars, aes(x=wt,y=mpg, color=as.factor(cyl))) +
      geom_point()+geom_smooth(method = "lm",formula = y~x)+
      theme_bw()+
      theme(legend.position=c(0.8,0.7))

p3 <- ggplot(mtcars, aes(x=wt,y=mpg, color=am)) +
      geom_point()+geom_smooth(method = "lm",formula = y~x)+
      theme_bw()+
      theme(legend.position=c(0.8,0.7))

suppressWarnings(multiplot(p1,p2,p3,cols=3))
```



Left plot shows that, on average, cars with automatic transmission consume more gasoline than cars with manual transmission. Middle figure demonstrates "mpg" dependence on "wt" for vehicles with different

number of cylinders. All 3 types show decrease in "mpg" as "wt" increases. Third plot represents mpg comparison between automatic and manual transmission at different weights. As shown, increase in "wt" makes both "am" type cars less fuel efficient, which makes sense. In addition, effect of change in weight is more pronounced for manual cars. We have to notice that this conclusions are very weak since they don't take into account other influential factors and solely based on a few variables. Thus, we need to use more advanced regression analysis and diagnostics to draw meaningful conclusion.

Coefficient Interpretation & Model Selection

Let's start by performing multivariate linear regression on a given data using mpg as a predictor and the rest of the variables as regressors.

```
net_fit <- lm(mpg ~ . ,data=mtcars)
kable(round(summary(net_fit)$coefficients,2), align = "c")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.30	18.72	0.66	0.52
cyl	-0.11	1.05	-0.11	0.92
disp	0.01	0.02	0.75	0.46
hp	-0.02	0.02	-0.99	0.33
drat	0.79	1.64	0.48	0.64
wt	-3.72	1.89	-1.96	0.06
qsec	0.82	0.73	1.12	0.27
vs1	0.32	2.10	0.15	0.88
am1	2.52	2.06	1.23	0.23
gear	0.66	1.49	0.44	0.67
carb	-0.20	0.83	-0.24	0.81

Coefficients can be interpreted in the following way: When using weight as a regressor, slope would represent how 1000lb increase in cars weight affects mpg keeping all other variables constant. Same rule applies to all other slope coefficients. For this set of data, slope for weight is $dmpg/dwt$=-3.71, which tells that mpg decreases as we increase wt. Intercept coefficients are not interesting.

P-values shown are based on hypothesis testing where

- Null hypothesis — slope = 0
- Alternative hypothesis — slope = Estimate

As we can see, all p-values are greater that statistical significance level(5%). At this point we could accept null hypothesis and give up, but that would be very bad decision to make. Let's see what we can do to improve our model. One of the major problems in multivariate regression is collinearity, where some variables are highly correlated with each other, which results in increased standard error and inaccuracy in the estimate of the slope. We can monitor collinearity by calculating "Variance Inflation Factor". Let' see what it gives us by including all variables into model.

```
library(car); round(vif(net_fit),2)
```

```
##   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## 15.37 21.62  9.83  3.37 15.16  7.53  4.97  4.65  5.36  7.91
```

We have high Variance inflation factors, which have detected collinearity problem in our model. Now, in order to improve the model, we will use "Stepwise Selection Method" to terminate unnecessary variables.

```
library(MASS)
stepwise <- stepAIC(net_fit, direction="both", trace=FALSE)
summary(stepwise)$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am1          2.935837  1.4109045  2.080819 4.671551e-02
```

```
summary(stepwise)$r.squared
```

```
## [1] 0.8496636
```

Now, let's carefully look at the table. We know that "am" is a categorical variable (0-automatic, and 1-manual) and when we use factor variable in our regression model, say for binary case, estimate for "am1" is just a change in the average "mpg" when we go from "am=0" to "am=1". **This means that manual cars are Δmpg=2.93 more fuel efficient than automatic cars.** Difference was little less than what we have now for unadjusted case($\Delta mpg = 2.52$), and the difference was much bigger when we calculated just an average for "am" as a factor. On the other hand R^2 =0.85 indicates that the model explains 85 the variability of the response data around its mean.

```
mean(mtcars[mtcars$am==1,]$mpg) -  mean(mtcars[mtcars$am==0,]$mpg)
```

```
## [1] 7.244939
```

So, "stepAIC" gave us best combination of variables which will give optimum performance, where for chosen variable combination all p-values are less than 0.05 in favor of alternative hypothesis. Now we can, once again do "VIF" calculation to check if collinearity issue persists.

```
Final_fit <- lm(mpg ~ wt+qsec+am, data = mtcars)
round(vif(Final_fit),2)
```

```
##   wt qsec   am
## 2.48 1.36 2.54
```

So, all "VIF" values are greater than 1 and less than 5, which makes them moderately correlated. Now, we can proceed to answer actual set of questions.

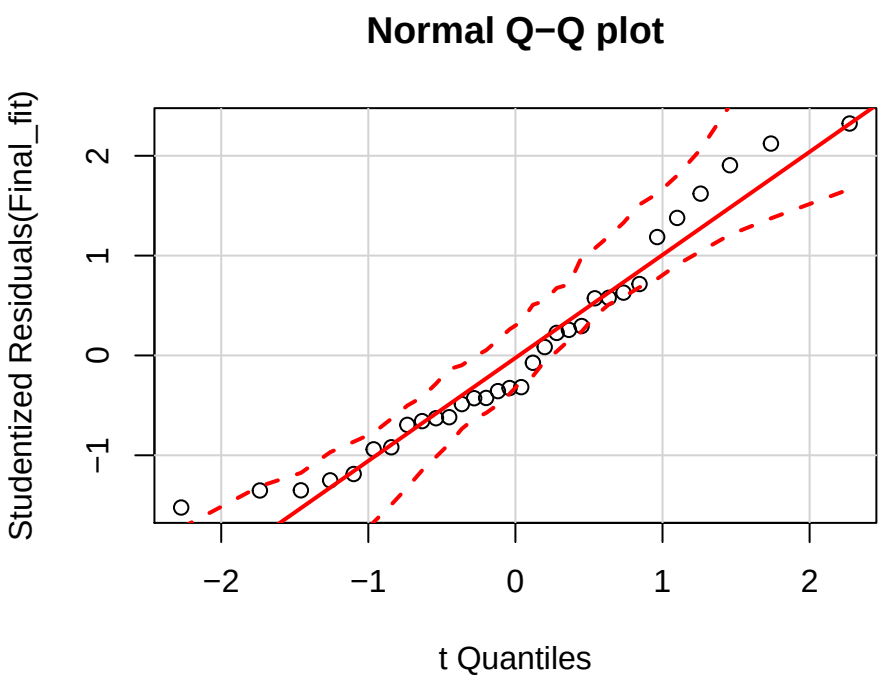
Residual Diagnostics

1. Residual help diagnose normality of the errors, which is usually done by plotting the residual quantiles versus normal quantiles.
2. Residual diagnostics helps us to detect outliers:
    - High leverage points (leverage is measured by hat diagonals (in R "hatvalues"). The hat values must be between 0 and 1 with larger values indicating greater (potential for) leverage.)

- High influence points (measured using “dffits” and “dfbetas” in R)

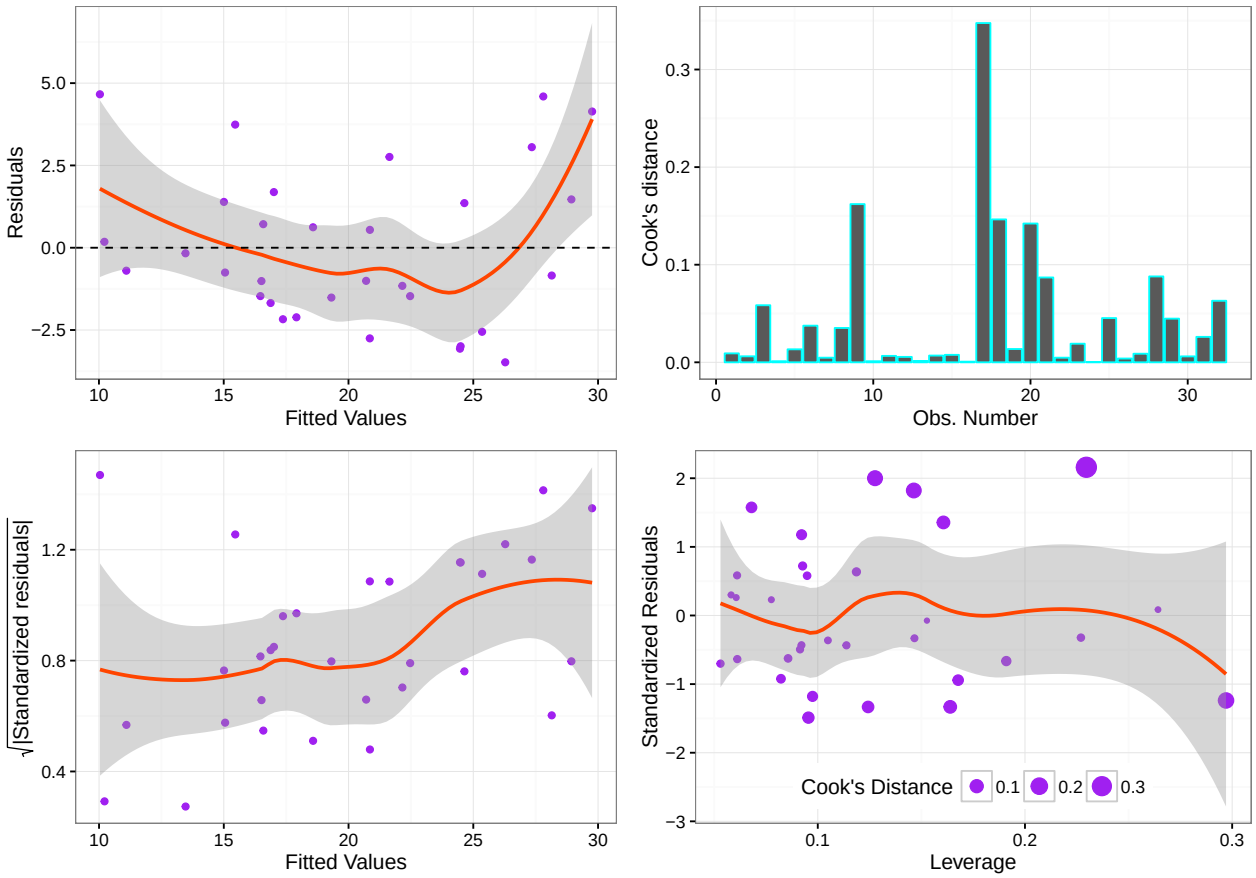
3. It helps to find if there is a systematic pattern in a data, where patterns in residual plots generally indicate poor aspect of model fit.
4. It helps to detect “Heteroskedasticity” (non constant variance).

```
qqPlot(Final_fit, main="Normal Q-Q plot")
```




```
library(broom)
df <- augment(Final_fit)

f1 <- ggplot(df, aes(x = .fitted, y = .resid)) + geom_point(color="purple") +
      theme_bw() + stat_smooth(color="orangered") +
      geom_hline(yintercept=0, col="black", linetype="dashed")+
      xlab("Fitted Values")+ylab("Residuals")

f2 <- ggplot(df, aes(.fitted, sqrt(abs(.std.resid))))+geom_point(color="purple") +
      stat_smooth(color="orangered")+xlab("Fitted Values") +
      ylab(expression(sqrt("|Standardized residuals|"))) +
      theme_bw()

f3 <- ggplot(df, aes(seq_along(.cooksd), .cooksd)) +
      geom_bar(stat="identity", position="identity",color="cyan") +
      xlab("Obs. Number")+ylab("Cook's distance") +
      theme_bw()

f4 <- ggplot(df, aes(.hat, .std.resid)) +
      geom_point(aes(size=.cooksd), na.rm=TRUE,color="purple") +
      stat_smooth(na.rm=TRUE, color="orangered") +
      xlab("Leverage")+ylab("Standardized Residuals") +
      scale_size_continuous("Cook's Distance", range=c(1,5)) +
```

```
      theme_bw()+theme(legend.position=c(0.5,0.1), legend.direction="horizontal")

suppressWarnings(multiplot(f1,f2,f3,f4,cols=2))
```



Two plots on the left, where we show "residuals" and "standardized residuals" vs "fitted values" plots indicate that there is neither systematic pattern nor "Heteroskedasticity" in the residual plots. Bottom right plot shows that 0<leverage<1 meaning we don't have high leverage problem. Cooks distance evaluates overall change in the coefficients when the i^{th} point is deleted. Mostly affected data point, which shows highest cooks distance is Chrysler Imperial with the value= 0.3475974.

Conclusion

Taking into account all analysis and assumptions based on provided data, I conclude that

- **Automatic cars, on average, consume slightly more fuel than manual cars.**
- **The best way to quantify the the MPG difference between automatic and manual transmissions for 'mtcars' dataset would be performing multivariate linear regression using only "wt" and "qseq" variables as regressors.**