RM CP Report

Zack Gouttel

14/07/2020

Executive Summary

This analysis is performed for the Motor Trend, a magazine about the automobile industry. it was achieved by looking at a data set of a collection of cars, we are exploring the relationship between a set of variables and miles per gallon (MPG) as outcome, in an attempt to answer the following questions:

"Is an automatic or manual transmission better for MPG?"

"Quantify the MPG difference between automatic and manual transmissions"

We performed exploratory data analyses, and used hypothesis testing and linear regression as methodologies to make inference in order to answer these questions. We also established both simple and multivariate linear regression analysis. However the result of the multivariable regression model is more promising as it includes the potential effect of other variables on MPG.

Using model selection strategy, we found out that among all variables *weight* and *quarter mile time* (acceleration) have significant *impact* in quantifying the *difference of MPG between automatic and manual transmission cars.*

#A First look to the data

We used the mtcars Dataset that was extracted from the 1974 Motor Trend US magazine, that comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). Below is a look into the first 15 columns of the dataset, each observation "row" represent a car model.

```r
library(knitr)
library(printr)
```

```
## Registered S3 method overwritten by 'printr':
##   method                from
##   knit_print.data.frame rmarkdown
```

```r
kable(head(mtcars,15),align = 'c')
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4

#Exploratory Analyses

by performing some initial explaratory data analysis, we get a better idea of the existing patterns between variables in the data set. Normally in regression analysis scatter plots are the most effective. Below we create nice pairwise scatter plots in order to investigate the relationship between all the variables in this data set.

```
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

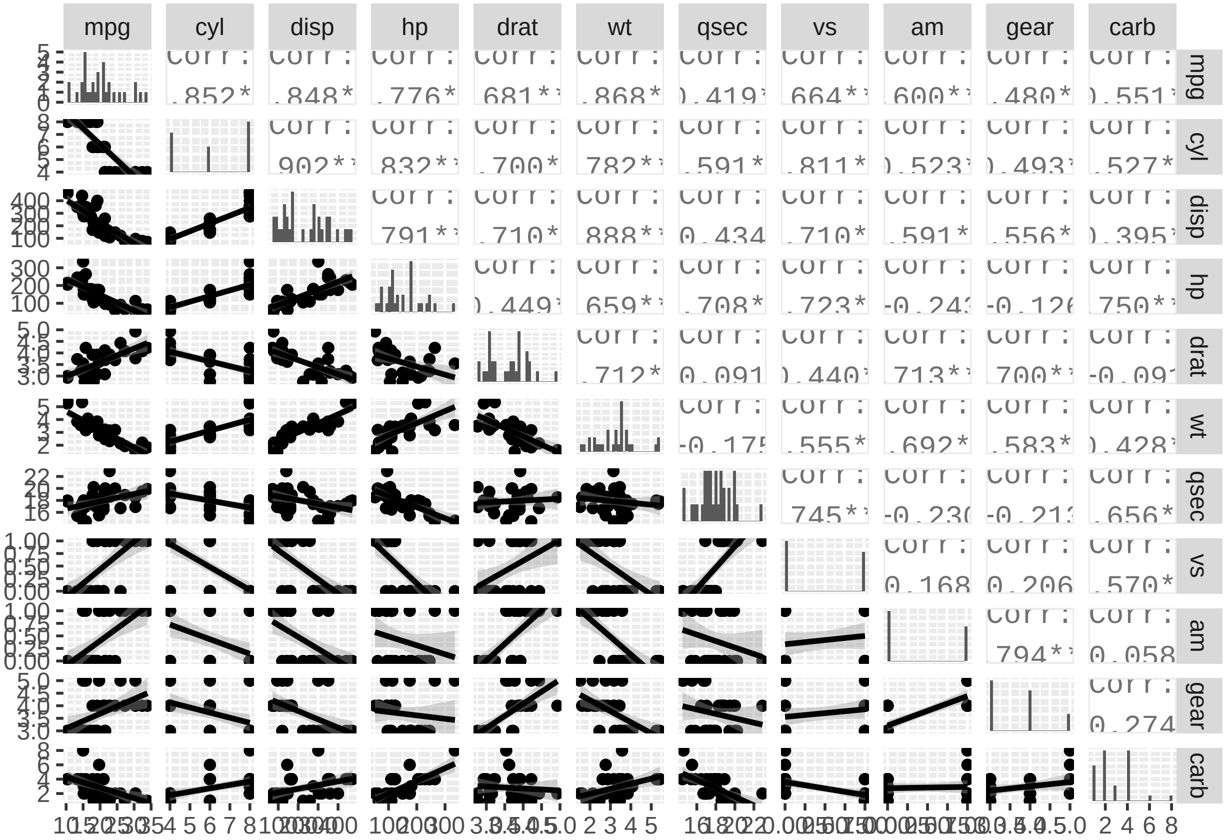
```
library(ggplot2)

ggpairs(mtcars,
        lower = list(continuous = "smooth",method = "loess", colour="blue"),
        diag=list(continuous="bar", colour="blue"),
        upper=list(corSize=15),
        axisLabels='show')
```

```
## Warning in check_and_set_ggpairs_defaults("diag", diag, continuous =
## "densityDiag", : Changing diag$continuous from 'bar' to 'barDiag'
```

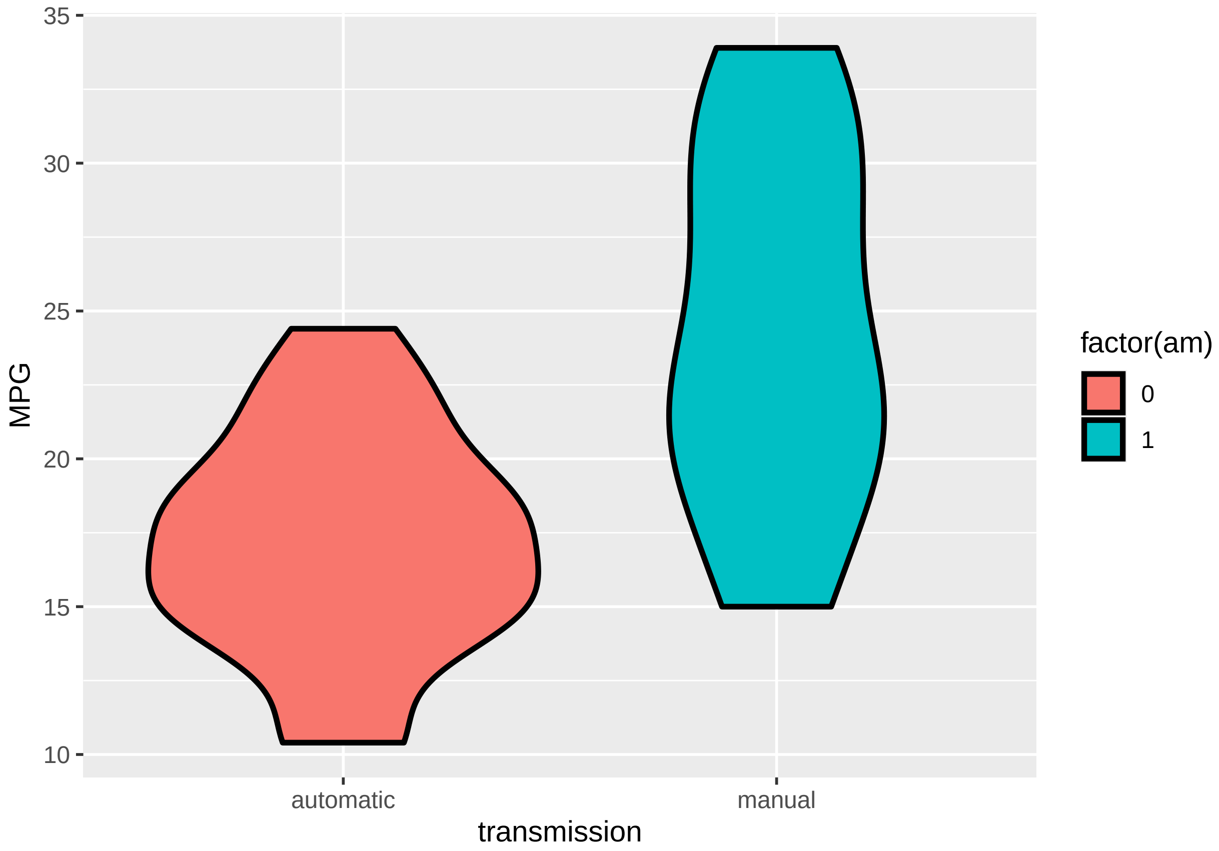
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



we can also create a Violin plot of MPG by automatic and manual transmissions to show the relationship that we may initially suggest as . In our dataset 0 represents an automatic transmission and 1 means a manual transmission.

```
ggplot(mtcars, aes(y=mpg, x=factor(am, labels = c("automatic", "manual")), fill=factor(am)))+
        geom_violin(colour="black", size=1)+
        xlab("transmission") + ylab("MPG")
```



from this initial exploration of the data through the violin plots we can say that automatic cars are less efficient in their fuel consumption, which is visible through the low mpg that is presented for automatic in this figure. However this may be dumbfound luck since it could have been that we picked out low end auto cars and high end manual transmission cars.

#Model Fitting and Hypothesis testing

To answer our first question, we test the hypothesis that cars with an automatic transmission use more fuel than cars with a manual transmission. To compare two samples to see if they have different means, we use two samples T-test.

```
test <- t.test(mpg ~ am, data= mtcars, var.equal = FALSE, paired=FALSE ,conf.level = .95)
result <- data.frame( "t-statistic"  = test$statistic,
                       "df" = test$parameter,
                        "p-value"  = test$p.value,
                        "lower CL" = test$conf.int[1],
                        "upper CL" = test$conf.int[2],
                        "automatic mean" = test$estimate[1],
                        "manual mean" = test$estimate[2],
                        row.names = "")
kable(x = round(result,3),align = 'c')
```

t.statistic	df	p.value	lower.CL	upper.CL	automatic.mean	manual.mean
-3.767	18.332	0.001	-11.28	-3.21	17.147	24.392

Here, the p-value shows that the probability for this apparent difference between the two groups happening by chance is very low. The confidence interval also describes how much lower the miles per gallon is in manual cars than it is in automatic cars. We can be confident that the true difference is between 3.2 and 11.3.

Simple Linear Model Regression

We can also fit factor variables as regressors and come up with thing like analysis of variance as a special case of linear regression models.we only take the am variable as a factor one here.

```
mtcars$amfactor <- factor(mtcars$am, labels = c("automatic", "manual"))
summary(lm(mpg ~ factor(amfactor), data = mtcars))$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
factor(amfactor)manual	7.244939	1.764422	4.106127	0.000285

All the estimates provided here are by taking the automatic car as reference level. 17.14 "intercept value" is simply the mean MPG of automatic transmission. The slope of 7.24 is the change in the mean between manual transmission and automatic transmission. You can verify that from the plot as well. The p-value of 0.000285 for the mean MPG difference between manual and automatic transmission is significant. Therefore we conclude that according to this model manual transmission is more fuel efficient.

##Multivariable Linear Model Regression

Modeling based on only one predictor variable is rarely sufficient or good enough as there may be other predictor variables affect MPG and therefore affect the difference in MPG by transmission. So the univariate model in this case is only one piece of the puzzle. Thus the need to use multivariable linear regression to develop a model that includes the effect of other variables.

Model Selection Procedure

what combination of predictors will best predict fuel efficiency? Which predictors increase our accuracy by a statistically significant amount? We might be able to guess that from some of the trends on the graph, but really we want to perform a statistical test to determine which predictors are significant, and to determine the ideal formula for predicting our outcome.

Including variables that we should't have increases actual standard errors of the regression variables.Thus we don't want to idly throw variables into the model. To confirm this fact, you can see below that if we include all the variables, none of them will be a significant predictor of MPG (judging by p-value at the 95% confidence level).

```
summary(lm(mpg ~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb, data = mtcars))$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3033742	18.7178844	0.6573058	0.5181244
cyl	-0.1114405	1.0450234	-0.1066392	0.9160874
disp	0.0133352	0.0178575	0.7467585	0.4634887
hp	-0.0214821	0.0217686	-0.9868407	0.3349553
drat	0.7871110	1.6353731	0.4813036	0.6352779
wt	-3.7153039	1.8944143	-1.9611887	0.0632522
qsec	0.8210407	0.7308448	1.1234133	0.2739413
factor(vs)1	0.3177628	2.1045086	0.1509915	0.8814235
factor(am)1	2.5202269	2.0566506	1.2254035	0.2339897
gear	0.6554130	1.4932600	0.4389142	0.6652064
carb	-0.1994193	0.8287525	-0.2406258	0.8121787

Detecting Colinearity

One of the major problems with multivariate regression is co-linearity ergo two or more predictor variables are highly correlated, and they are both entered into a regression model, it increases the true standard error inducing very unstable estimates of the slope. We can assess the collinearity by variance inflation factor (VIF). Lets look at the variance inflation factors if we throw all the variables into the model.

```r
library(car)
```

```
## Loading required package: carData
```

```r
fitvif <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb, data = mtcars)
kable(vif(fitvif),align = 'c')
```

	x
cyl	15.373833
disp	21.620241
hp	9.832037
drat	3.374620
wt	15.164887
qsec	7.527958
factor(vs)	4.965873
factor(am)	4.648487
gear	5.357452
carb	7.908747

Values for the VIF that are greater than 10 are considered large. We should also pay attention to VIF values between 5 and 10. At these point we might consider leaving only one of these variables in the model.

Step-by-Step Selection Method

Among available methods we decided to perform a step-by-step selection to help us select a subset of variables that best explain the MPG. Please note that we also treat the vc variable as a categorical variable.

```r
library(MASS)
fit <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb, data = mtcars)
step <- stepAIC(fit, direction="both", trace=FALSE)
summary(step)$coeff
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
factor(am)1	2.935837	1.4109045	2.080819	0.0467155

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

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

In addition to transmission,we can see that the weight of the vehicle as well as acceleration speed have the highest relation to explaining the variation in mpg. The adjusted R^2 is 84% which means that the model explains 84% of the variation in mpg, indicating a highly predictive model.

###Nested Likelihood ration testing

In order to verify the result of the step-by-step selection model, we also perform nested likelihood ratio tests

```r
fit1 <- lm(mpg ~ factor(am), data = mtcars)
fit2 <- lm(mpg ~ factor(am)+wt, data = mtcars)
fit3 <- lm(mpg ~ factor(am)+wt+qsec, data = mtcars)
fit4 <- lm(mpg ~ factor(am)+wt+qsec+hp, data = mtcars)
fit5 <- lm(mpg ~ factor(am)+wt+qsec+hp+drat, data = mtcars)
anova(fit1, fit2, fit3, fit4, fit5)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.8966	NA	NA	NA	NA
29	278.3197	1	442.576902	72.5359307	0.0000000
28	169.2859	1	109.033768	17.8700375	0.0002579
27	160.0665	1	9.219469	1.5110205	0.2299925
26	158.6386	1	1.427847	0.2340163	0.6326111

the result is consistent with step-by-step selection model and adding any more variables in addition to wt, am and qsec will dramatically increase the variation in the model, and the p-value immediately becomes insignificant.

Final Model Fit

Now that we have confirmed our final model, we may proceed to fitting it

```r
finalfit <- lm(mpg ~ wt+qsec+factor(am), data = mtcars)
summary(finalfit)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
factor(am)1	2.935837	1.4109045	2.080819	0.0467155

Quantifying the Change in MPG between Manuel and Auto Transmisson cars:

all the variables now are statistically significant. This model explains 84% of the variance in miles per gallon (mpg). when we read the coefficient for am, we say that, *on average, manual transmission cars have 2.94 MPGs more than automatic transmission cars.* However **this effect was much higher than when we did not adjust for weight and qsec**.