Regression Models Course Project

Konstantin Serditov

12/26/2015

Executive summary

This report is devoted to exploring the relationship between set of variables and miles per gallon (MPG) in mtcars dataset. We try to answer two main questions: "Is an automatic or manual transmission better for MPG" and "Quantify the MPG difference between automatic and manual transmissions".

Regression analysis shows that MPG change with respect to transmission type can be estimated between -0.7 to 4.9 (95% confidence interval) and its influence on MPG is by far smaller than influence of other variables. So it is not possible to make useful answer on the question using the dataset.

However, useful linear regression model based on 'hp', 'wt' and 'am' was build and it fits data points with R-squared value 0.8399 and p-value 2.908e-11.

Mtcars data exploration

Using following commands (output is supressed due to verbosity):

```r
data(mtcars)
```

```r
?mtcars
str(mtcars)
summary(mtcars)
head(mtcars)
```

Note that cyl, vs, am, gear, carb are factor variables by nature:

Is there difference in MPG values and Transmission types? See box plot on Figure 1:

```r
boxplot(mtcars$mpg ~ as.factor(mtcars$am))
```

Looks like there is some but we have to look on the whole picture. Please see Figure 2 in the appendix which shows correlation matrix for all presented variables:

```r
corrplot(cor(mtcars), type="lower", method = "number",
         title="Figure 2: Correlation matrix for mtcars")
```

There is also strong correlation with other variables, namely, cyl, disp, hp, drat, wt, vs, am.

Regression analysis

Using Figure 2 exclude gear and qsec from further review. Also, note perfect correlation between cyl and disp and exclude cyl as well. Start with the following model:

```r
summary(lm(mpg ~ disp + hp + drat + wt + vs + am + carb, data = mtcars))$coeff
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 26.539283517 6.78753904  3.9100009 0.0006612587
## disp         0.004628565 0.01544223  0.2997342 0.7669604876
## hp          -0.025950201 0.02052136 -1.2645459 0.2181767405
## drat         1.075662975 1.49989581  0.7171585 0.4801984640
## wt          -2.563931423 1.49112767 -1.7194580 0.0984040922
## vs           1.775150509 1.64358405  1.0800485 0.2908552324
## am           2.527642096 1.67802077  1.5063235 0.1450340746
## carb        -0.402973614 0.67408913 -0.5978046 0.5555697878
```

Eliminate not significant variable using highest p-value (>0.05). On this step it is 'disp'. Repeat the process several times (not shown here due to page limit). After this process, best model which includes 'am' looks like this:

```r
fit <- lm(mpg ~ hp + wt + as.factor(am), data = mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt + as.factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    34.002875   2.642659  12.867 2.82e-13 ***
## hp             -0.037479   0.009605  -3.902 0.000546 ***
## wt             -2.878575   0.904971  -3.181 0.003574 **
## as.factor(am)1  2.083710   1.376420   1.514 0.141268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

Model summary shows slope 2.08 for transmission type coefficient. However, 95% confidence interval for it is

```r
confint(fit, 'as.factor(am)1', level=0.95)
```

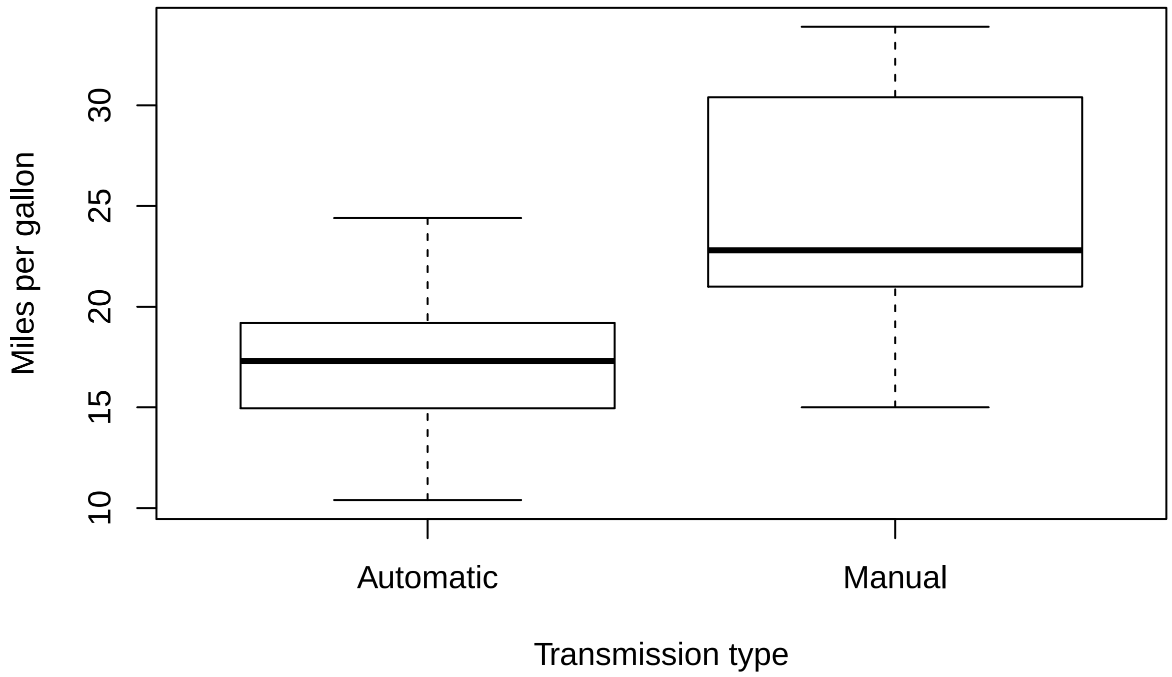
```
##                     2.5 %   97.5 %
## as.factor(am)1 -0.7357587 4.903179
```

So MPG change with respect to transmission type is estimated between -0.7 to 4.9. P-value for 'am' variable is too high and this means that contribution of hp and wt is much more significant than influence of transmission type on MPG.

Diagnostic plots are presented on Figure 3. There is no obvious pattern on Residuals vs Fitted and Scale-Location plots. Points on Q-Q plot mostly lays on line, however two outliers exists as well as on Cook's distance plot, so we may want to investigate and exclude them.

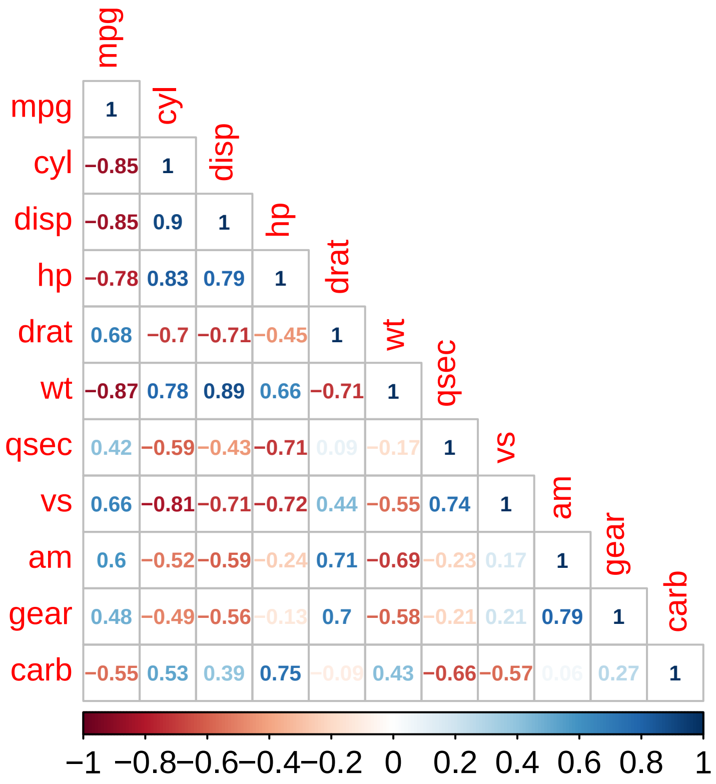
Appendix: supporting figures

Figure 1: Box plot MPG vs Transmission type



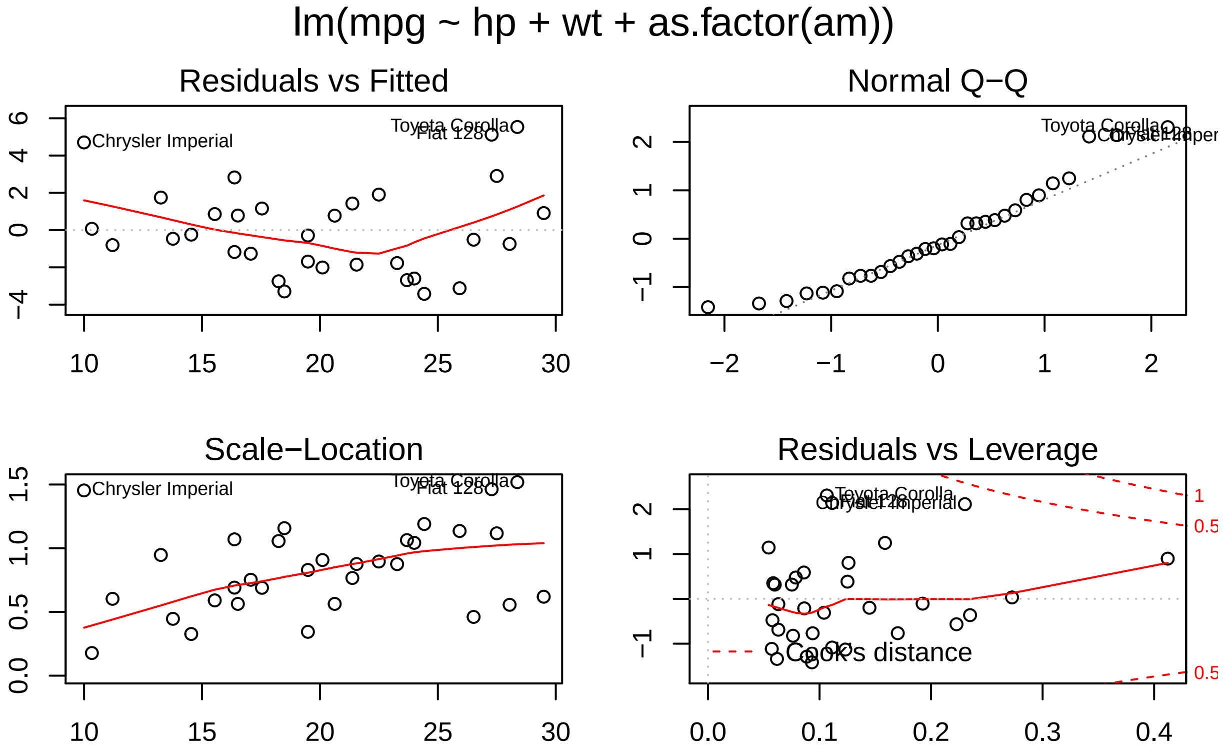
```
## Warning: package 'corrplot' was built under R version 3.2.3
```

Figure 2: Correlation matrix for mtcars



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1										
cyl	−0.85	1									
disp	−0.85	0.9	1								
hp	−0.78	0.83	0.79	1							
drat	0.68	−0.7	−0.71	−0.45	1						
wt	−0.87	0.78	0.89	0.66	−0.71	1					
qsec	0.42	−0.59	−0.43	−0.71	0.09	−0.17	1				
vs	0.66	−0.81	−0.71	−0.72	0.44	−0.55	0.74	1			
am	0.6	−0.52	−0.59	−0.24	0.71	−0.69	−0.23	0.17	1		
gear	0.48	−0.49	−0.56	−0.13	0.7	−0.58	−0.21	0.21	0.79	1	
carb	−0.55	0.53	0.39	0.75	−0.09	0.43	−0.66	−0.57	0.06	0.27	1

Figure 3. Model diagnostics



lm(mpg ~ hp + wt + as.factor(am))
Residuals vs Fitted
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Toyota Corolla
Chrysler Imper
Scale−Location
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Cook's distance