Regression Models: effect of transmission type on mileage per gallon in the mtcars dataset

Stefano Merlo

04/11/2015

Overview

This paper is to explain relationship between the transmission type of a car (manual vs automatic) and the mileage per gallon, using R and the dataset `mtcars`.

Findings are that the manual transmission is better for the mileage per gallon outcome, and the difference between the two transmission is 7.2 miles per gallon.

Exploratory data analysis

In the sample analyzed, `59.4%` of the cars have automatic transmission, while `40.6%` have manual transmission.

In average, cars with manual transmission perform `24.39` miles per gallon, while automatic transmission ones perform `17.15` miles per gallon.

Regression analysis

Simple linear regression model

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

The estimated slope of the regression line is positive (`7.2449393`): since the transmission type is a factor predictor, and the 1 value is associated with the manual transmission, we conclude that having a positive coefficient results in an **increment of the estimated mpg when we're considering manual transmission** (intercept + slope X 1); on the other case, when considering the automatic transmission, the slope is multiplied by 0, so the final result is smaller (intercept + slope X 0).

We also estimated that automatic transmission is performing `17.147` miles per gallon, while the manual transmission has '`7.245` miles per gallon more.

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.1474     1.1246 15.2475    0e+00
## am            7.2449     1.7644  4.1061    3e-04
```

From the low p-values we reject the null hypothesis that there is no relationship between the two variables.

Validation

```
##                2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## am           3.64151 10.84837
```

From the calculation, there's 95% of possibility that actual values are included in the interval of our model.

To have a further look at the accuracy, let's plot the residuals (see Appendix).

We can notice that residuals are distributed along the 0 horizontal line. The absence of any pattern give us confidence that the simple linear regression model fits well the predictor / response analysis. But the absolute distance from the X-axis is high, so there may be rooms for improvement.

Now let's quantify the variance explained from the model with the R squared.

```
## [1] "R squared:  0.36"
```

only `36%` of the variance is explained. Let's explore a different method, the multivariate regression analysis.

Multivariate regression model

Looking at only a predictor can be reductive; we can fit a different model by including more variables in it using the formula notation of the lm function.

In Appendix there's the correlation heatmap: we can use to evaluate which predictors are not highly cross-correlated (yellow) and include in the model those that are more independent (red).

First fit is done by including predictors without reciprocal interactions:

```
fit2 <- lm(formula = mpg ~ am + wt + hp + disp + carb, data = mtcars)
```

```
## [1] "R squared:  0.845604004231635"
```

This R squared tells us that `84.6%` of the variance is explained in our model.

But let's try to include also relationships between predictors by slightly change the formula notation (replace + with * between predictors).

```
fit3 <- lm(formula = mpg ~ am * wt * hp * disp * carb, data = mtcars)
```

```
## [1] "R squared:  0.99488111320941"
```

```
## [1] "Is 'am' slope coefficient positive?  TRUE"
```

Wow. `99.5%` of the variance is explained in the new model.

We can consider this new model far more accurate than the simple linear one to predict the outcome from these predictors (see Appendix for the new residuals plot), but it's also more complicated and beyond the scope of this analysis.

What really matter is that the coefficient for the predictor 'am' is still positive: this validates the first result, that the manual transmission increases the mpg.

Conclusion

A simple linear regression model shows that the manual transmission os better for the mileage per gallon, and the increase is quantified in `7.245` mpg. But we also noticed that this model is poor performing, since, even if valid from the confidence interval and p-value analysis, the R squared shows that the model explains only the `36%` of the variance.

Adding new predictors, and their relationships among each other, increase drammatically the accuracy to a `99.5%` of the variance explained, but even if this new model will probably be good for a prediction analysis (predict a response by given new observations) it's not adding value to the scope of this analysis to answer the two simple questions at the beginning.

Appendix

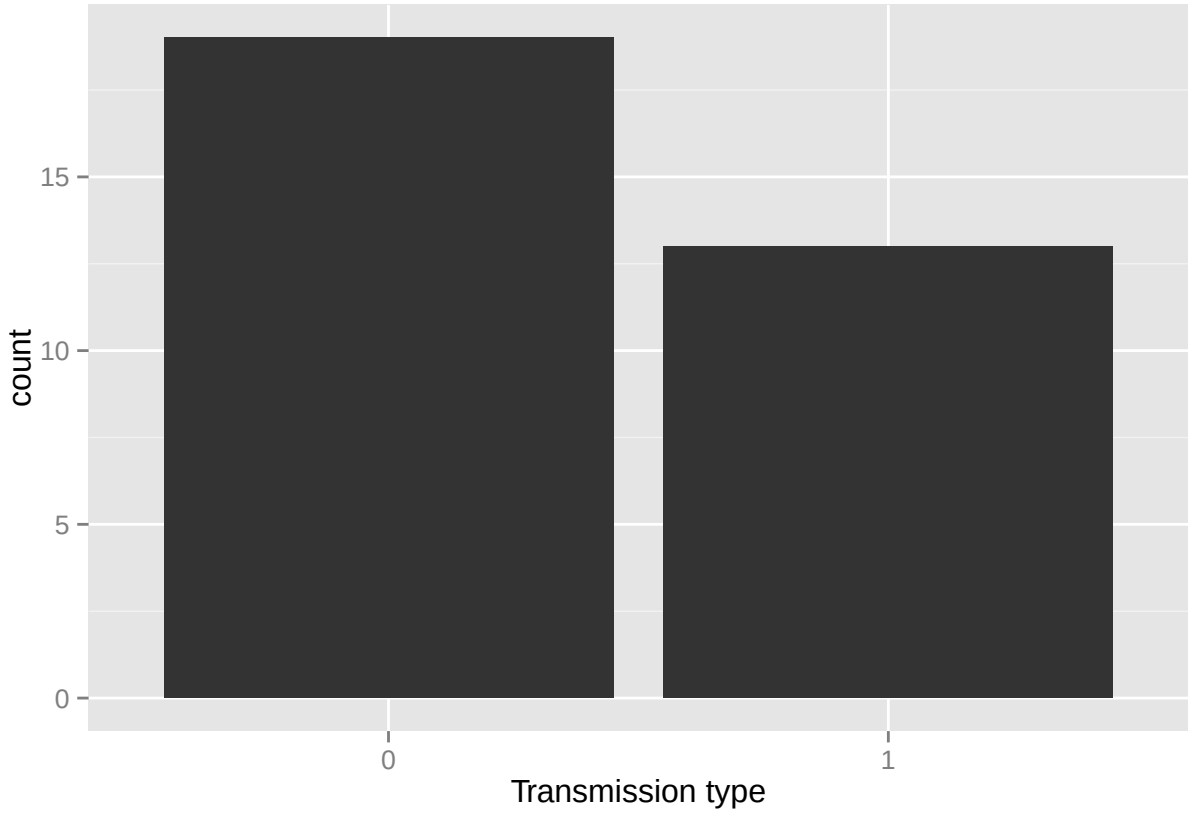
mtcars dataset specs

http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html

Share of manual and automatic transmission in the dataset

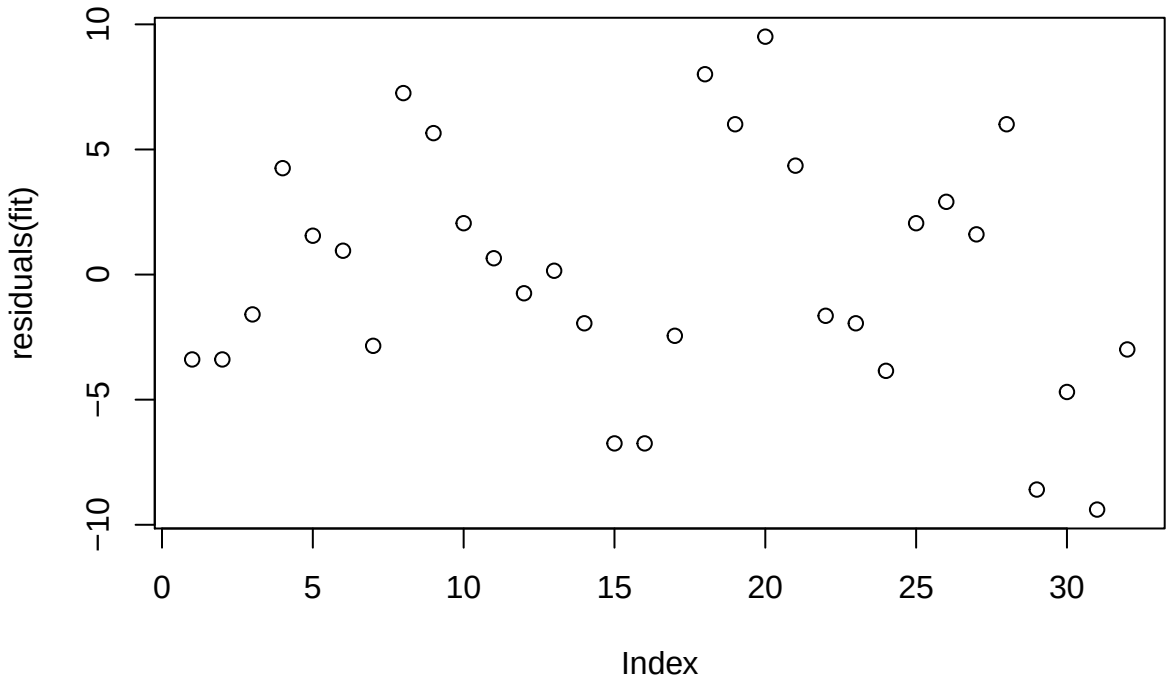
Automatic transmission is labeled with a 0, while manual is 1.

```
qplot(as.factor(mtcars$am), xlab="Transmission type")
```

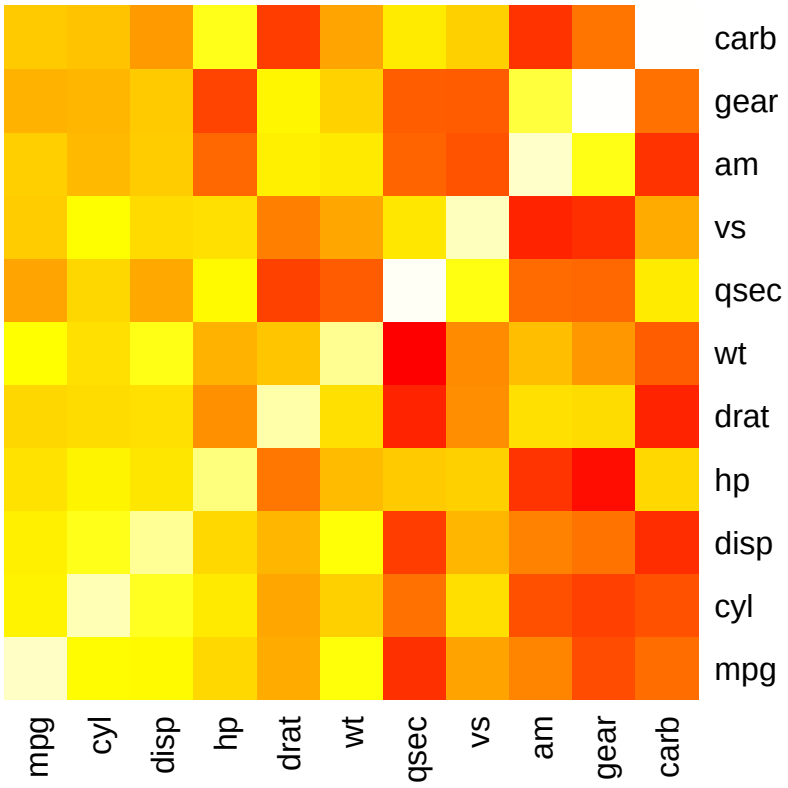


Simple linear regression: residuals plot

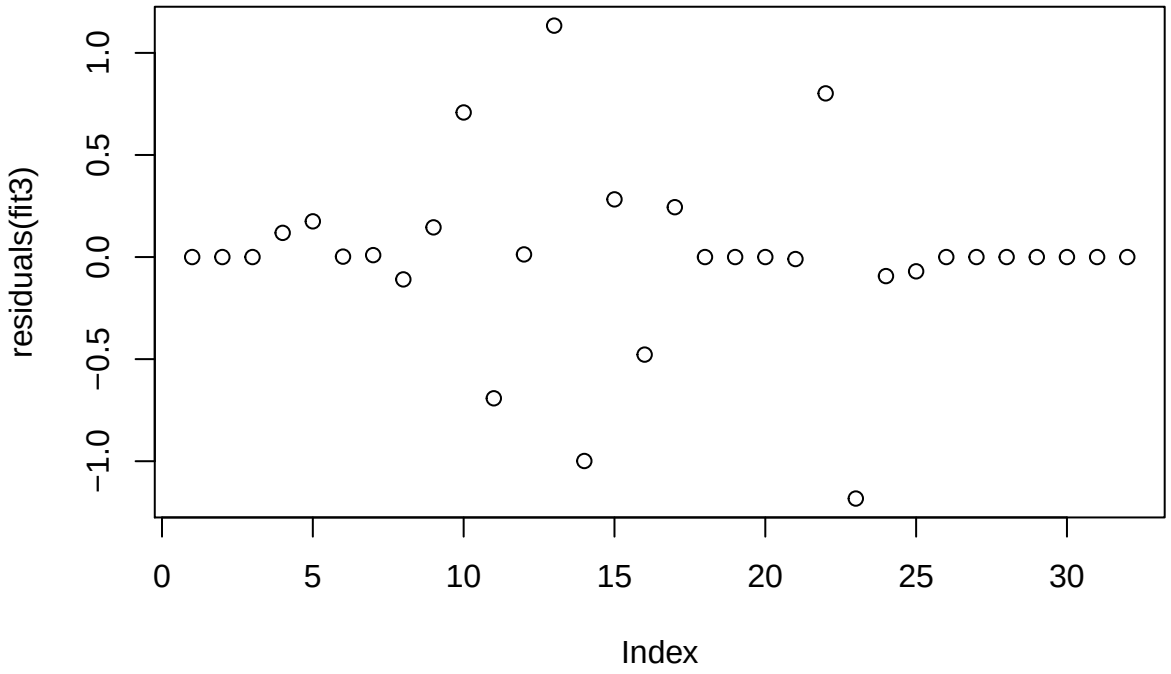
```
plot(residuals(fit))
```



Correlation heatmap



Residuals plot for the enhanced model (multivariate regression)



Complete Knitr markdown document

- https://github.com/erpreciso/regmod-coursera