Regression Models - Course Project

Pete

May 3, 2018

Executive Summary

Executives at Motor Trend, a magazine about the automobile industry, are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome), in a data set of a collection of cars. 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

Exploratory data analysis

A brief exploratory data analysis shows that:

1. Automatic transmissions are more frequent than manual, at the time when the dataset was built (see Figure 1 in Appendix)
2. It appears that heavier cars (`wt` variable) are more likely to have automatic transmission (and allow lesser `mpg` – mile per gallon, although this is somewhat obvious); This is a very important finding, because one might believe that manual trasmission gives better `mpg` but according to the data this is only due to the fact that manual trasmission are used on lighter cars! See Fig. 2 in Appendix for details. On the other hand, cars with manual transmission seems to have better `mpg` even with equal horse-power (`hp` variable). See again Fig. 2 in Appendix for details.

Regression analysis

First of all, based on the exploratory analysis, we have performed a t-test against the null hypothesis that there is no correlation between trasmission type and `mpg`. We have obtained a p-value equal to `0.001374`, meaning that we could reject the null hypothesis.

```
##
##  Welch Two Sample t-test
##
## data:  DT[auto == "automatic"]$mpg and DT[auto == "manual"]$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
```

Therefore we have fitted a first linear model, with outcome `mpg` and only predictor the transmission type. Even though this model says that manual transmissions brings a 7.2x improvement on `mpg`, the R^2 value says that this model can only explain 36% of the variance.

```
##
## Call:
## lm(formula = mpg ~ auto, data = DT)
##
## 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 ***
## automanual     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
```

Therefore we fitted a second model, using a multivariate linear regression. The outcome is obviously again `mpg` and the predictors are `auto`, `cyl`, `disp`, `hp` and `wt`. First, we looked at the ANOVA test between the previous model and the second one: the resulting p-value is about `4e-8`, suggesting the second models is much better. Moreover, this models has R^2 such that it can explain 85.5% of the variance.

```
##
## Call:
## lm(formula = mpg ~ auto + cyl + disp + hp + wt, data = DT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## automanual   1.55649    1.44054   1.080  0.28984
## cyl         -1.10638    0.67636  -1.636  0.11393
## disp         0.01226    0.01171   1.047  0.30472
## hp          -0.02796    0.01392  -2.008  0.05510 .
## wt          -3.30262    1.13364  -2.913  0.00726 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

We also looked at the residual plots for this model, verifying they are normally distributed. See Fig. 3.

Conclusions

We conclude that (1) manual transmission seems better in terms of `mpg`, and (2) the difference between automatic and manual transmission is about 1.55x `mpg`. It is worth noticing that this result holds with the 85.5% variance uncertainty given by the R^2 value.

Appendix

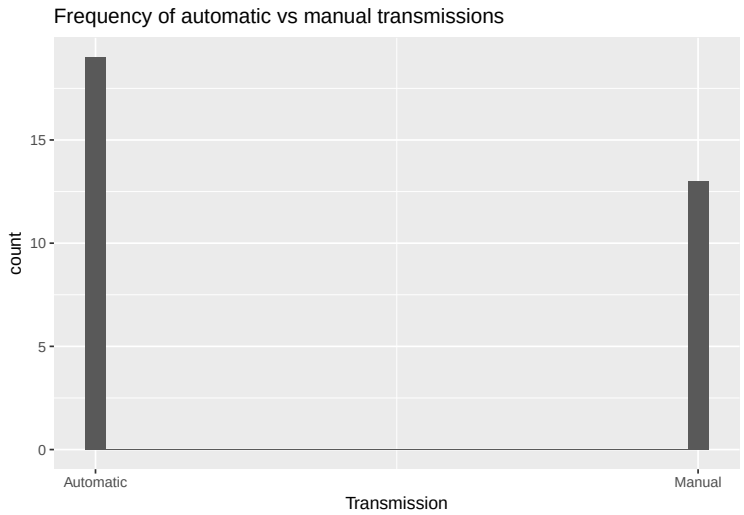

Figure 1:

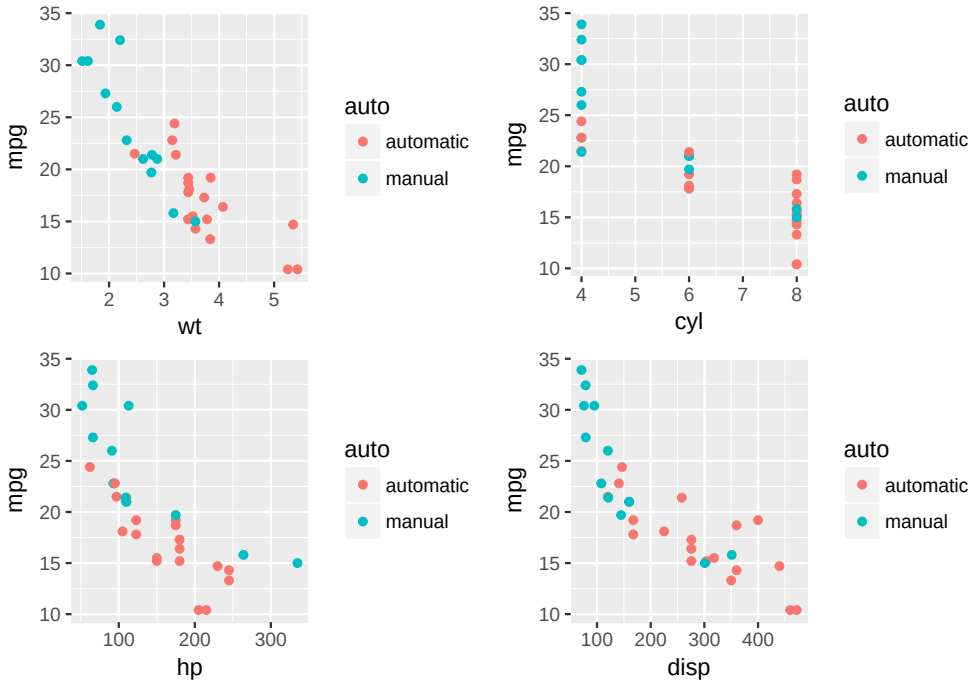

Figure 2:

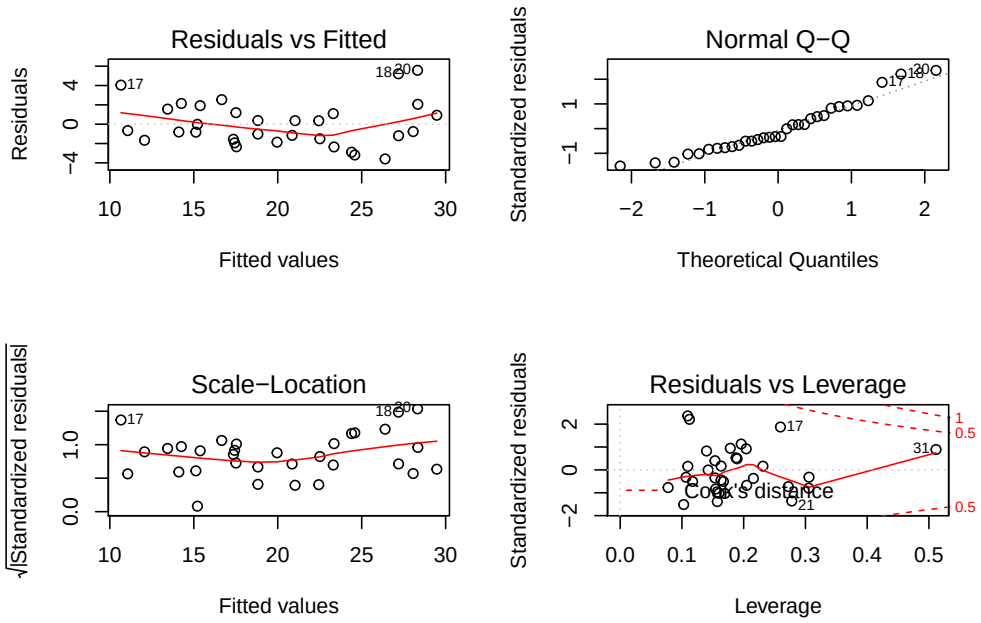
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Figure 3: