Coursera Regression Models - Motor Trend Project

Raphael Carvalho

10/02/2019

Questions to be answered

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- “Is an automatic or manual transmission better for MPG”
- “Quantify the MPG difference between automatic and manual transmissions”

Analysis

Loading the packages and datasets…

```
## Observations: 32
## Variables: 11
## $ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2,…
## $ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4,…
## $ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140…
## $ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 18…
## $ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92,…
## $ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.1…
## $ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.…
## $ vs   <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1,…
## $ am   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1,…
## $ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4,…
## $ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1,…
```

Transforming some variables into factors…

Getting the difference in MPG between Manual and Automatic…

To do that, we’ll first compare the mean MPG for both_

```
aggregate(mpg~am, data = mtcars, mean)
```

am <fctr>	mpg <dbl>
Automatic	17.14737
Manual	24.39231

2 rows

As you can see, manual cars have an average MPG 7.25 higher than automatic cars. Looking at this results we can hypothetize that manual cars are more efficient when it comes down to MPG. But, to determine if this difference that we observed is statistically significant, we should run a T-test, as seen below:

```
D_automatic <- mtcars[mtcars$am == "Automatic",]
D_manual <- mtcars[mtcars$am == "Manual",]
t.test(D_automatic$mpg, D_manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  D_automatic$mpg and D_manual$mpg
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

As you can see by the test results, the p-value is very close to zero, meaning that we should take the alternative hypothesis that is that the true difference in means is not equal to zero. So, to make things tangible, let's run a linear regression model to quantify this difference:

```
model <- lm(mpg ~ am, data = mtcars)
summary(model)
```

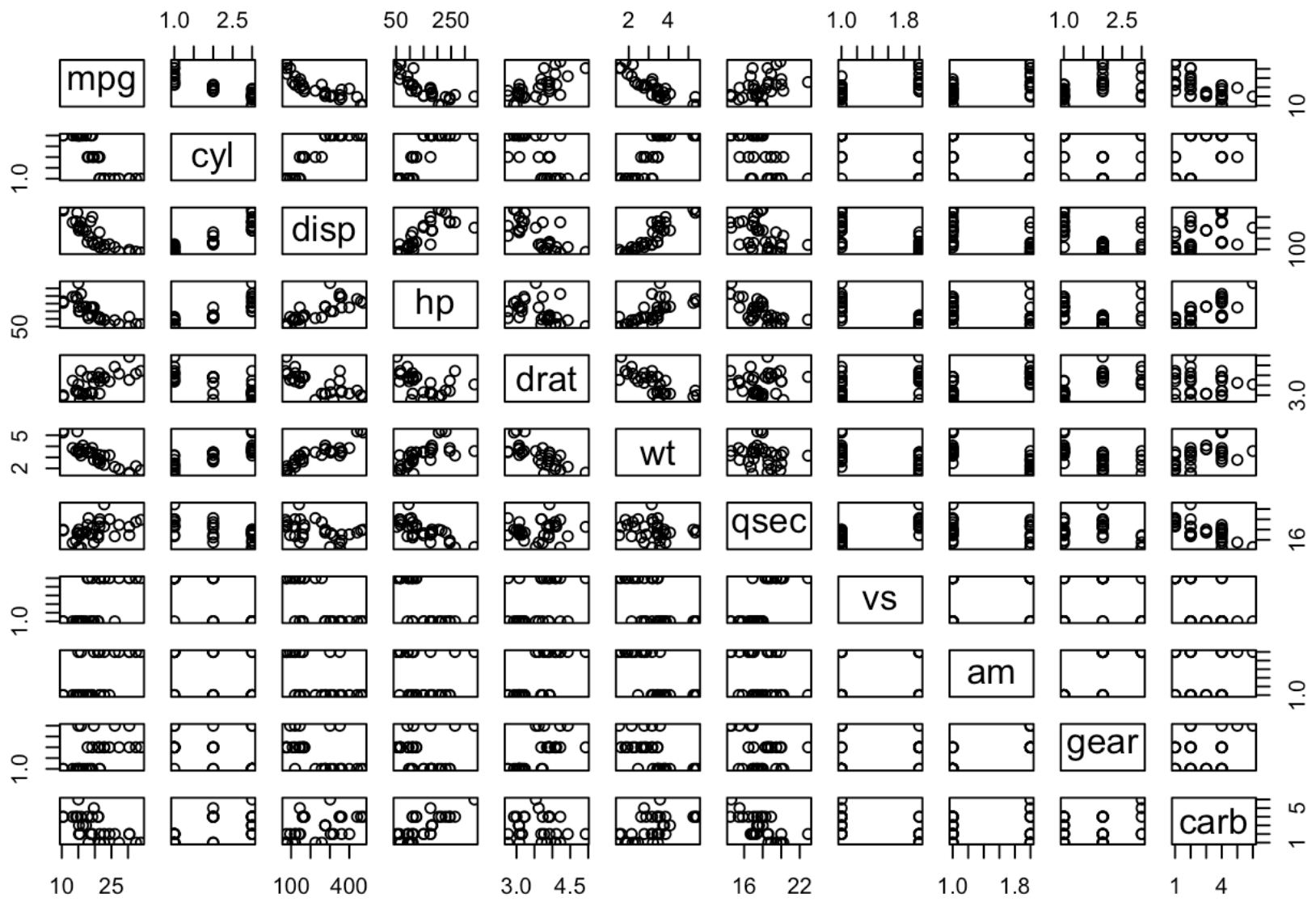
```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Looking on the results of the model, we can see that if our car is automatic, we should expect a MPG of 17.147 (intercept), otherwise, we should add a slope of 7.245, giving us 24.392 MPG, just as we saw when we compared the means for each group of cars.

Despite all variables been signifficant to the model, we can see that we've got a R^2 of .36. This tells us that the model only explain 36% of the variance of the data. This is not good at all. Great models explain 70% or more of the data variance. So, in order to get better results, we should add more variables that should help more of the data's variability to our model.

In order to select good variables, we need to spot those that have a strong correlation with the variable that we are trying to predict. We can do that by analysing the plot below

```
pairs(mpg ~ ., data = mtcars)
```



From this we see that cyl, disp, hp, wt have the strongest correlation with mpg. So, we build a new model using these variables and compare them to the initial model with the anova function.

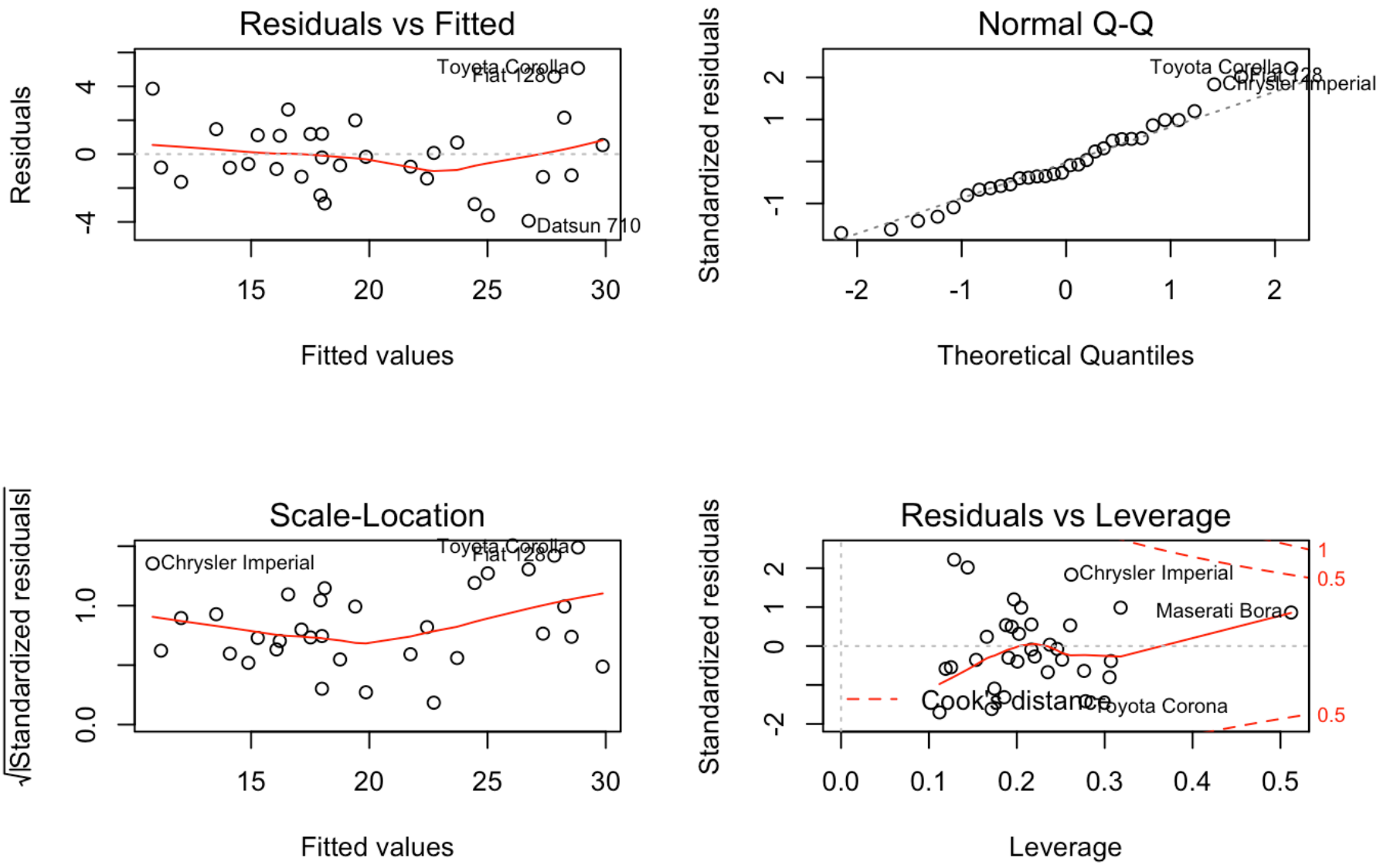
```
multivariable_model <- lm(mpg~am + cyl + disp + hp + wt, data = mtcars)
anova(model, multivariable_model)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	30	720.8966	*NA*	*NA*	*NA*	*NA*
2	25	150.4088	5	570.4878	18.96457	8.636804e-08

2 rows

Analysing the results, we've got a p-value very close to zero, meaning that we can claim the model with more variables is significantly better than our first model. We double-check the residuals for non-normality and can see they are all normally distributed and homoskedastic.

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



At last, here's the summary of the new model:

```
summary(multivariable_model)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## amManual     1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## hp          -0.032480   0.013983  -2.323   0.0286 *
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

The new model explains 86.64% of the variance and as a result, cyl, disp, hp, wt did affect the correlation between mpg and am. Thus, we can say the difference between automatic and manual transmissions is 1.81 MPG (by looking at the estimate value for variable amManual).