Regression Models project: Exploration of factors affecting MPG

Nick Plummer

20/08/2018

Exectutive Summary

Analysis of the `mtcars` dataset, examining the fuel efficiency and 10 aspects of design and performance for 32 cars produced in 1973-1974, demonstrated that cars with a manual transmission (n=13) were on average 7.2mpg more efficient than automatic cars (n=19). However, transmission type is not a particularly good predictor of fuel efficiency, with the number of cylinders and weight of the vehicle both being much stronger predictors of fuel efficiency than transmission type.

Exploratory analysis

MPG vs transmission

```
data(mtcars)
fit0 <- lm(mpg ~ am, data = mtcars)
print(xtable(fit0), comment=FALSE)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.1474	1.1246	15.25	0.0000
am	7.2449	1.7644	4.11	0.0003

There is a straightforward linear relationship between efficiency and transmission type ($p<0.001$), with manual cars' mean efficiency 7.25 mpg higher than automatic. The low R^2 means transmission only accounts for 34% of the variability in efficiency in this model.

MPG vs other factors

We can use an analysis of variance model (calls `lm` for each stratum) to compare all possible preductor variables (removing `qsec`, time to cover 1/4 mile, as a dependent variable). However there will potentially be correlation between variables (e.g. a car with more cylinders is likely to develop more horsepower), which can be assessed by plotting the mtcars dataset, and using `cor` matrices (see appendix for table and plot).

```
cor <- cor(mtcars)
cor[upper.tri(cor)] <- NA
# plot(mtcars)
# print(xtable(cor), comment=FALSE)
```

This demonstrates a strong relationship between displacement and cylinders (0.9), displacement and weight (0.89) and weight and cylinders (0.78). A link between displacment and cylinders seems physically plausible; the correlation with weight more due to heavier cars requiring a larger engine.

```
# Set factor variables correctly
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs, labels=c("V-engine", "Straight engine"))
```

```
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))

# Run aov
all_fit <- aov(mpg ~ .-qsec, data = mtcars)
print(xtable(all_fit), comment=FALSE)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cyl	2	824.78	412.39	54.24	0.0000
disp	1	57.64	57.64	7.58	0.0141
hp	1	18.50	18.50	2.43	0.1383
drat	1	11.91	11.91	1.57	0.2286
wt	1	55.79	55.79	7.34	0.0155
vs	1	1.39	1.39	0.18	0.6744
am	1	13.37	13.37	1.76	0.2034
gear	2	3.39	1.70	0.22	0.8026
carb	5	17.62	3.52	0.46	0.7977
Residuals	16	121.64	7.60		

The analysis of varience model suggests that `cyl` (number of cylinders), `hp` (horsepower) and `disp` (engine displacement in cubic inches) seem most strongly correlated with MPG ($p<0.05$),

The next step will be to build models using combinations of the variables identified in the previous section, and including the link with transmission type.

```
fit1 <- lm(mpg ~ cyl, data=mtcars)
fit2 <- lm(mpg ~ wt, data = mtcars)
fit3 <- lm(mpg ~ cyl + wt, data = mtcars)
fit4 <- lm(mpg ~ disp, data = mtcars)
fit5 <- lm(mpg ~ disp + cyl, data = mtcars)
fit6 <- lm(mpg ~ disp + cyl + wt, data = mtcars)
```

The model with the highest adjusted R squared value is `fit3`, which models fuel efficiency as a function of number of cyclinders and weight (`R^2` = 0.8200146), lower than models which include displacement. This agrees with our previous analysis that although cylinders, weight and displacement all have a signficant relationship to fuel efficency, displacment is strongly related to cylinders so is a confounder.

```
fit7 <- lm(mpg ~ cyl + wt +am, data = mtcars)
print(xtable(fit7), comment=FALSE)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.7536	2.8135	12.00	0.0000
cyl6	-4.2573	1.4112	-3.02	0.0055
cyl8	-6.0791	1.6837	-3.61	0.0012
wt	-3.1496	0.9080	-3.47	0.0018
amManual	0.1501	1.3002	0.12	0.9089

Addition of a transmission term to the model decreases its accuracy (`R^2` = 0.8134405). The coefficient for transmission type is decreased from that in `fit0`, predicting that a car with a manual transmission would have an improved fuel efficiency of 0.18 mpg over an automatic transmission with the same number of cylinders and weight.

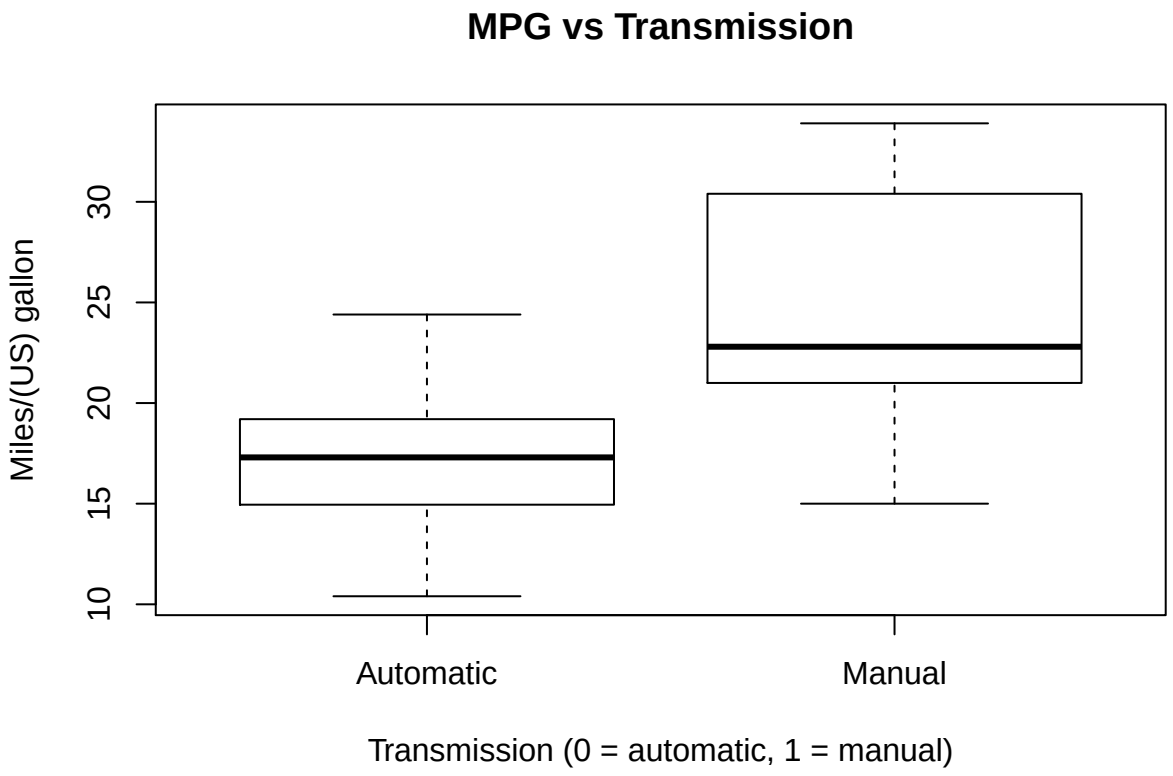

Figure 1: MPG vs transmission

Appendix

MPG vs transmission

```
boxplot(mpg ~ am, data = mtcars,
  xlab="Transmission (0 = automatic, 1 = manual)",
  ylab="Miles/(US) gallon",
  main="MPG vs Transmission")
```

MPG vs other factors

```
plot(mtcars)
```

```
print(xtable(cor), comment=FALSE)
```

```
par(mfrow=c(2,2))
plot(fit3)
```

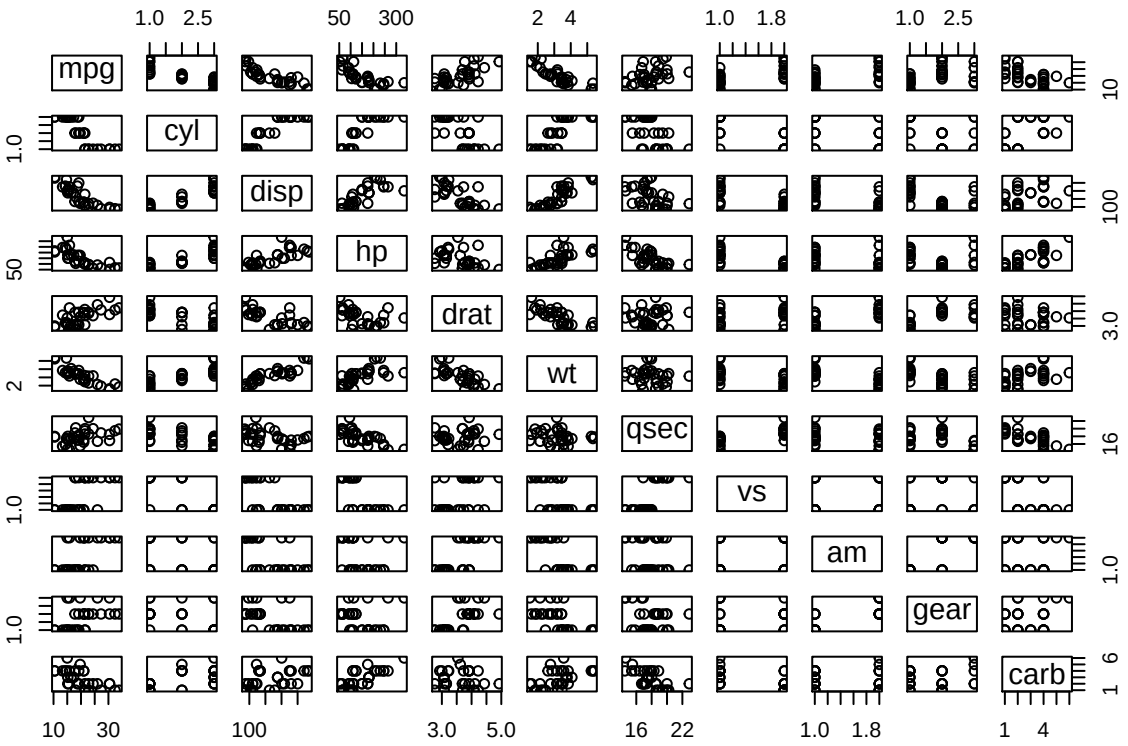

Figure 2: Graphical representation of correlations within mtcars

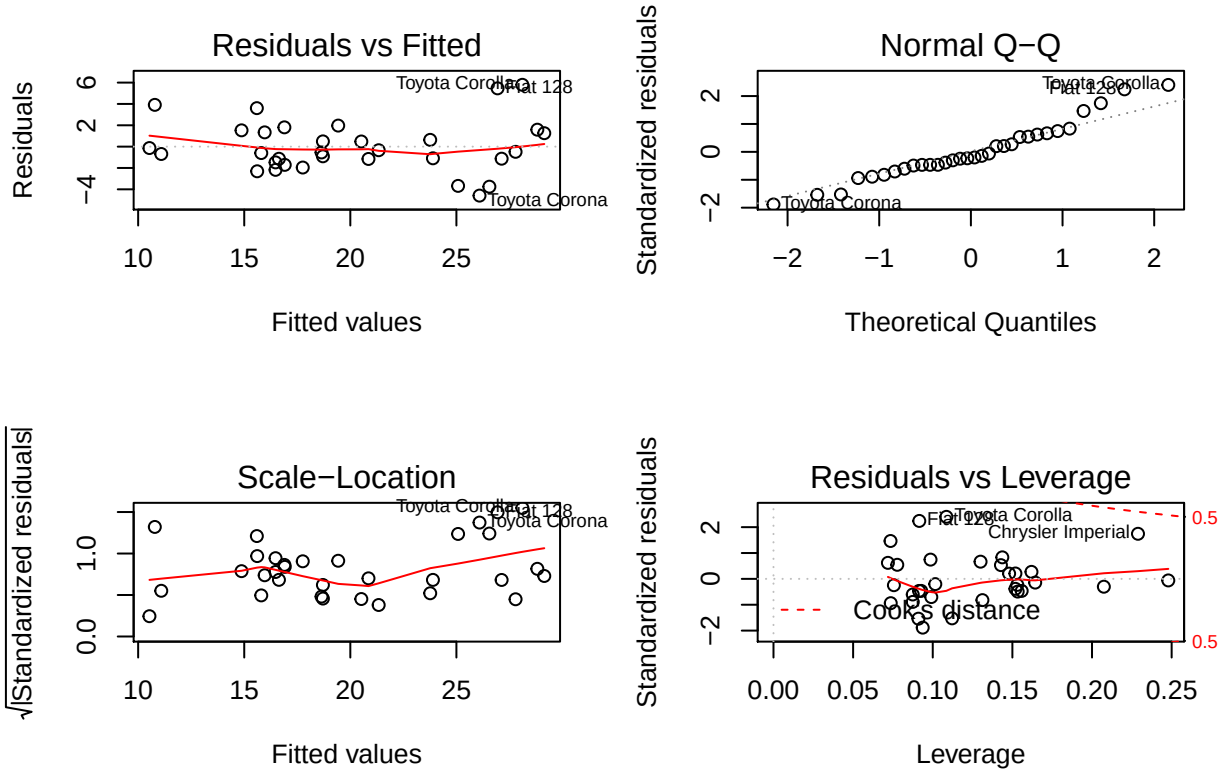

Figure 3: Diagnostic plots of residuals for model 3

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00										
cyl	-0.85	1.00									
disp	-0.85	0.90	1.00								
hp	-0.78	0.83	0.79	1.00							
drat	0.68	-0.70	-0.71	-0.45	1.00						
wt	-0.87	0.78	0.89	0.66	-0.71	1.00					
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1.00				
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1.00			
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1.00		
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1.00	
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1.00