Regression Models Course Project

Hans Ehrencrona

11 March 2016

Executive summary

This document represents the final project in the Coursera/Johns Hopkins Regression Models course. The overall objective was to analyze the `mtcars` data in the R `datasets` package using regression models and exploratory analyses. Specifically, the following two questions were asked: "Is an automatic or manual transmission better for MPG?", and "Quantify the MPG difference between automatic and manual transmissions."

Manual transmission was associated with a 7.2 miles/gallon improvement compared to automatic (P = 0.001). However, in multivariable regression analysis holding weight, horse power and number of cylinders constant, manual transmission showed a non-significant positive association with MPG (1.8 (+/- 2.9) MPG improvement, P = 0.21). In this model, weight and horse power seem to be the most important predictors.

Exploratory analysis and statistical inference

Due to space constraints, I can not show all code in this report. For those interested I refer to the Rmd file at my GitHub repository. Several external packages were loaded for this analysis. Details on the dataset can be found through `?mtcars`. I transformed `mtcars` into `carsFactor`, changing the following variables from numeric to factor: `cyl, vs, am, gear, carb`.

I constructed a boxplot of MPG by transmission type (Appendix, Fig 1) that shows clearly increased MPG for cars with manual transmission. This relationship can be evaluated by `t.test(mpg ~ am, data = carsFactor)`), which demonstrates a highly significant difference of the means, P = 0.0014. Next, I examined the correlation between MPG and the other variables in `mtcars`:

```
##         wt        cyl       disp         hp       carb       qsec
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840
##       gear         am         vs       drat        mpg
##  0.4802848  0.5998324  0.6640389  0.6811719  1.0000000
```

Several variables show a high correlation with MPG, as can also be visualised in a pairwise plot for `MPG`, variables with cor>0.7 (`cyl, disp, hp, wt`) and colour coded for `am` (Appendix, Fig 2a). This is further emphasized in Fig 2b, a scatter plot of MPG vs weight by transmission. From this exploratory analysis, it seems clear that it is oversimplistic to only study the relation between `mpg` and `am`.

Linear regression and model selection

Table 1: fitBase <- lm(mpg ~ am, data = carsFactor)

	Estimate	Std Err	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
amManual	7.244939	1.764422	4.106127	0.000285

This simple linear regression model shows a high correlation between transmission and MPG. The coefficent $\beta 0$ represents the mean MPG for automatic transmission (`am` = 0), and $\beta 1$ is the mean difference between automatic and manual. However, the adjusted R^2 is only 0.33, meaning that a large fraction of the variance remains unexplained by this model. Compare this to the full model:

```r
fitAll <- lm(mpg ~ ., data = carsFactor)
```

In the full model (results suppressed), the adjusted R^2 is 0.78, so a larger fraction of the variance is explained. However, no variables are significant, and inclusion of too many varibles can lead to variance inflation and over-fitting. Going back to the correlation above, the variables with cor>0.7 (`cyl, disp, hp, wt`) all make sense to include in a model, since the weight and engine efficency should have an impact on MPG. Different models were compared:

```r
fitAddWt <- update(fitBase, .~. + wt)
fitAddWtCyl <- update(fitAddWt, .~. + cyl)
fitAddWtCylDisp <- update(fitAddWtCyl, .~. + disp)
fitAddWtCylDispHp <- update(fitAddWtCylDisp, .~. + hp)
```

```r
anova(fitBase, fitAddWt, fitAddWtCyl, fitAddWtCylDisp, fitAddWtCylDispHp)
```

In the ANOVA (results suppressed) we can see statistically significant differences from the base model until we add `disp`, but `hp` again seems to contribute. This leads us to the final model:

```r
fitFinal <- update(fitAddWtCylDisp, .~. -disp + hp)
```

```r
anova(fitBase, fitFinal)
anova(fitFinal, fitAll)
```

These ANOVA results (results suppressed) demonstrate both a significant difference from the base model, and at the same time no significant difference from the full model. Good! Let's take a look at the final model summary together with 95% CI estimates:

Table 2: fitFinal <- lm(mpg ~ am + wt + cyl + hp, data = carsFactor)

	Estimate	Std Err	t value	Pr(>|t|)	95% CI (lower)	95% CI (upper)
(Intercept)	33.7083239	2.6048862	12.940421	0.0000000	28.3539037	39.0627441
amManual	1.8092114	1.3963045	1.295714	0.2064597	-1.0609336	4.6793564
wt	-2.4968294	0.8855878	-2.819404	0.0090814	-4.3171812	-0.6764776
cyl6	-3.0313445	1.4072835	-2.154040	0.0406827	-5.9240572	-0.1386318
cyl8	-2.1636753	2.2842517	-0.947214	0.3522509	-6.8590220	2.5316713
hp	-0.0321094	0.0136926	-2.345025	0.0269346	-0.0602549	-0.0039639

In the final model, it is clear from the $\beta 1$ coefficient and the 95% CI that manual transmission is associated with a 1.8 (+/- 2.9) MPG increase, holding weight, horse power and number of cylinders constant. This result does not reach statistical significance at the $\alpha = 0.05$ level, as demonstrated both by the estimated p-value and the 95% CI. In this model, the other predictors seem to be more important. Finally, the adjusted R^2 of the final model is 0.84 (with multiple $R^2 = 0.87$). This represents a clear improvement in the fraction of the variance explained as compared to the base model, and also an improvement of the adjusted R^2 compared to the full model.

Diagnostics and conclusion

For diagnostic plots, see Appendix, Fig 3. These results are in summary acceptable. There is a slight curve in the residuals vs fitted and scale-location plots, but no important bias or heteroscedasticity can be seen. The QQ plot also shows a weak s-curve, but not enough to violate the distributional assumption. There are no outliers that exert a high leverage. The final model does not show serious variance inflation:

```
vif(fitFinal)
```

```
##         GVIF Df GVIF^(1/(2*Df))
## am  2.590777  1        1.609589
## wt  4.007113  1        2.001778
## cyl 5.824545  2        1.553515
## hp  4.703625  1        2.168784
```

In conclusion, multivariable regression demonstrated that the positive effect of manual transmission on MPG seen in simple linear regression was no longer significant when controlled for important confounders. It might be added that the exploratory plots in Fig 2a and b suggest that important interactions between several predictors may exist, which could be investigated further (outside the scope of my report). Finally, it should be noted that the `mtcars` dataset was collected in 1974, and therefore these results are probably not relevant for modern cars.

APPENDIX

Fig 1. Boxplot of MPG by transmission type

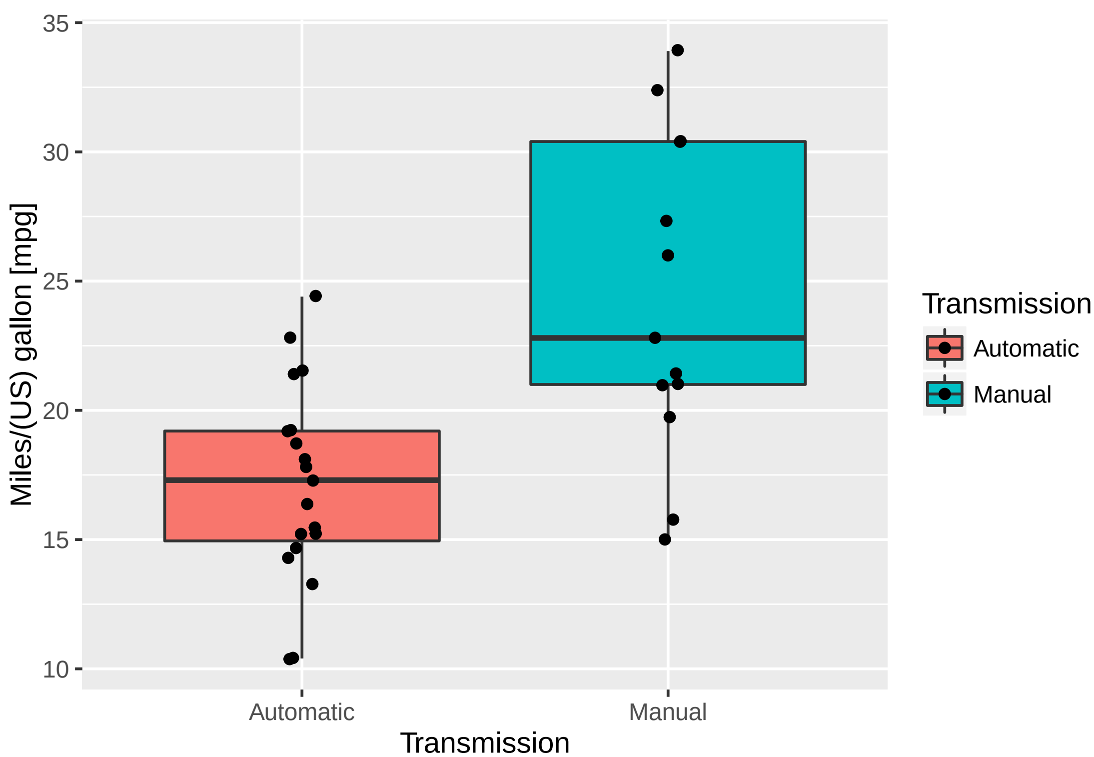

Fig 2a. Pairwise plot of selected variables from the `mtcars` dataset, colour coded by transmission

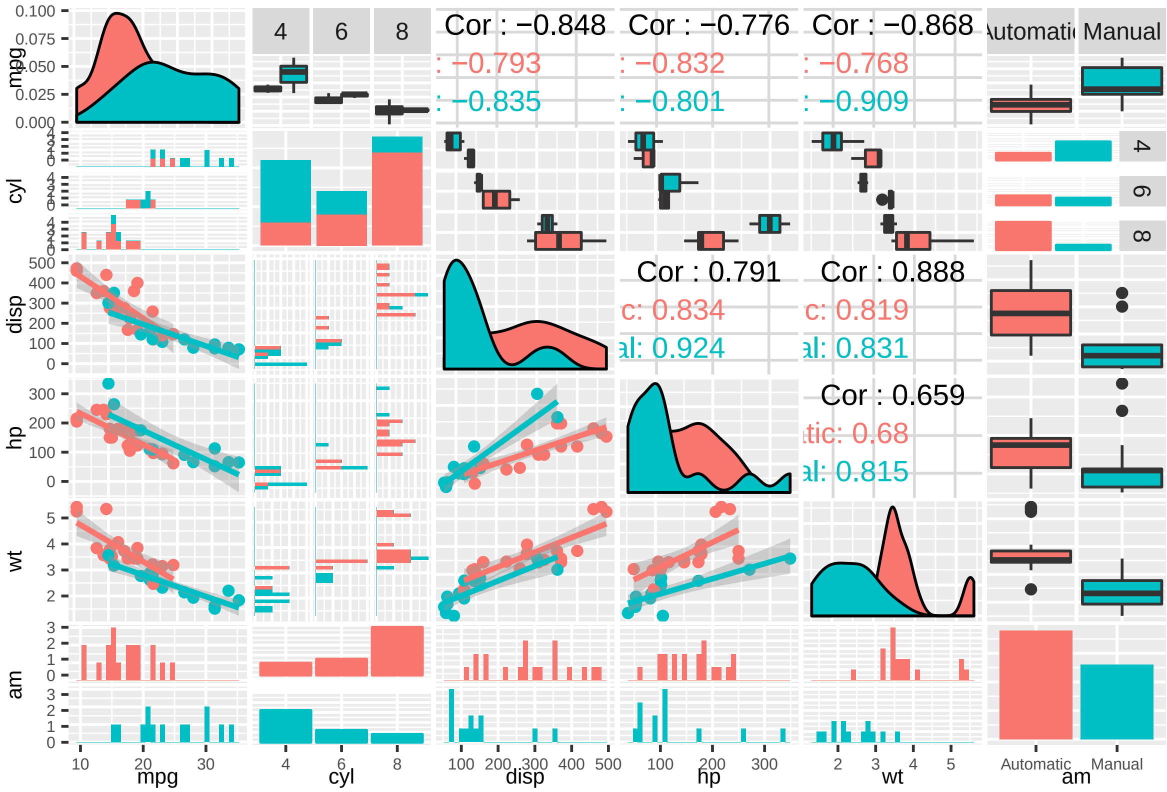

Fig 2b. Scatter plot of MPG vs weight by transmission

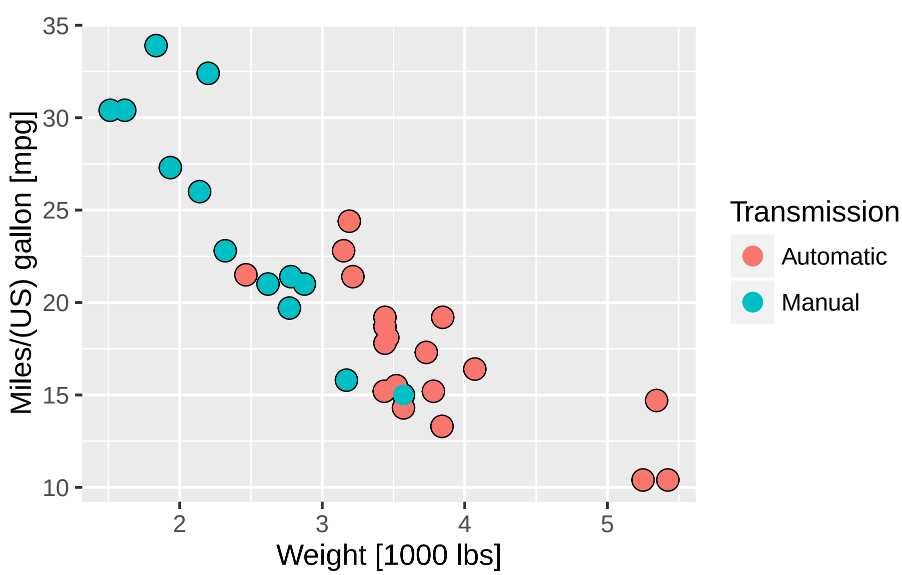

Fig 3. Diagnostic plots