**Comparing Dependent Variable Model Transformations by Leave-one-out Cross-Validation**

*Written by:* ***David Harris***

[*LinkedIn*](https://www.linkedin.com/in/davidharristunks/) *Profile*

*17 April 2020*

[*https://www.linkedin.com/pulse/comparing-dependent-variable-model-transformations-r-david-harris*](https://www.linkedin.com/pulse/comparing-dependent-variable-model-transformations-r-david-harris)

When deciding on a dependent variable transformation to enable optimal predictive power, a key concern is that the R-squared values (which are commonly used to measure the predictive fit of different models) are no longer comparable. Since R-squared is the proportion of the variance in the dependent variable that is explained by the variation in the independent variables, by performing transformations on the dependent variable we have also transformed their variance structures. As such, we can no longer gain insight into the relative predictive power of the different models by comparing their R-squared values.

**Box-Cox Power Transformations**

Although one may use the Box-Cox power transformation to get an indication of which transformation will help our model fit the data, this method will not help us with a measure of how well the transformed model performs for prediction, nor will it indicate whether using the mean or median predictor when rewinding the transformation will yield the best predictions. The following example highlights the issue relying too heavily on the Box-Cox transformation alone by employing leave-one-out cross-validation to calculate a comparable predictive power measure.

**Leave-one-out Cross-validation**

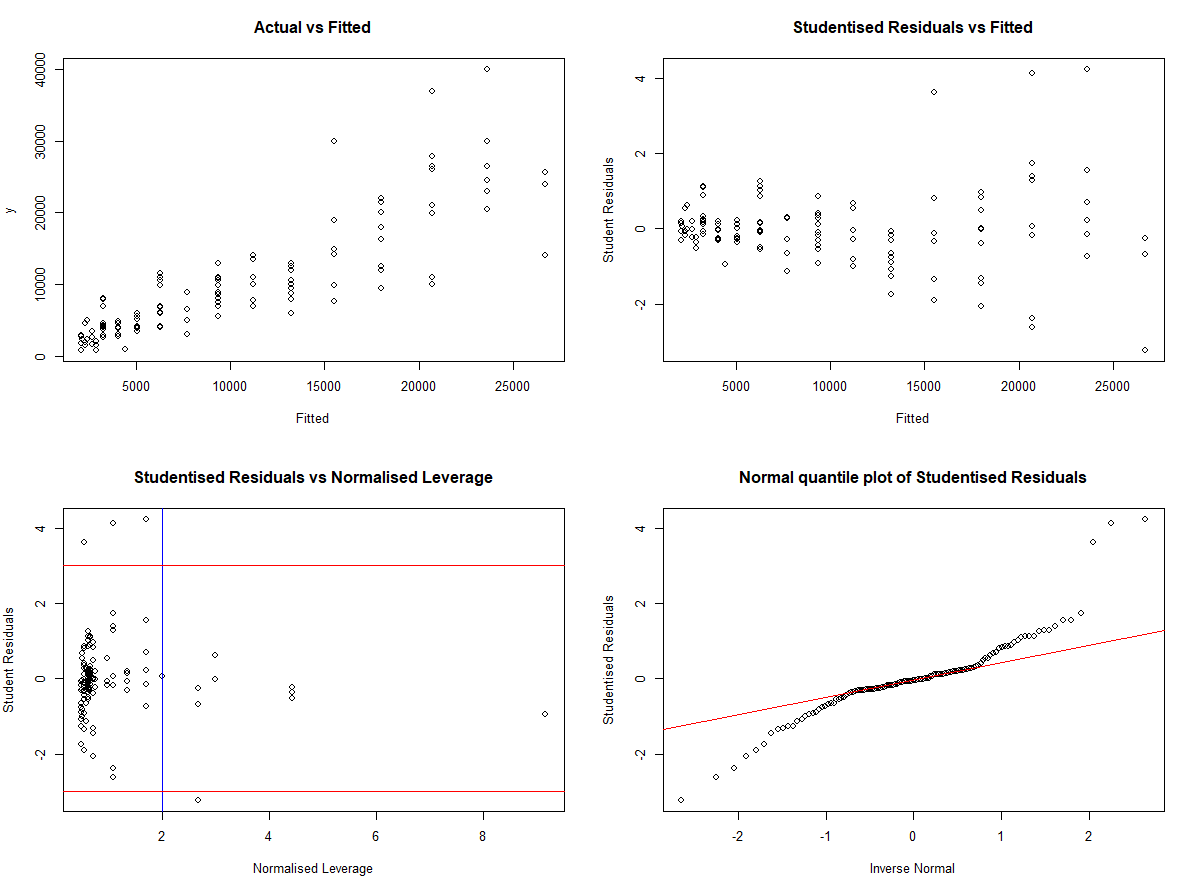
Cross-validation can be used as a means to enable comparison of the different dependent variable transformations. The leave-one-out method of cross-validation uses one observation from the sample dataset to be used as the validation data, using the remaining observations as training data. Each model is used to predict the value of the validation data point by fitting the training data, then a prediction error is calculated by subtracting the predicted value from the validation data observation.

To ensure we obtain comparable measures, it is essential that the fitted value produced by the transformed models are converted back to the “level” form to obtain a predicted value that is in the same units as the original observation data. Using different methods for this conversion enables the comparison of the mean and median predictors. This process is repeated, until each observation from the original sample has been used as a validation data point, then by summing the absolute values1 of the prediction errors we obtain a measure of how well the transformed model in question makes out-of-sample predictions.

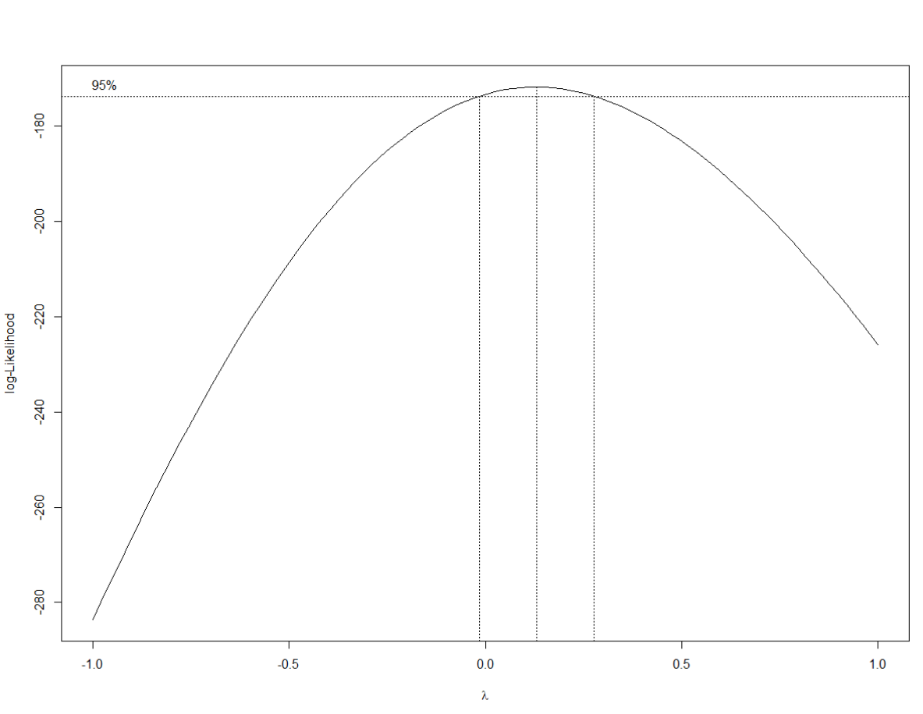
Once the cross-validation measure has been calculated for each transformation (and applicable predictor type), the model that produces the lowest measure is deemed to have the greatest predictive power.

**Example**

With the attached mazda.csv data set we wish to develop a model that will predict the price of a car, given its age. Following initial observations, it is determined that the inclusion of an age-squared term as a second variable in the model will help to alleviate non-linearity concerns. The below diagnostic plots are then produced.δα

Model:

Although the inclusion of this term has helped linearise the diagnostics graphs, we are left with 2 concerns, namely the increasing variance for larger fitted values (heteroskedasticity) and non-normal residuals. We know that both of these issues can be addressed by transforming the dependent variable, and perform a Box-Cox power transformation to see which type of dependent variable transformation may be suitable.



The Box-Cox Power transformation graph above suggests that transforming our Price variable by a log-transformation would be suitable (since 0 lies within the 95% confidence interval of *lambda*, *λ*). However we know that a square-root transformation can also help alleviate the heteroskedasticity and achieve residuals that more closely follow a normal distribution. Therefore we decide to test both a log-transformation and a square-root transformation of Price (as well as our initial linear model as a control) using leave-one-out cross-validation to determine which model will likely produce the best out-of-sample fit.

Using the accompanying R script we perform the LOO cross-validation for the three models. To ensure comparability of our cross-validation measure, we must convert out predicted values back to their “level” value (that is *ln*(Price) to Price and *sqrt*(Price) to Price) before comparing them to the original observations being used as the validation data. Here we can choose either the mean or median predictor. For the model where no transformation is made, the mean and median predictors will be the same, however for the log- and square-root models they are not. Therefore we can test these also via cross-validation to see which predictor obtains the lowest cross-validation measure by use of the below formulas when converting our fitted values back to the “level” prediction value. The accompanying R script performs each of these conversions within the respective cross-validation loop in calculating the comparable predictive power measures.

Log-transformation:

Square-root transformation:

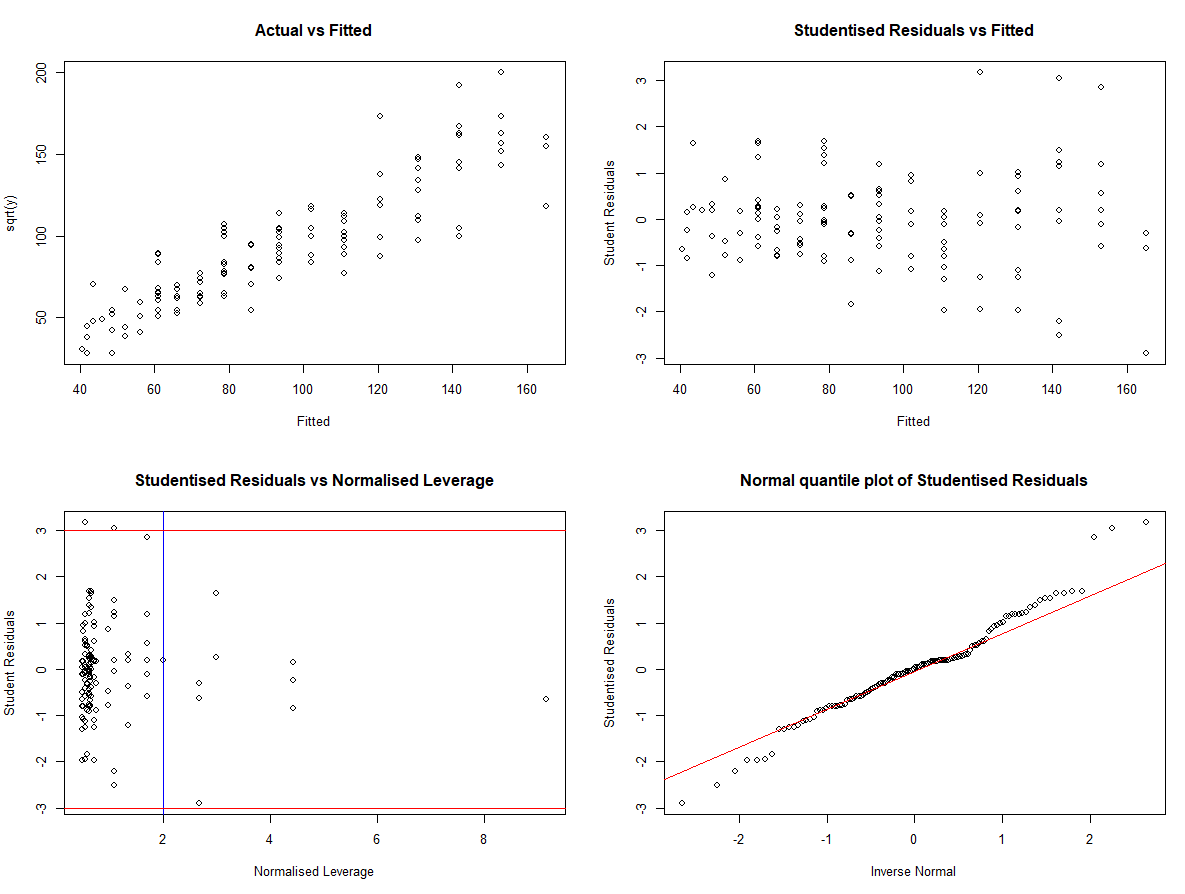
where fitted is our predicted value and MSE is the mean-squared error of the model.

We then can compare the calculated measures to see which model and predictor performs best in terms of cross-validation. From the below table, we see that it is in fact the square-root transformed model that performs best in terms of cross-validation, with the median predictor obtaining the lowest LOOCV measure.

|  |  |
| --- | --- |
| Model | LOOCV Measure |
| Linear (No transformation) | 2812.269 |
| Log-transformation (Median) | 2745.285 |
| Log-transformation (Mean) | 2733.893 |
| **Square-root transformation (Median)** | **2720.082** |
| Square-root transformation (Mean) | 2730.395 |

Taking a final look at the diagnostics plot of the preferred model we can see that the square-root transformation has significantly alleviated our concerns regarding heteroskedasticity and non-normal residuals. As such we can use the square-root transformed model and median predictor to produce predictions and prediction intervals relatively comfortably.

Model:



**Conclusion**

This guide has outlined why comparing R-squared valued is not valid when comparing models with different transformation of the dependent variable. The guide suggests that the use of a Box-Cox power transformation can help identify suitable transformations of the dependent variable, however the Box-Cox transformation alone will not ensure our model performs optimally when making out-of-sample predictions. This guide and attached R script will enable users to perform leave-one-out cross-validation to test and compare the predictive power of models with different transformations of the dependent variable, and indicate whether the mean or median predictor is most appropriate.

**Attachment A** - R Script

**Attachment B** - Mazda.csv data file

*Thanks for reading all the way to the end of the article! I’d love to hear any comments about the above. Feel free to leave a message, or reach out to me through*[*LinkedIn*](https://www.linkedin.com/in/davidharristunks/)*.*