Supplementary Material for “The Box-Cox Transformation: Review and Extensions”

Anthony C. Atkinson
The London School of Economics, London WC2A 2AE, UK,[*]
Marco Riani[†] and Aldo Corbellini,[‡] Dipartimento di Scienze Economiche e Aziendale and Interdepartmental Centre for Robust Statistics, Università di Parma,
43100 Parma, Italy

February 16, 2020

1 Introduction

This supplementary material consists of scatter plots of the simulated data analysed in Section 9 of the paper, followed by three data analyses that were not included in the final version of the paper. Section 5 presents results from applying AREG to the cleaned difference data from §10 of the paper.

The simulated data, after the generation of outliers, are in Figure 1. The right-hand panel of Figure 6 in the paper repeats this scatterplot showing the 164 observations identified as outliers.

2 Gasoline Data

The data in Table 1 of Chen *et al.* (2002) are 107 readings with response the distance driven and explanatory variable the amount of gasoline consumed. Figure 2 shows the fan plot for these data for $\lambda_0 = -0.5, 0, 0.5, 1$ and 1.5. There are two features of this plot: one is that there is a single influential observation (unit 77),

[*] e-mail: a.c.atkinson@lse.ac.uk
[†] e-mail: mriani@unipr.it
[‡] e-mail: aldo.corbellini@unipr.it

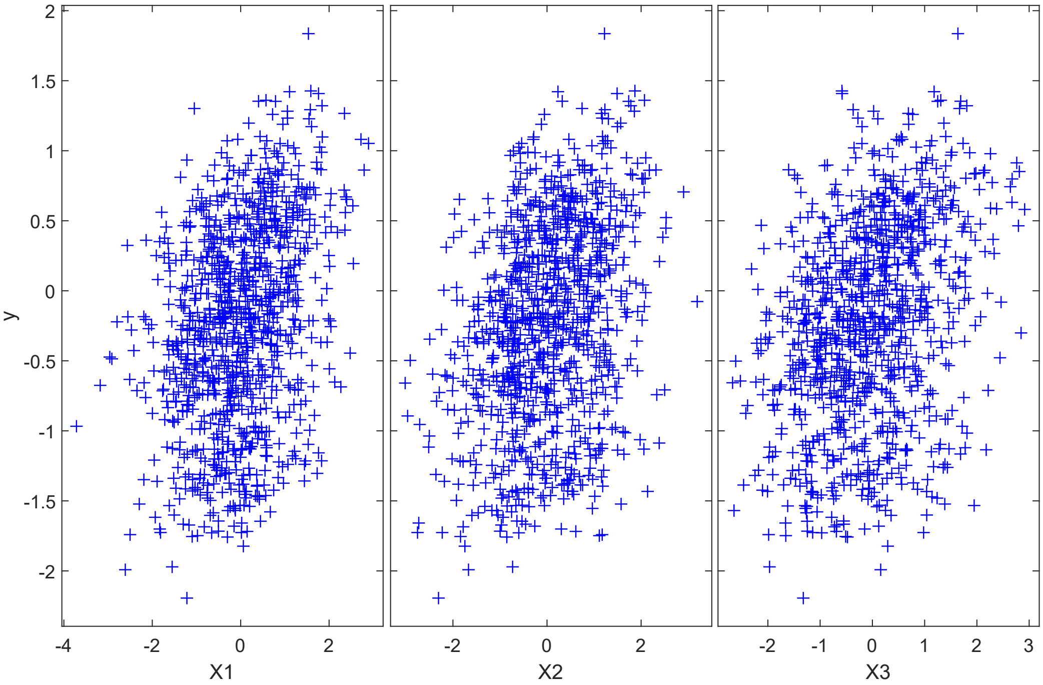

Figure 1: Simulated data of §9: scatterplots of y against $x_1 - x_3$ for the 1,000 simulated observations, including 200 shifted observations

which enters at the end of the search when $\lambda_0 = -0.5, 0$ and 0.5. This observation has by far the lowest values of both x and y. The other is that this observation has a strong effect on the estimated transformation parameter, leading to rejection of values of $-0.5, 0$ and 0.5 for λ; the indication is that the value of $\hat{\lambda}$ will be close to 1.5 as indicated in the top-left panel of Figure 1 of Chen *et al.* (2002). When this observation is deleted, there is no longer any evidence that the data need transformation.

Figure 3 shows, on the log scale, the changes in the estimate of the slope of the regression as λ goes from 0.8 to 2.6. Both plots are close to straight lines. For the unnormalized transformation in the upper panel the values, not logged, go from 3.42 to 145,359, whereas, for the normalized transformation $z(\lambda)$, the change is only from 8.58 to 11.59. Our plots reveal this strong effect, of the type which caused such concern for Bickel and Doksum (1981). The effect of the plots of Chen *et al.* (2002) is muted by the automatic computer rescaling of panels similar to ours.

The lower panel of Figure 3 shows the greatly reduced variability with λ of the parameter estimate from the normalized transformation. As Box and Cox (1982) comment "Of course, the gross correlation effects would be avoided if ... the investigation had been conducted in terms of $z(\lambda)$." They continue "There

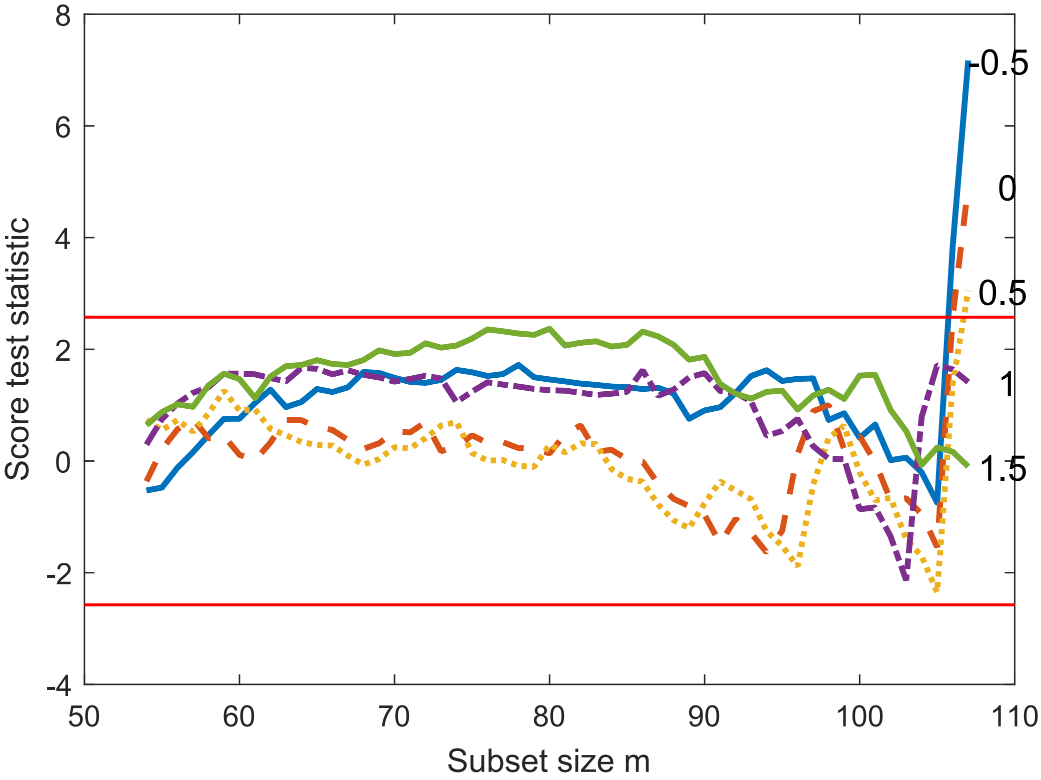

Figure 2: Gasoline data. Fan plot showing that the evidence for the transformations depends on only one observation

are numerous aspects of transformation that merit further study. These include in particular the further development of simple ways of assessing *transformation potential;* that is, of providing some more formal measure of the ability of particular data to provide useful information about a class of transformation". We would claim that Figure 2 provides a very clear indication that the gasoline data have no transformation potential. Although the figure is cogent and easy to interpret, perhaps the simplicity of its calculation could be debated.

The paper of Chen *et al.* (2002) was focused on aspects of theoretical statistics with these data used solely as a motivating example. There was no attempt in either the paper or the discussion to try to elucidate the structure of the data which pops out when appropriate methods are used.

3 Poison Data

The poison data analysed by Box and Cox (1964) provide a clear example of data with strong transformation potential; the analysis produces a paradigmatic fan

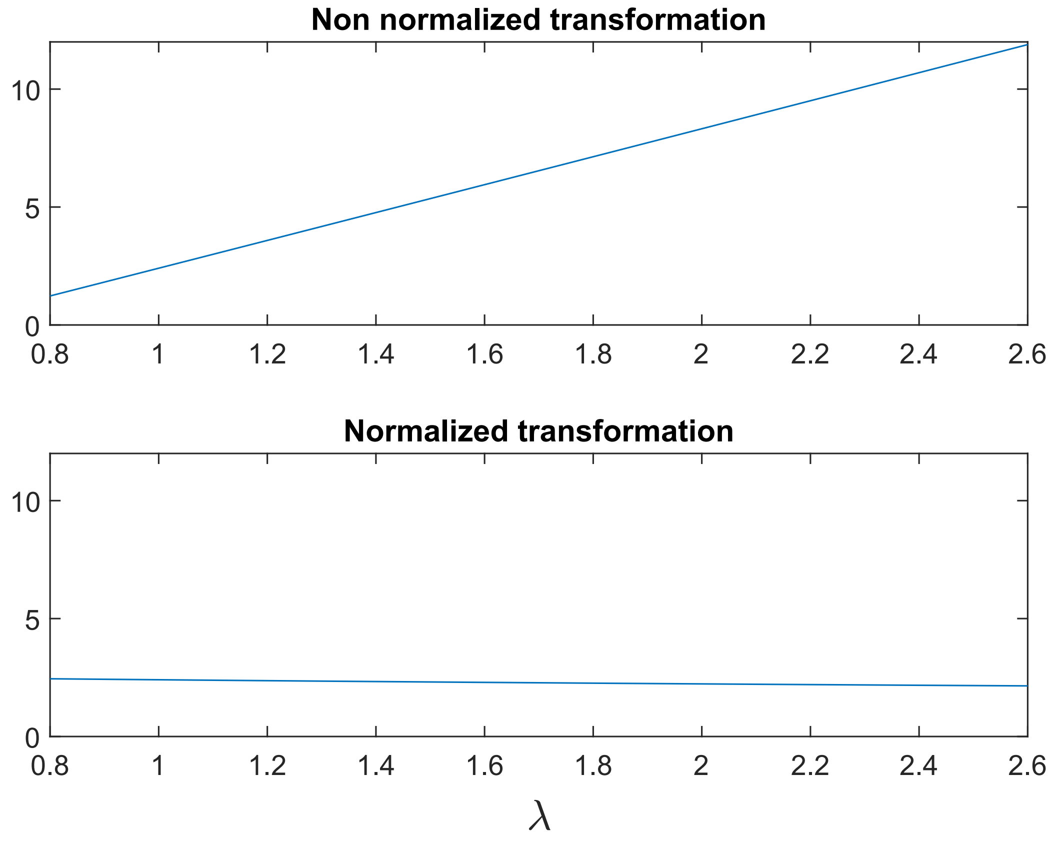

Figure 3: Gasoline data; estimate, on the log scale, of the regression parameter $\beta(\lambda)$ as λ varies. Upper panel response is $y(\lambda)$, lower panel $z(\lambda)$

plot. We compare the Box-Cox transformation with those from ACE and AVAS both for the original data and for data modified to have four outliers.

The data are the survival times of animals in a 3×4 factorial experiment, the factors being three poisons and four treatments. Each combination of the two factors is used for four animals, the allocation to animals being completely randomized. There are thus 48 observations. The data presumably come from Box's work during World War II on antidotes to nerve gases (Box, 2013, p.28). We fit a model without interactions so that $p = 8$. The fan plot in Figure 4 shows trajectories of the score statistic for six values of λ_0 fanning out as the search progresses. There are no abrupt changes such as were caused by the single outlier in Figure 2. Thus all the data support the values of -1 and -0.5 for λ_0, the value of -1 being chosen on grounds of scientific interpretability. The conclusion is that death rate is the property with a simple additive structure, rather than survival time.

We now contaminate the data. In part of a study of the effect and identifica-

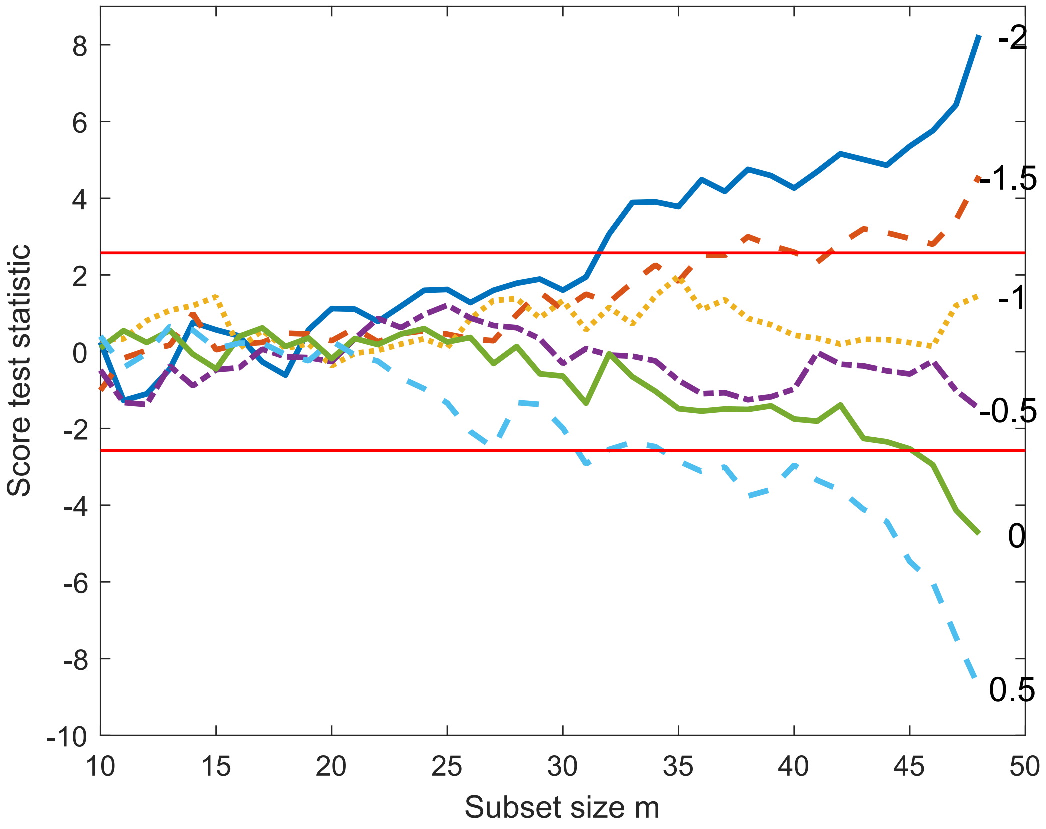

Figure 4: Poison data. Fan plot suggesting $\lambda = -1$ or -0.5

tion of multiple outliers, Atkinson and Riani (2000, p.104) modified four small observations in the poison data, the intention being to produce large outliers on the reciprocal scale, which have little effect on the untransformed data and so influence $\hat{\lambda}$ towards 1. The fan plot for these contaminated data is in Figure 5. The effect is dramatic. For three values of λ_0, the four outliers enter at the end of the search causing the trajectories for $\lambda_0 = -1$ and -0.5 to move outside the 99% bands; earlier in the search the values of the statistics for $\lambda_0 = -1$ lie in the centre of the band. The plot shows that a plausible estimate for λ based on all the data would be 0.25.

We also analysed the two sets of data with ACE and AVAS. In all our comparisons we specified monotonic transformations for ACE; In AVAS the response transformation is always monotonic. We summarize these comparisons in Figure 6 by superimposing plots of the transformations found for the original and contaminated data for the three transformation methods. The left-hand panel is for the Box-Cox transformation, where λ has changed from -1 to 0.25. Compared

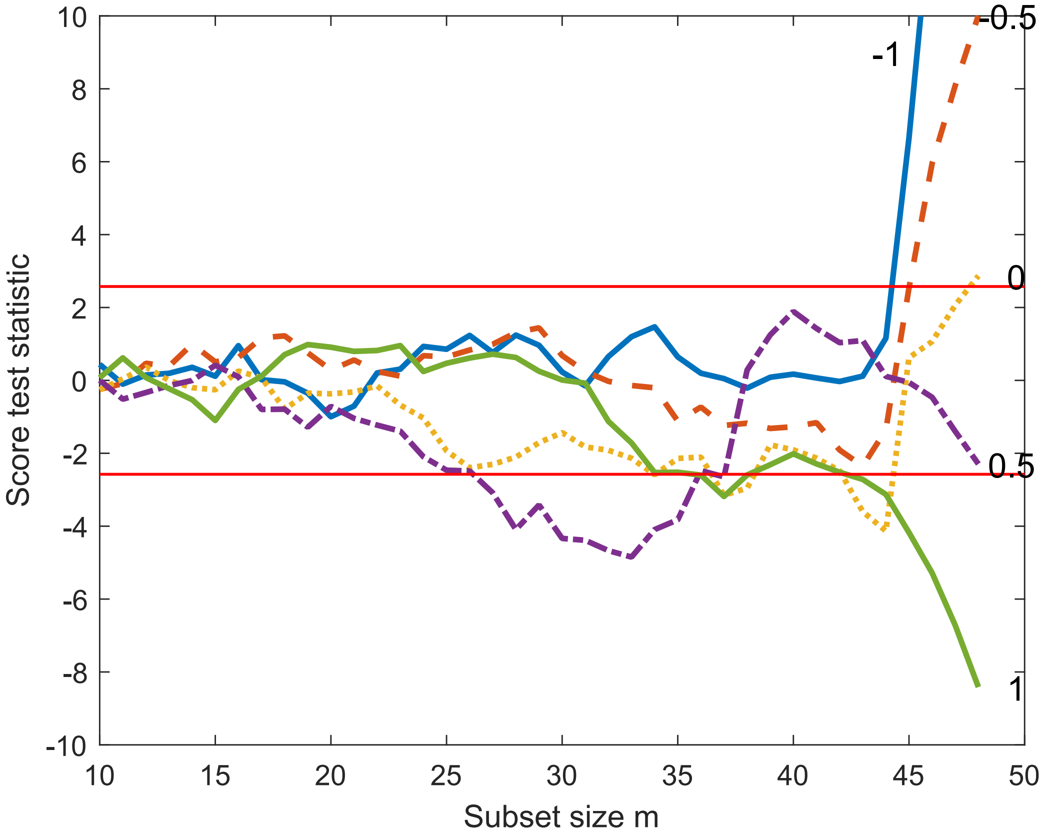

Figure 5: Contaminated poison data. Fanplot.

with the curve for -1 the curve for the contaminated data is almost horizontal; there is only a slight transformation of the data. The other two panels have a different vertical scale. The centre panel shows the two curves for AVAS, with that for the contaminated data again being less curved, but roughly corresponding to a Box-Cox value of $\lambda = 1/3$ (solid line). The curves for ACE in the right-hand panel are quite different. The four outliers have been Winsorized and brought in to the value of the fifth smallest observation. Otherwise, the two curves are virtually identical. However, the smallest observations in the original data have also been Winsorized. It is clear from the figure that the Box-Cox transformation produces a stronger transformation of the original data than do the other two methods. Compared with the Box-Cox transformation, ACE and AVAS seem to give similar transformations whether the data are contaminated or not, with AVAS transforming the data slightly more strongly. The plot of residuals against fitted values for ACE from both sets of data (not shown) do show some increase of variability with fitted value, an indication of under transformation.

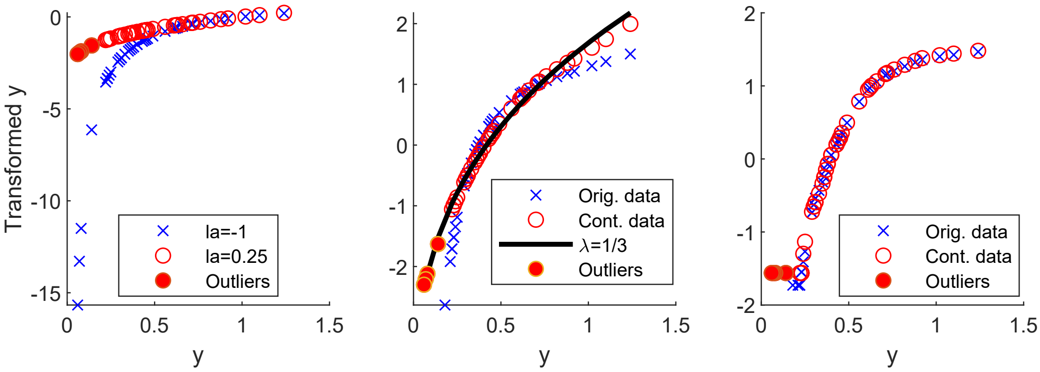

Figure 6: Contaminated poison data; comparison of transformations. In all panels the four contaminated observations are shown by filled circles. Left-hand panel, Box-Cox transformation; Centre-panel, AVAS and Box-Cox transformation with $\lambda = 1/3$. Right-hand panel, ACE. Note the distinct vertical scale for the Box-Cox transformation.

4 Balance Sheet Data

Atkinson *et al.* (2020) analyse 1,405 observations on the profit or loss of individual firms; 407 make a loss. Their analysis uses the extended fan plot and the forward search to detect outliers and obtain estimates of λ_N and λ_P. We briefly summarise their analysis, before turning to the use of AVAS and ACE.

There are five explanatory variables. Regression on all five produces a fitted model for which $R^2 = 0.511$. The fan plot for the overall statistic suggests the hypothesis of no transformation $(\lambda_0 = 1)$ is acceptable, although there is an abrupt increase in the value of the statistic towards the end of the search perhaps indicating the presence of outliers. The extended fan plot for testing $\lambda_0 = 1$ clarifies this structure; positive and negative observations apparently need different transformations with a sharp increase in the values of all three statistics at the end of the search.

The analysis again proceeded by trial and error over a coarse grid of values to find estimates of the two transformations parameters, checking potential transformations with extended fan plots for $\lambda_0 = 1$ for the transformed data. The resulting transformation had $\lambda_P = 0.5$ and $\lambda_N = 1.5$. The forward search was used to identify outlying observations in this scale, a procedure similar to that for the John and Draper difference data of §10. The automatic procedure for outlier detection (Riani *et al.*, 2009) identified a total of 42 outliers. The value of R^2 for regression on the transformed data with the outliers deleted is 0.684, compared with 0.511 for the original data.

We now consider non-parametric response transformations, starting with AVAS.

The top left-hand panel of Figure 7 is a plot of the transformed against original y which shows a sigmoid logistic shape with an inflection point near $y = 0$. However, the diagonal patterns in the plot of residuals versus fitted values show that regression with this transformed response does not remove all the structure in the data. The value of R^2 for this regression is only 0.526; the transformation has achieved little.

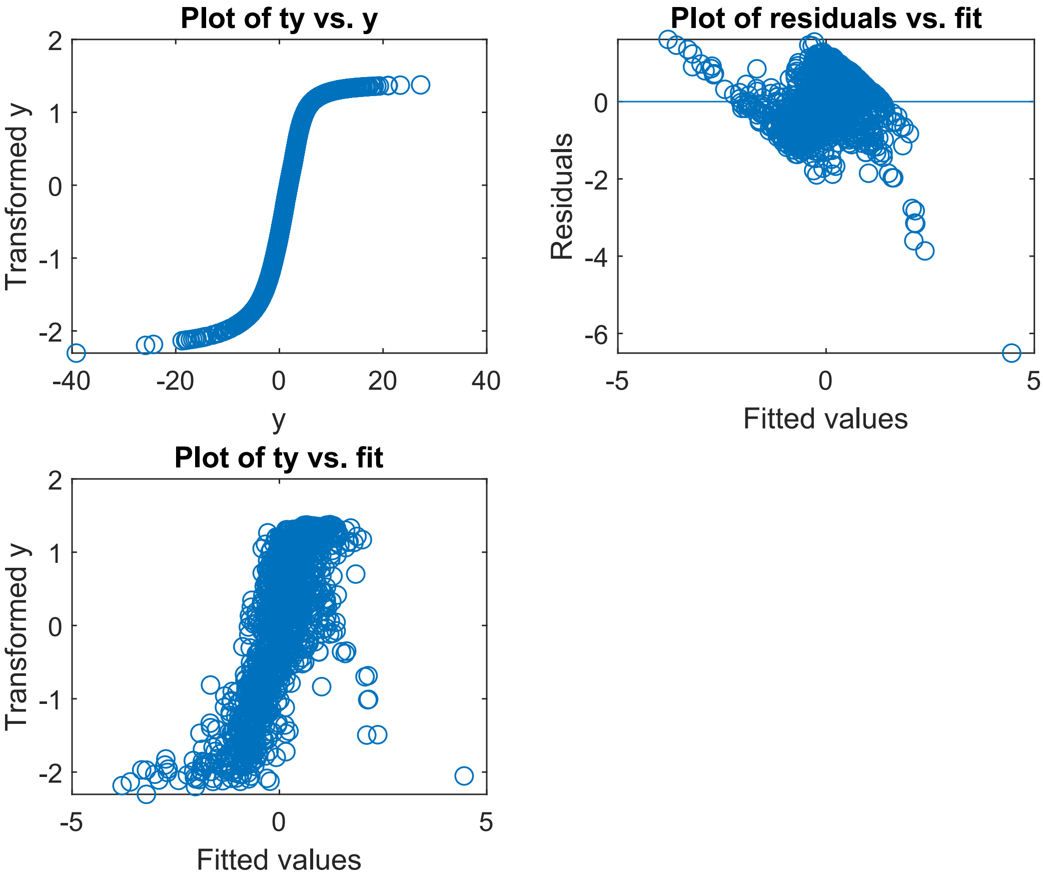

Figure 7: Balance sheet data. Properties of data transformed by AVAS

Figure 8 shows the parallel results for ACE. The plot of transformed versus original y again shows an inflection at y near zero, but now there are two virtually straight parts and a set of constant transformed values for the lowest values of y. These constant values give rise to a diagonal band in the plot of residuals against fitted values. The value of R^2 for regression on this response is 0.558, an improvement on AVAS, but far from the value of 0.684 from parametric transformation and outlier removal.

We conclude with some further comparisons of parametric and non-parametric

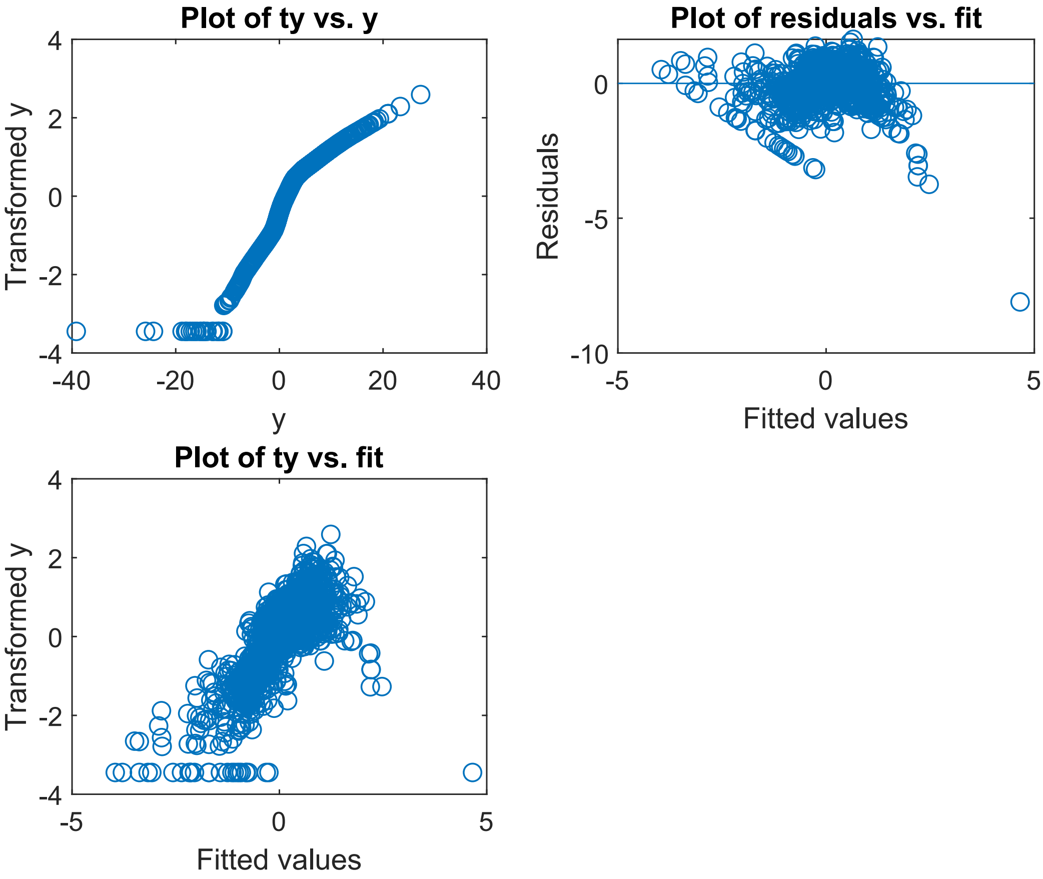

Figure 8: Balance sheet data. Properties of data transformed by ACE

transformations. For comparison with Figures 7 and 8, Figure 9 shows the residuals plotted against fitted values for the data cleaned of outliers, transformed with $\lambda_P = 0.5$ and $\lambda_N = 1.5$. The 42 outliers are shown as filled circles. For the remaining observations the plot shows there is, as there should be, no relationship between residuals and fitted values.

Figure 10 compares the plots of transformed against original y; that for AVAS is in the left-hand panel. The results from the extended Yeo-Johnson transformation with $\lambda_P = 0.5$ and $\lambda_N = 1.5$ are plotted as crosses. They show a concave portion for positive y and a convex portion for negative values. The AVAS values are plotted as circles, with filled circles for the 42 outliers. The AVAS curve provides a poor approximation to that of the parametric transformation. A better approximation is provided by the curve from ACE in the right-hand panel, particularly once the outliers are disregarded. This improved approximation is reflected in the larger value of R^2 for ACE than for AVAS.

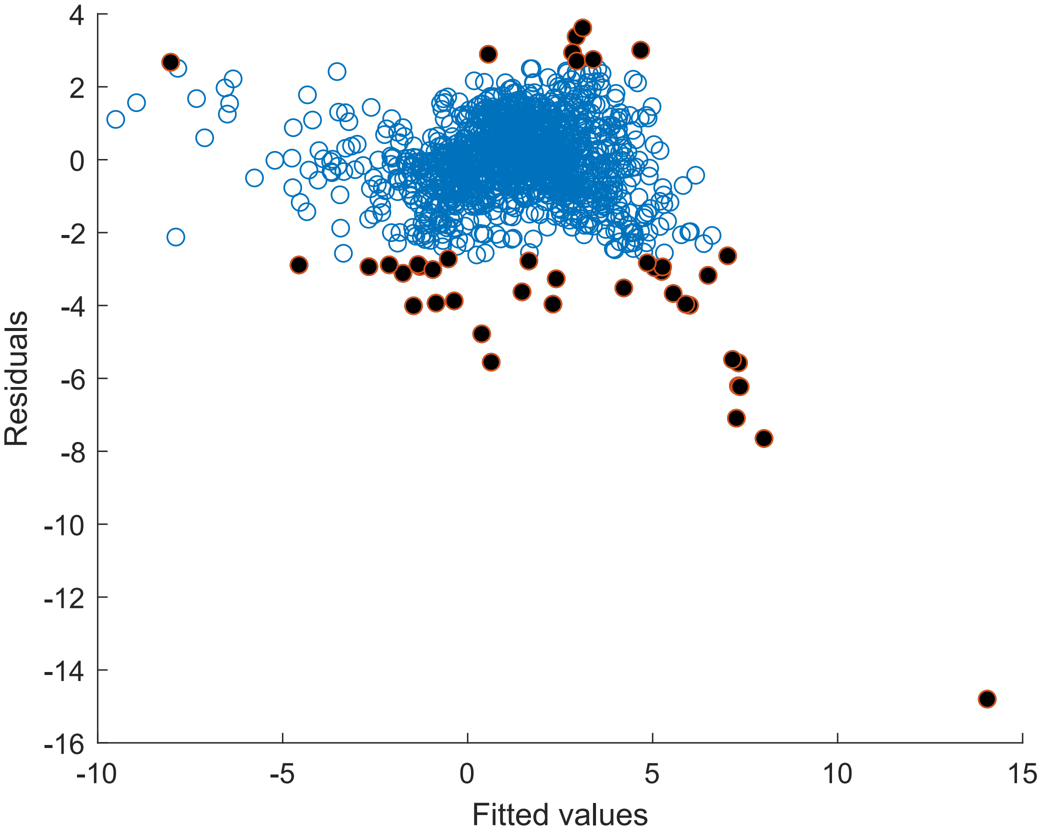

Figure 9: Cleaned balance sheet data; scaled residuals against fitted values. The values for the 42 deleted observations are shown by filled circles

5 John and Draper Difference Data

The analysis in Section 10 of the paper of the difference data due to John and Draper (1980) leads to the identification of six extreme outliers. The values of R^2 from the subsequent analyses of the "cleaned" data, that is after the removal of the outliers, are in Table 2 of the paper, in which the highest values come from the use of ACE. The cleaned data were also analysed using AREG. The plots of transformed against original y for $k = 3 - 6$ in the panels of Figure 11 show that AREG did not yield a monotonic transformation for any of these values for the number of knots in the splines used for smoothing.

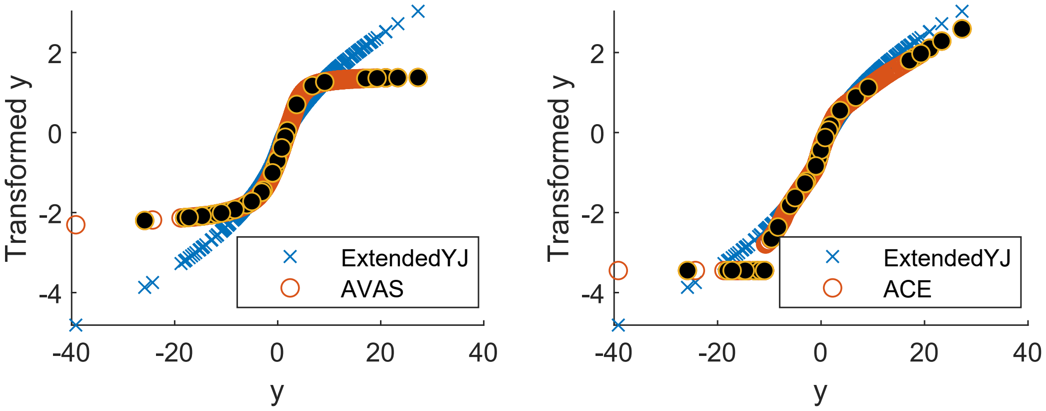

Figure 10: Balance sheet data: comparison of non-parametric transformations with the extended Yeo-Johnson transformation with $\lambda_N = 1.5$ and $\lambda_P = 0.5$. Left-hand panel AVAS, right-hand panel ACE. For the non-parametric transformations, the 42 outliers are shown by filled circles

References

Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer–Verlag, New York.

Atkinson, A. C., Riani, M., and Corbellini, A. (2020). The transformation of profit and loss data. *Applied Statistics*, **69**. DOI: https://doi.org/10.1111/rssc.12389.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, **76**, 296–311.

Box, G. E. P. (2013). *An Accidental Statistician*. Wiley, Chichester.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, **26**, 211–252.

Box, G. E. P. and Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, **77**, 209–210.

Chen, G., Lockhart, R. A., and Stephens, M. A. (2002). Box-Cox transformations in linear models: large sample theory and tests of normality (with discussion). *The Canadian Journal of Statistics*, **30**, 177–234.

John, J. A. and Draper, N. R. (1980). An alternative family of transformations. *Applied Statistics*, **29**, 190–197.

Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, **71**, 447–466.

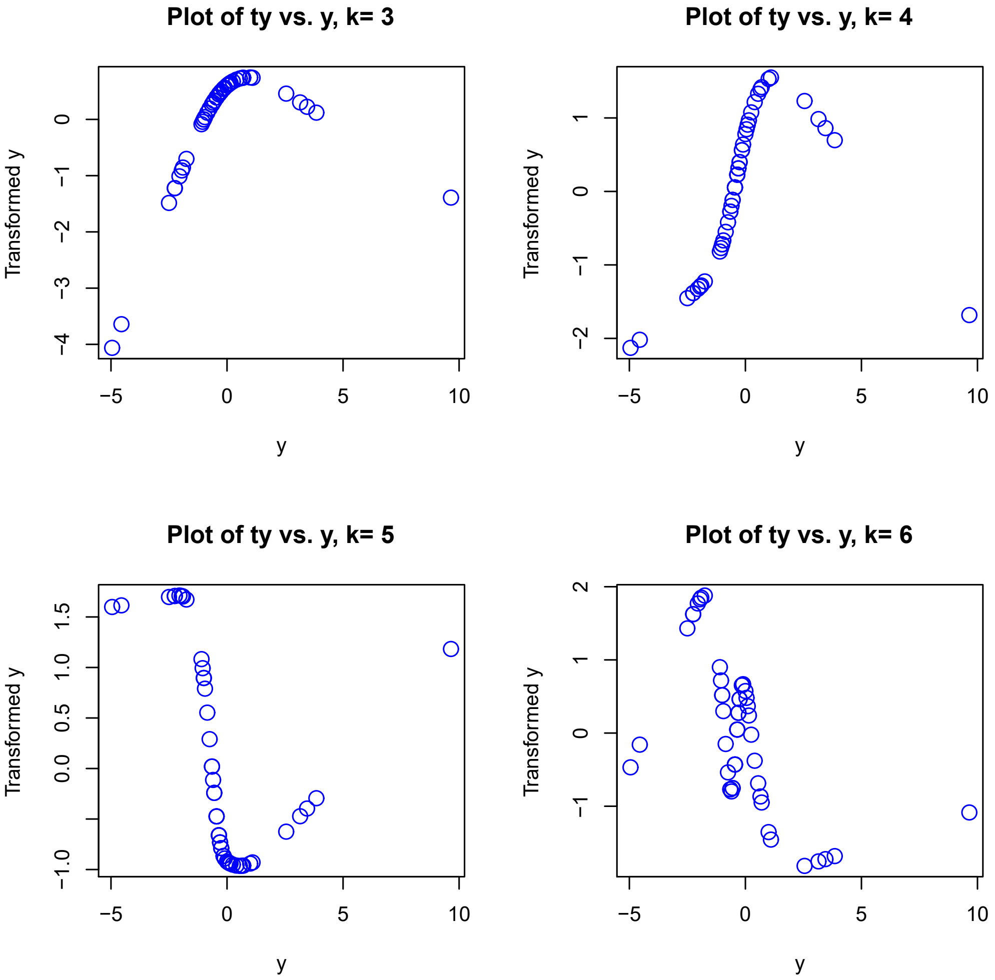

Figure 11: Cleaned difference data: comparison of transformed response against untransformed values from analyses with AREG for four values of the number of knots k. The default value is $k = 4$