

Regression Models Course Project

Andrew E. Davidson

May 23, 2015

I ran into problems running Knit PDF in R Studio

Unfortunately my report is not formatted as nicely as I might like. Knit did not work on my Mac. As a work around I set the Knit format to html and used my browser to convert it into PDF. This makes the report take up much more space

Explore the mtcars data set

Probably the most useful thing to do is read the data set description. Using `? mtcars` we see the data set contains 11 variables, including such as engine size, number of gears, ... We'll need to adjust our models to account for these variable.

set up factors

Using `str` it turns out `cyl`, `gear`, `am`, and `carb` are numeric, not factors. We'll need to convert them

```
str(mtcars)
```

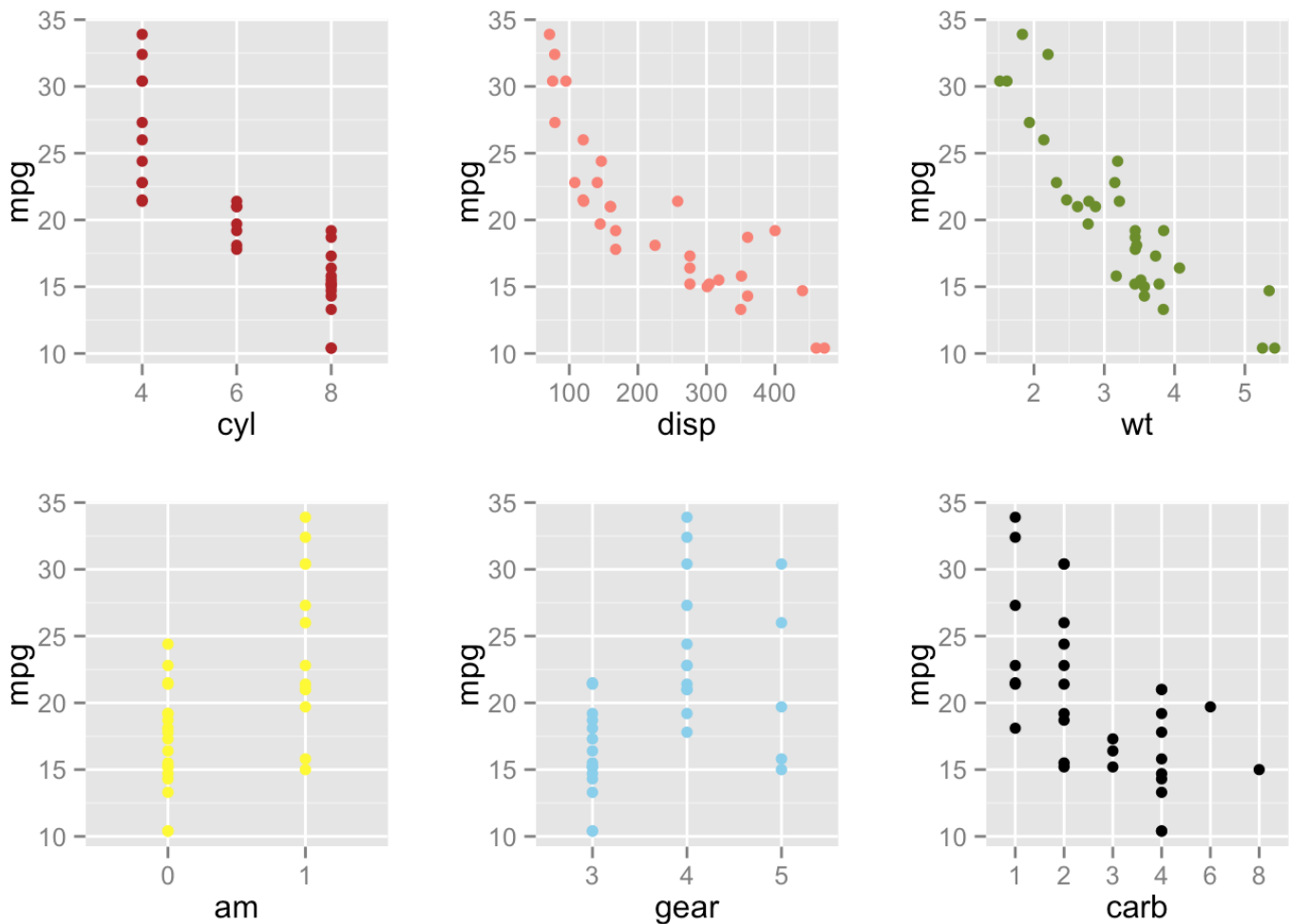
```
## 'data.frame':   32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 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 : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
mtcars$am <- factor(mtcars$am)
mtcars$cyl <- factor(mtcars$cyl)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

Lets take a look at some of the raw sample data

##		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
##	Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
##	Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
##	Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Using scatter plots we can identify variables that might be good predictors of miles per gallon. We are looking for variables that seem to have strong linear tendency. For example displacement. We are also looking for variables that might be confounding. For example we would expect the more gears a car has the better the miles per gallon would be. Just looking at the scatter plot, it looks like cars with 4 gears might do better than cars with either 3 or 5 gears suggesting that something else is going on. The confounding effect may have to do with how the test was run. Did they take the car out on track and run at a high speed for long time? If so we would expect cars with 5 gears to consistently outperform the others holding things like weight, engine size, ... constant



Is an automatic or manual transmission better for MPG?

To solve this problem, we need to create a linear model that predicts MPG. We can not simply using the factor variable am, (automatic or manual transmission) as a predictor. We'll need to adjust for other factors such as weight, number of gears, engine size, ... To keep things simple we will assume there are no interaction between variables. That is to say our model does not have predictors variable of the form $wt * disp$.

First we need to decide what variables to use in our model. If we create a linear model using all the variable and look at the col $Pr(>|t|)$ we do not find a beta that appears to be significant. The best predictor seems to be vehicle weight. the $Pr()$ value is still lower than we would normally like. The Estimate for wt is -4.5, which means for every 1000 lbs increase in vehicle weight we would expect MPG to decrease by 4.5. (Assuming we did not change any of the other variables like engine size, transmission, ...)

```
summary(lm(mpg ~ ., data=mtcars))$coefficients
```

##	Estimate	Std. Error	t value	$Pr(> t)$
## (Intercept)	23.87913	20.06582	1.19004	0.25253
## cyl6	-2.64870	3.04089	-0.87103	0.39747
## cyl8	-0.33616	7.15954	-0.04695	0.96317
## disp	0.03555	0.03190	1.11433	0.28267
## hp	-0.07051	0.03943	-1.78835	0.09393
## drat	1.18283	2.48348	0.47628	0.64074
## wt	-4.52978	2.53875	-1.78426	0.09462
## qsec	0.36784	0.93540	0.39325	0.69967
## vs	1.93085	2.87126	0.67248	0.51151
## am1	1.21212	3.21355	0.37719	0.71132
## gear4	1.11435	3.79952	0.29329	0.77332
## gear5	2.52840	3.73636	0.67670	0.50890
## carb2	-0.97935	2.31797	-0.42250	0.67865
## carb3	2.99964	4.29355	0.69864	0.49547
## carb4	1.09142	4.44962	0.24528	0.80956
## carb6	4.47757	6.38406	0.70137	0.49381
## carb8	7.25041	8.36057	0.86722	0.39948

Whats interesting is if we look at the unadjust model. That is to say the model with a single predictor variable of weight, the is t value is significant! The size of the effect is also much stronger.

```
summary(lm(mpg ~ wt, data=mtcars))$coefficients
```

##	Estimate	Std. Error	t value	$Pr(> t)$
## (Intercept)	37.285	1.8776	19.858	8.242e-19
## wt	-5.344	0.5591	-9.559	1.294e-10

The next step to try a couple of different models. That is to say choose different covariants and see if we can improve the t values. I got the most significant results with predictors of wt and cyl. Adding gear, drat, and disp did not reduced significance so in the interested of having the simplest possible model I did not include them in my final model

Because am is our factor variable with the lowest level of the coveraignts in our model, R will automatically select it as the reference variable. That is to say the expected value in mpg for cars with automatic transmissions.

```
model <- lm(mpg ~ am + wt + cyl + carb, data=mtcars)
summary(model)$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	30.8849	3.805	8.1175	4.630e-08
## am1	1.8736	1.779	1.0530	3.038e-01
## wt	-2.3318	1.130	-2.0635	5.106e-02
## cyl6	-2.8225	1.860	-1.5172	1.435e-01
## cyl8	-5.2330	2.014	-2.5982	1.641e-02
## carb2	-0.4656	1.446	-0.3219	7.506e-01
## carb3	-0.3510	2.289	-0.1534	8.795e-01
## carb4	-2.3001	1.902	-1.2090	2.395e-01
## carb6	-3.7769	3.461	-1.0911	2.870e-01
## carb8	-4.2009	3.608	-1.1644	2.567e-01

Conclusion

The expect MPG for cars with automatic transmission is 30.89. Having a manual transmission increases the MPG by 1.87.

Graph of Residuals

If our model is good, we would expect our residual point graph to appear random. That is to say we should not find any sort of pattern. (Our model looks good)

```
r <- resid(model)
df <- data.frame(mtcars$mpg, r)
colnames(df) <- c("mpg", "residual")
g <- ggplot(data=df, aes(y=mpg, x=residual))
g <- g + geom_point(size=5, colour="blue")
g
```

