Coursera Regression Course Project

Cody Frisby

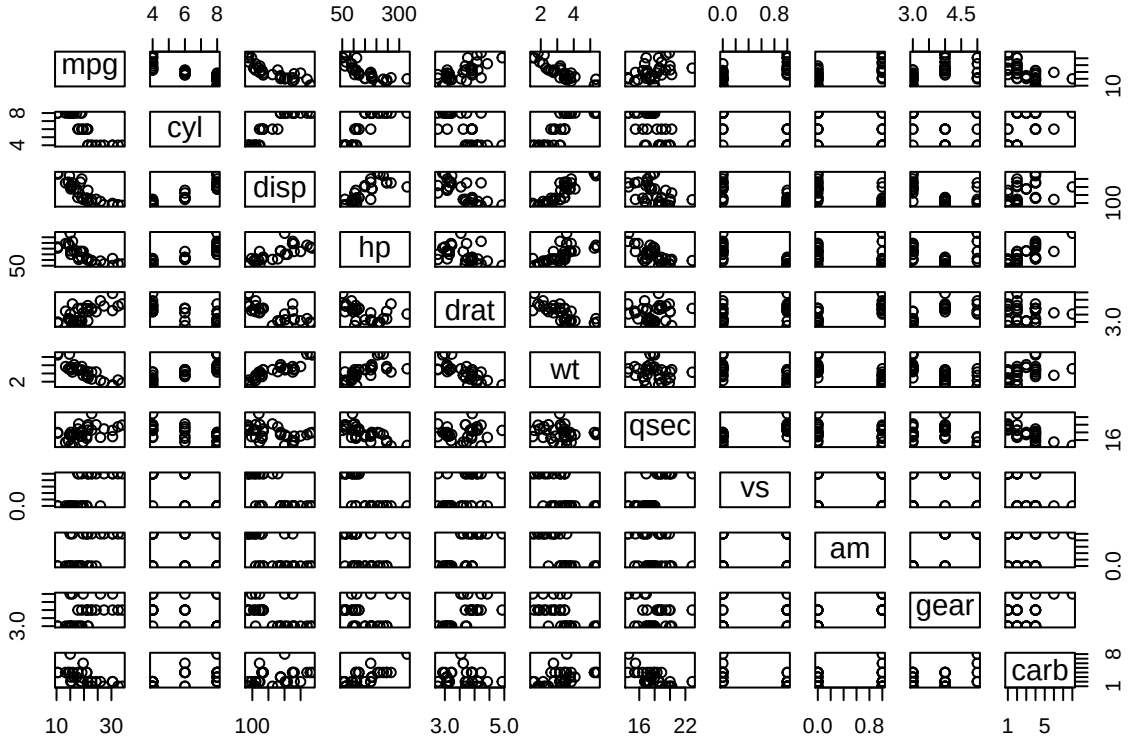
April 10, 2016

Executive Summary

This study contains data that was extracted from the 1974 *Motor Trend* magazine. There are 32 total observations with 11 total variables (?mtcars in the R console). Here I attempt to answer the question of whether manual or automatic is better, and what the difference is if any, for miles per gallon and what variables are significant in predicting **mpg**.

Exploratory Analysis

Let's start by looking at a scatter plot matrix of the data.
Note: the R code will be at the end of the document. I will be using echo=FALSE in my knitr document to hide the code inline.



Next, I fit a model using all the variables (no interactions) and took a look at the variance inflation factors.

cyl	15.373833
disp	21.620241
hp	9.832037
drat	3.374620
wt	15.164887
qsec	7.527958
vs	4.965873

am	4.648487
gear	5.357452
carb	7.908747

We can see that a few of the variables are going to need to be removed from the final model. Using the leaps package and regsubsets() function from that package I narrowed down the best models using BICs. My final model is

$$mpg_i = \beta_0 + \beta_1 wt_i + \beta_2 qsec_i + \beta_3 am_i + \varepsilon_i$$

and when fit to the data our prediction function is

$$mpg = 9.6177805 - 3.9165037 wt + 1.225886 qsec + 2.9358372 am$$

1) Is an automatic or manual transmission better for MPGs?

The answer to this question is **manual** transmission. A **manual** transmission will get 2.9358372 more **mpg** than an **automatic**, all other variables held constant.

2) Difference between auto and manual?

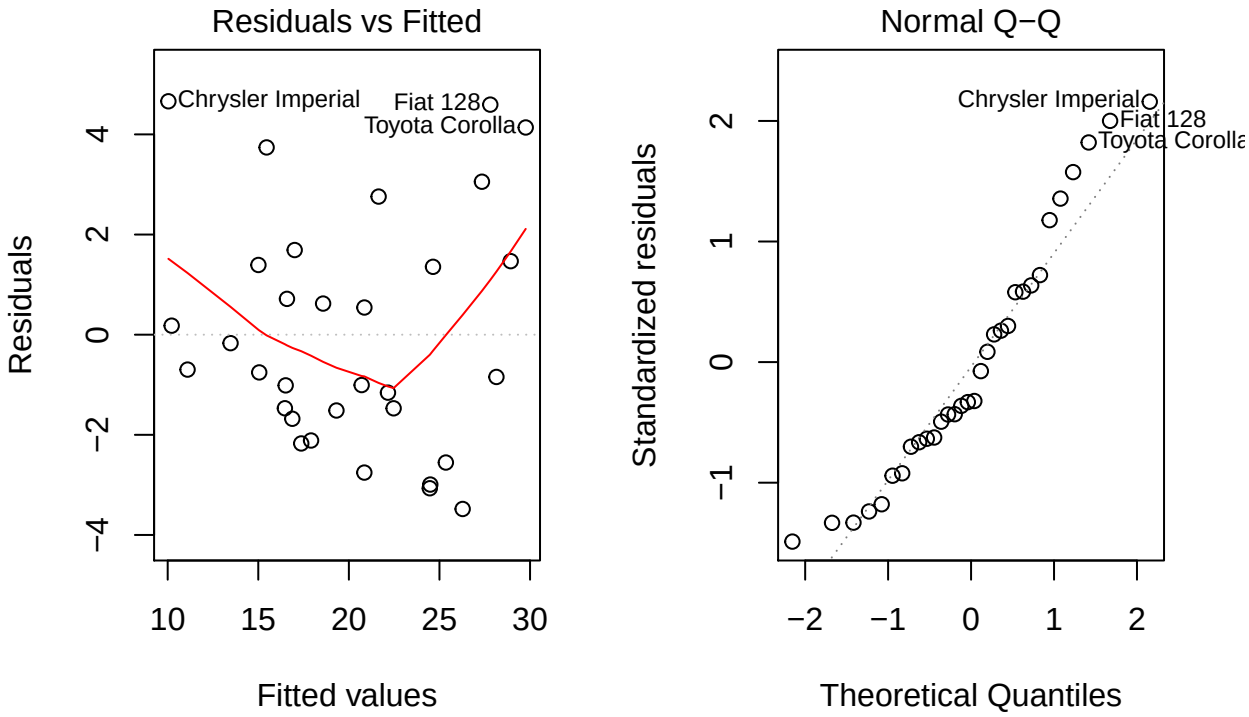
Firstly, I display the coefficients of the model I chose.

	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
am	2.935837	1.4109045	2.080819	0.0467155

Here we can see that our model includes **wt** (the car's mass in 1000 lbs), **qsec** (the cars 1/4 mile time), and **am** (auto or manual transmission). There is a negative effect from **wt** meaning that as the cars' mass increases its expected **mpg**s decrease by -3.9165037. Similarly, as a cars quarter mile time increases its expected **mpg**s increase by 1.225886. And lastly, when the car has a manual transmission, its **mpg**s are expected to be 2.9358372 higher than when it has an automatic transmission, holding the other variables constant. A 95% confidence interval is [0.0457303, 5.8259441], showing that at worst **manual** will beat **auto** by only 0.0457303.

Model Adequacy

Here, I take a look at the model residuals checking the assumptions of linear models.



Our residual plots look OK. The assumptions of equal variance, linearity, and normality are OK.

Conclusion

Although all the terms in our model are statistically significant, at the $\alpha = 0.05$ level, the confidence interval for **am** is quite wide, 0.0457303, 5.8259441. I chose this model over the other possibilities because it was the most parsimonious model in the top 10 (based on BIC) that contained **am** with the lowest BIC, Cp, and close to the highest adjusted R^2 value.

I display the top ten models (BIC) for reference:

	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
3 (1)					*	*		*		
2 (1)	*				*					
2 (2)			*		*					
2 (3)					*	*				
3 (2)	*		*		*					
3 (3)	*				*					*
4 (1)			*		*	*		*		
4 (2)					*	*		*		*
3 (4)			*		*			*		
3 (5)	*				*	*				

R code:

```r
# first let's get the data into R
df <- mtcars
plot(df) # scatterplot matrix, exploratory
fit.all <- lm(mpg ~ ., df)
# here we look at the variance inflation factors of all the
# variables in the model.
knitr::kable(car::vif(fit.all))
# here we take a look at all subsets using the leaps package
# ten best per number of predictors
mods <- leaps::regsubsets(x = df[,2:11], y = df[,1], nbest=10)
# extract the BIC, Cp, and adj R^2 from the mods object
mods.s <- cbind(summary(mods)$bic, summary(mods)$cp,
                      summary(mods)$adjr2)
colnames(mods.s) <- c("BIC", "Cp", "AdjR2")
# create an R object of all subsets
mods.v <- summary(mods)$outmat
# creates a position vector, ordering by BIC ascending
best.bic <- order(mods.s[,1])
# I will go will the model lm(mpg ~ wt+qsec+am)
# fit the "best" model based on BIC
fit <- lm(mpg ~ wt+qsec+am, df)
# store the coefs of our model
betas <- summary(fit)$coef[,1]
knitr::kable(car::vif(fit.all))
knitr::kable(summary(fit)$coef)
# residual plots
par(mfrow=c(1,2))
plot(fit, which = c(1,2))
confint(fit)[4,]
knitr::kable(mods.v[best.bic[1:10], ])
```