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

Abstract

The Box-Cox power transformation family for non-negative responses in linear models has a long and interesting history in both statistical practice and theory, which we summarize. The relationship between generalized linear models and log transformed data is illustrated. Extensions investigated include the transform both sides model and the Yeo-Johnson transformation for observations that can be positive or negative. The paper also describes an extended Yeo-Johnson transformation that allows positive and negative responses to have different power transformations. Analyses of data show this to be necessary. Robustness enters in the fan plot for which the forward search provides an ordering of the data. Plausible transformations are checked with an extended fan plot. These procedures are used to compare parametric power transformations with nonparametric transformations produced by smoothing.

AMS 2000 subject classif cations: Primary 62F03, 62F35, 62J05, 62J20; secondary 68U05.

Keywords:

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

ACE; AVAS; constructed variable; extended Yeo-Johnson transformation; forward search; linked plots; robust methods.

1 Introduction

This paper is principally concerned with the Box-Cox transformation of the response in linear regression models. The extensions include the transformation of both sides of the model, the transformation of responses that can be both positive and negative and comparisons with nonparametric alternatives. It is taken as given that robust procedures are necessary, since outlying observations can have an appreciable effect on the estimated transformation.

The use of transformations in the simplif cation of distributions has a long history. Cox (1977) instances Fisher's z transformation of the correlation coeff - cient (Fisher, 1915). Probit analysis for binomial proportions (Bliss, 1934) is also a transformation to normality. General discussions of the history, purposes and development of transformations are in the review article Cox (1977) and two related articles taken from the Encyclopedia of Statistics (Atkinson and Cox, 1988; Taylor, 2004). Box and Cox (1964) emphasise the effect of transformations to normality on the systematic part of the model. The transformation should provide simple, more revealing analyses that lead to sharper inferences. An extensive survey of literature from the f rst quarter century of the Box-Cox transformation is Sakia (1992). Hoyle (1973) lists 19 transformations, several of which are special cases of the Box-Cox transformation. The monograph of Carroll and Ruppert (1988) ranges widely over topics in the statistical transformation of data.

The Box-Cox transformation is described in §2 together with some of the inferential problems arising from this seemingly simple model. The use of the Box-Cox transformation is illustrated in §3 by the analysis of data on mental illness. The results are compared with those from a generalized linear model, that is a model in which the linear predictor, rather than the response, is transformed. Section 4 covers the transform both sides method of Carroll and Ruppert (1988) which can preserve the relationship between the response and a theoretical model whilst achieving homogeneity of variance. The section also describes nonparametric alternatives to the Box-Cox transformation, as well as other transformations, including extensions of the Box-Cox transformation.

These procedures are based on aggregate statistics, calculated over the whole sample. However, estimation of the transformation parameter can be particularly sensitive to outliers and an incorrect transformation can indicate spurious outliers that disappear under the correct transformation. In §5 we discuss robust methods and, in §6.2, recall the fan plot that illuminates the effect of individual observations on the estimated transformation. Section 7 illustrates the use of these robust

techniques in the analysis of the illness data and compares the results with those from nonparametric transformations.

Yeo and Johnson (2000) extended the Box-Cox transformation to a one-parameter family that allows the transformation of both positive and negative observations. §8.2 describes the further extension of this transformation by Atkinson *et al.* (2020) to allow different transformation parameters for positive and negative observations, together with robust procedures for testing whether the different parameter values are necessary. The paper concludes with an analysis of data on differences (John and Draper, 1980) which illustrates the need for this extended transformation as does the analysis in §4 of the supplementary material.

2 The Box-Cox Transformation

The Box-Cox transformation for non-negative responses is a function of the parameter λ. The transformed response is

$$y(\lambda) = (y^\lambda - 1)/\lambda \qquad (\lambda \neq 0); \qquad \log y \qquad (\lambda = 0), \tag{1}$$

with $\lambda = 1$ corresponding to no transformation, $\lambda = 1/2$ to the square root transformation, $\lambda = 0$ to the logarithmic transformation and $\lambda = -1$ to the reciprocal transformation, thus avoiding a discontinuity at zero.

The development in Box and Cox (1964) is for the normal theory linear model

$$y(\lambda) = X\beta(\lambda) + \epsilon, \tag{2}$$

where X is $n \times p$, $\beta(\lambda)$ is a $p \times 1$ vector of unknown parameters and the variance of the independent errors $\epsilon_i \quad (i = 1, ..., n)$ is $\sigma^2(\lambda)$. The aim of the transformation is to produce a response for which the variance of ϵ_i is constant with an approximately normal distribution. The linear model ideally should also be simple and additive, for example avoiding interaction and quadratic terms. All three aims are satisf ed in the examples given by Box and Cox (1964), as they are in the analysis of numerous other data set, such as those in Atkinson and Riani (2000, Chapter 4).

To estimate λ it is necessary to allow for the change of scale of $y(\lambda)$ with λ. The likelihood of the transformed observations relative to the original observations y includes the Jacobian

$$J = \prod_{i=1}^{n} \left| \frac{\partial y_i(\lambda)}{\partial y_i} \right|. \tag{3}$$

For the power transformation (1), $\partial y_i(\lambda)/\partial y_i = y_i^{\lambda-1}$, so that

$$\log J = (\lambda - 1) \sum \log y_i = n(\lambda - 1) \log \dot{y},$$

where $\dot{y}$ is the geometric mean of the observations. A simple form for the likelihood is found by working with the normalized transformation

$$z(\lambda) = y(\lambda)/J^{1/n} = (y^\lambda - 1)/\lambda \dot{y}^{\lambda-1}. \tag{4}$$

For given λ the parameters are estimated by least squares:

$$\hat{\beta}(\lambda) = (X^T X)^{-1} X^T z(\lambda) \quad \text{and} \tag{5}$$

$$s^2(\lambda) = \{z(\lambda) - X\hat{\beta}(\lambda)\}^T \{z(\lambda) - X\hat{\beta}(\lambda)\}/(n-p) = R(\lambda)/(n-p). \tag{6}$$

If an additive constant is ignored, the profile loglikelihood, partially maximized, over $\beta(\lambda)$ and $\sigma^2(\lambda)$, is

$$L_{\max}(\lambda) = -(n/2)\log\{R(\lambda)/(n-p)\}, \tag{7}$$

so that $\hat{\lambda}$ minimizes $R(\lambda)$.

For inference about plausible values of the transformation parameter λ, Box and Cox suggest likelihood ratio tests using (7), that is, the statistic

$$T_{LR} = 2\{L_{\max}(\hat{\lambda}) - L_{\max}(\lambda_o)\} = n\log\{R(\lambda_0)/R(\hat{\lambda})\}. \tag{8}$$

Although Box and Cox (1964) find the estimate $\hat{\lambda}$ that maximizes the profile loglikelihood, they are careful to stress in their §2 that they are concerned not merely to find a transformation which justifies assumptions, but rather to find, where possible, a metric in terms of which the findings may be succinctly expressed. Typically in linear models, the main interest is in the factor effects, the choice of λ being only a preliminary step. They state "we shall need to fix one, or possibly a small number, of λ's and go ahead with the detailed estimation and interpretation of the factor effects on this particular transformed scale. We shall choose (the estimate $\tilde{\lambda}$) partly in the light of the information provided by the data and partly from general considerations of simplicity, ease of interpretation, etc."

This formulation has led to some controversy in the statistical literature. Bickel and Doksum (1981) and Chen *et al.* (2002) ignore the suggested procedure. They show for regression models with response $y(\lambda)$ that, when the transformation parameter is poorly determined, the variability of the estimated parameters in the linear model can be greatly increased if λ is estimated by $\hat{\lambda}$ rather than by $\tilde{\lambda}$. Box and Cox (1982) and Hinkley and Runger (1984) query the scientific usefulness of such estimates of parameters on an unknown measurement scale. They further comment that the effects observed by Bickel and Doksum would be greatly reduced if the investigation had been conducted in terms of $z(\lambda)$ rather than $y(\lambda)$. McCullagh (2002a), in comments on Chen *et al.*, is very clear about the Box-Cox procedure for choosing λ. In the same discussion Reid (2002) comments

"The Box-Cox model is very useful for the theory of statistics, as a moderately anomalous model in the sense that blind application of conventional theory leads to absurd results." Details are in McCullagh (2002b) and Taylor and Liu (2007). Cox and Reid (1987) use the Box-Cox model as an example of their method for obtaining approximate parameter orthogonality, here between λ and the parameters of the linear model.

The practical procedure is analysis in terms of $z(\lambda)$ leading to $\hat{\lambda}$ and hence to a, hopefully, physically interpretable estimate $\tilde{\lambda}$ chosen from a grid of plausible values. Carroll (1982) argues that the grid needs to become denser as n increases. Indeed, for the small examples of Box and Cox, inverse or logarithmic transformations are indicated. But for the 509 observations on loyalty card usage in Perrotta *et al.* (2009), the value 1/3 is rejected when outliers are removed, but the value 0.4 is acceptable. A final point is that, for comparisons across sets of data, parameter estimates need to be calculated using $y(\tilde{\lambda})$ to avoid dependence on $\dot{y}$.

3 Mental Illness Data: Transformations and the Generalized Linear Model

Kleinbaum and Kupper (1978, p.148) describe observational data on the assessment of mental illness of 53 patients. We compare the Box-Cox transformation with an analysis using a generalized linear model with various Box-Cox links.

A psychiatrist assigns values for mental retardation and degree of distrust of doctors in newly hospitalized patients. After six months of treatment, a value is assigned for the degree of illness of each patient. We explore the Box-Cox transformation of degree of illness with regression on the two initial assessments. The maximum likelihood estimate of λ is 0.046, with 95% confidence limits from the profile log-likelihood of -0.307, 0.404. The data support the log transformation, $\lambda = 0$. There is significant regression on both variables with a t value of 2.88 for the relationship with the initial assessment of retardation and -2.21 for distrust of doctors. The QQ-plots of residuals show an appreciable improvement in normality after transformation.

In the Box-Cox model the transformed response follows a linear model. On the other hand, in generalized linear models the linear model is transformed by the link function. For positive skew continuous data, an alternative to the Box-Cox transformation is a gamma GLM. The canonical link for this GLM is the inverse function, but the log link often provides a good fit to data. There is a strong relationship between the linear model fitted to the logged response and the GLM with a log link. We illustrate this relationship for the Mental Illness data.

With $E(Y) = \mu$ and the linear predictor $\eta = x^T\beta$, the link function relates

Table 1: Mental illness data. Deviances from GLM analyses for three values of the parameter of the Box-Cox link

Link	Deviance
-1	19.18
0	19.56
1	20.12

the two by $\eta = g(\mu)$. For the gamma model the variance is quadratically related to the mean; the variance function $V(\mu) = \mu^2$. We use the Box-Cox link $g(\mu) = (\mu^\lambda - 1)/\lambda$: for $\lambda = -1$ we obtain the reciprocal link, with the log link for $\lambda = 0$. Table 1 gives the deviances from f tting the gamma model with $\lambda = -1, 0$ and 1. Although the reciprocal link yields the smallest deviance, there is no signif cant increase in deviance if one of the other links is used. The insensitivity of data analyses to the exact specif cation of the gamma link is well established - for example the analysis by McCullagh and Nelder (1989, p.377) of their car insurance data. Further discussion of the relationship between the gamma and lognormal models is in McCullagh and Nelder (1989, Chapter 8). Atkinson and Riani (2000, Chapter 6) use the goodness of link test of Pregibon (1980) to provide a fan plot for the parameter in the Box-Cox family of link functions.

We now consider the relationship between the two models f tted to the Mental Illness data. The coeff cient of variation of the untransformed data is taken as constant

$$\text{var}(Y) = \sigma^2\{E(Y)\}^2 = \sigma^2\mu^2,$$

so that σ is the coeff cient of variation of Y. The variance-stabilizing transformation is log(Y). For small σ^2 the approximate moments of log(Y) are

$$E\{\log(Y)\} = \log(\mu) - \sigma^2/2 \quad \text{and} \quad \text{var}\{\log(Y)\} = \sigma^2.$$

If the systematic part of the model is multiplicative on the original scale, coeff - cient estimates of the parameters and of their precision may be obtained by transforming to the log scale and using ordinary least squares. If the exact distribution of Y is known, maximum likelihood estimation for the known distribution should be used. Firth (1988) compares the log-normal and gamma models under reciprocal mis-specif cation, the gamma distribution performing slightly better.

Figure 1 shows the comparison of f tted values for the linear model after log transformation of y with those from the gamma model for two Box-Cox links. In the left-hand panel for the reciprocal link the relationship between the two sets of f tted values is slightly convex. The right-hand panel shows the straight-line relationship for the log link. The plot for the identity link ($\lambda = 1$) is not shown. As is to be expected, it is slightly concave.

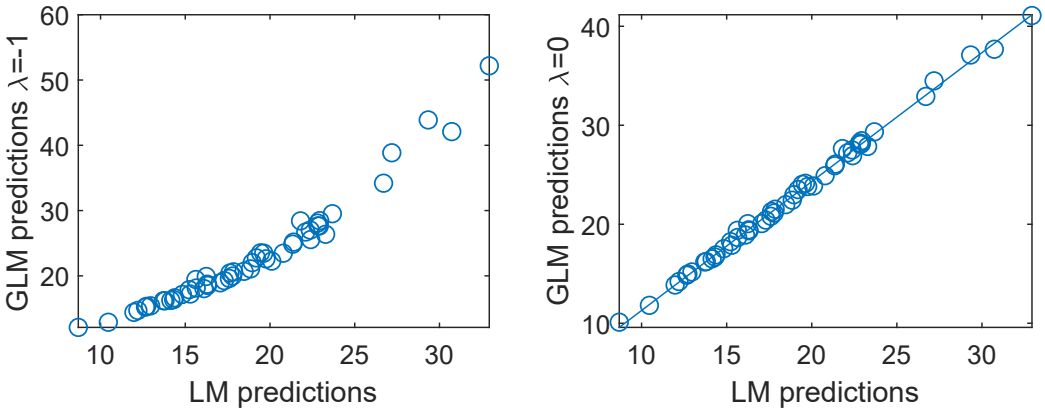

Figure 1: Mental Illness data. Comparison of f tted values from gamma and log-normal models. Left-hand panel, reciprocal link, right-hand panel log link

The close relationship between the gamma and log-normal f ts depends on σ^2 being suff ciently small. For the mental illness data the estimated value of σ^2 is 0.4. Results of a small simulation (Atkinson, 1982) on the choice between the two models for the chimpanzee learning data of Brown and Hollander (1977) lead McCullagh and Nelder to comment that discrimination between the two models may be diff cult even for σ^2 as large as 0.6. The relationship also depends on the observations having a common variance. Wiens (1999) provides an example of two-group data in which the relationship fails to hold due to different variances in the two groups after log transformation.

4 Further Transformations

4.1 Transform Both Sides

There is sometimes a strong, often theoretically derived, relationship between the response and the model $\eta(x, \beta)$, combined with variance heterogeneity. Box-Cox transformation of the response to achieve stability of variance can destroy the relationship between E(Y) and $\eta(x, \beta)$. For example, the kinetic models of chemistry provide deterministic relationships of concentrations of reactants and products on time and temperature. A well-known simple example is the Michaelis-Menten model for enzyme kinetics in which the response goes from zero to an asymptotic value $V_{\max}$. Transforming the response to y^λ would result in a different range for the transformed response.

Carroll and Ruppert (1988, Chapter 4) developed a transform both sides model for such problems, motivated by theoretical models for sockeye salmon breeding. The transformation model is

$$(y^\lambda - 1)/\lambda = \{\eta(x, \beta)^\lambda - 1\}/\lambda + \epsilon, \tag{9}$$

where the independent errors are normally distributed. As with the Box-Cox transformation, the parameters λ and β are found by minimizing the residual sum of squares in the regression model which includes the Jacobian of the transformation, again $\dot{y}$.

The theoretical procedure is to minimize the residual sum of squares using $y(\lambda)/\dot{y}^{\lambda-1}$, or equivalently $y(\lambda)/\dot{y}^\lambda$, as the response and the similarly transformed value of η as the model. Carroll and Ruppert comment that, unless λ is f xed, it is not possible to use standard nonlinear regression routines for this minimization as such routings typically do not allow the response to depend upon unknown parameters. They reformulate the problem in terms of a 'pseudo model', estimation of which converged rapidly in our application.

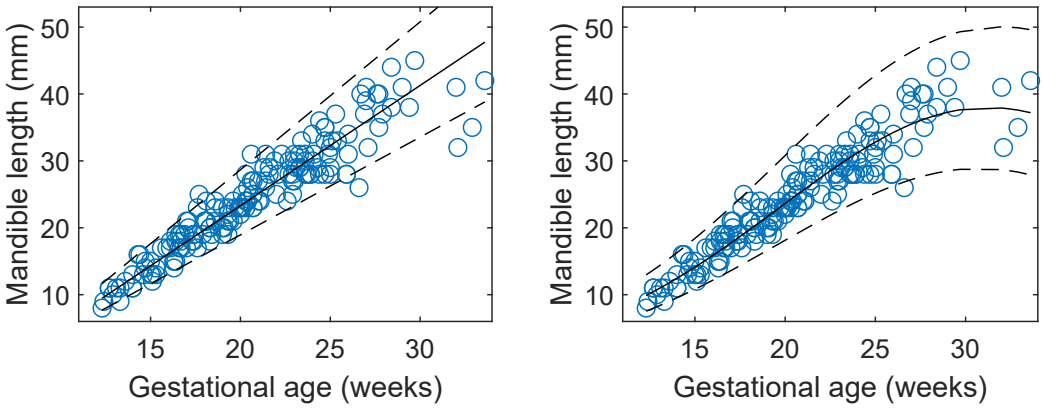

Figure 2: 99% prediction intervals for back-transformed mandible length data. Left-hand panel, transform both sides, $\lambda = 0$. Right-hand panel, logarithmic Box-Cox transformation with a quadratic model.

As our example we use data on mandible length in foetuses used by Royston and Altman (1994) to illustrate the use of fractional polynomials as explanatory variables in regression models. There are 158 observations on foetuses of age x less than 28 weeks. There are also nine measurements with $x > 28$, which the clinicians felt formed a different group with excessive measurement error. The plot of the data in left-hand panel of Figure 2 suggests that mandible length increases linearly with gestational age and that the variance likewise increases.

Royston and Altman overcome the increase in variance by use of the log transformation, but the relationship between log(mandible length) and age is then curved. We use the transformation of both sides to obtain a homoscedastic model in which the linear relationship is preserved.

Regression of all 167 observations on age assuming homoscedasticity indicates only two outliers, rather than 9. We proceed on the assumption that there will be no outliers when we have allowed for heteroscedasticity. The estimate for the Box-Cox transformation of the response suggests a value of 0.75 for λ - a compromise between preserving linearity and a transformation to homoscedasticity. With a logged response there is very strong evidence for inclusion of a quadratic term. The transform both sides model with regression just on age also indicates the log transformation with $\hat{\lambda} = -0.08$.

The left-hand panel of Figure 2 shows the 99% prediction interval for the back-transformed response from the transform both sides model. The fitted model retains the desired linearity and the prediction interval increases with gestational age in line with the heteroscedasticity of the observations. The right-hand panel shows a similar interval for the back-transformed quadratic regression with log y as the response. This panel shows that, although the quadratic model fits well to the majority of the observations, there is increasing curvature for values of $x > 28$.

It is clear that the transform both sides model is to be preferred for predictions over quadratic regression. The difference from predictions using the fractional polynomial model of Royston and Altman is not so obvious. However, the method of transforming both sides preserves the linear relationship between length and age and, more generally, the ability to combine theoretical models with transformation to normality.

4.2 Nonparametric Transformations

The Box-Cox transformation produces a smooth relationship between $y(\lambda)$ and the original y which is determined by the value of λ. The extended Yeo-Johnson transformation of §8.1 for observations that can be positive or negative, likewise produces a smooth relationship but depending on two parameters. These parametric transformations may be too restrictive. A nonparametric alternative is to use some form of smoothing to estimate the relationship, allowing for greater flexibility. This may have advantages in the analysis of a specific set of data, with disadvantages if the aim is to compare different sets of observations which are subject to the same transformation. Figure 9 illustrates the extra information provided by one nonparametric transformation.

The general model is

$$g(y, \kappa) = \eta(x, \beta) + \epsilon, \tag{10}$$

where κ might be a vector of parameters def ning a spline transformation and ϵ is not necessarily normally distributed.

Ramsay (1988) uses a monotone spline to estimate $g(.)$, with the regression parameters estimated by least squares. An advantage is that the Jacobian of the transformation is found by straightforward differentiation of the spline. His analysis of the wool data from Box and Cox produces a transformation very close to that from the parametric analysis. The loglikelihood is not signif cantly improved by increasing the complexity of the spline through the addition of extra knots. Song and Lu (2012) adopt a Bayesian approach. They use penalized Bayesian splines to transform the response to approximate normality by maximizing a normal likelihood with prior distributions for the parameters of $\eta(x, \beta)$. Their plot of the transformation for U-shaped data[1] is sigmoid, far from the convex or concave shapes attainable from the Box-Cox transformation.

The semiparametric method of Foster *et al.* (2001) assumes that the Box-Cox transformation holds, that is $g(y, \kappa) = y(\lambda)$ but that in (10) the distribution of ϵ is unknown. An estimating equation, combined with a grid search over values of λ, provides estimates of λ and β. The covariance matrix of the parameter estimates is found using a resampling method. The parameter estimates and their standard errors for the wool data are virtually indistinguishable from those of Box and Cox. However, the semiparametric method shows appreciable improvement for prediction when the error distribution is not normal. Cai *et al.* (2005) use related methods for the transformation of censored survival times.

Two nonparametric methods use the "supersmoother" (Friedman and Stuetzle, 1982) instead of the spline smoothing of Ramsay (1988). Both methods can transform explanatory variables and response. The assumed model is a generalized additive model, that is one with transformations of both response and explanatory variables but without interactions. Both rely on repeated application of the univariate smoother. In ACE (alternating conditional expectations) Breiman and Friedman (1985) maximize a measure of correlation between all variables; in regression the response variable is not treated as being different from the explanatory variables. Tibshirani (1988) describes a related method in which the transformation for the response is intended to yield additivity and variance stabilization (AVAS). The asymptotic variance stabilizing transformation is estimated for the response. Hastie and Tibshirani (1990, Chapter 7) provide a description of both ACE and AVAS with an emphasis on response transformation and the mathematical relationship to the Box-Cox transformation. Subroutine AREG of the R-package Hmisc (Harrell Jr, 2019) replaces the smoother in ACE with restricted cubic smoothing splines, with a controllable number of knots.

[1] In their Figure 2 these data have a minimum of -1.5. We are assured that, for the calculation of the Box-Cox transformation in their Table 1, the data were shifted to be non-negative.

For the transform both sides model Wang and Ruppert (1995) assume that the observations are normally distributed and use kernel density estimation to determine $g(y, \kappa)$. Since in the transform both sides model the form of the relationship between y and $\eta(x, \theta)$ is known, least squares can be used to estimate β. A companion article (Nychka and Ruppert, 1995) uses splines.

We use ACE, AVAS and AREG as our nonparametric alternatives to the Box-Cox transformation, ACE and AVAS being the most studied of the nonparametric transformations. We exclude the transformation of explanatory variables. The original progams for both ACE and AVAS are written in 'classical' Fortran, without comments and with many non-informative variable names. This Fortran code also provides the basis of the R package acepack (Spector *et al.*, 2016). We have rewritten the programs in Matlab. These new programs have been thoroughly compared with the Fortran programs and validated to give identical numerical results to the originals and incorporated into our toolbox for robust analysis. The output of ACE and AVAS are a set of transformed responses, scaled to have unit variance. Unlike ACE and AVAS, which are free of adjustable parameters, AREG requires the specif cation of the number k of knots in the splines. The aggregate statistic for comparison of models for all three is the value of R^2 which Wang and Murphy (2005) convert into BIC values. Harrell provides routines for bootstrap evaluation of the variances of the estimated linear model parameters obtained from AREG.

Marazzi *et al.* (2009) review papers on the Box-Cox transformation, from the standpoint of computational feasibility and, particularly, robustness. None of the methods have high breakdown and all, for example Carroll and Ruppert (1985), breakdown for outliers at leverage points. In their discussion of Breiman and Friedman (1985), Buja and Kass (1985) comment on the need to develop diagnostics and robust forms of ACE. Some diagnostic information can be obtained by comparing parametric and nonparametric transformations on data before and after the removal of outliers, which we exemplify in §10.

4.3 More Transformations

Extensions of the Box-Cox transformation

For values of λ other than zero, the distribution of $y(\lambda)$ is truncated. For $\lambda > 0$, $y(\lambda)$ is bounded below at $-1/\lambda$; for $\lambda < 0$ it is bounded above at the same value. Only exponentiation of the log normal distribution yields a normal distribution on the whole real line. Yang (2006) introduced a dual transformation $y(\lambda) = (y^\lambda - y^{-\lambda})/2\lambda, (\lambda \neq 0)$ with the logarithmic transformations at $\lambda = 0$, which removes the bound in the Box-Cox transformation.

Zhang and Yang (2017) describe a method for applying the Box-Cox transformation to huge data sets. The necessity is to avoid storing all the data in the

computer before performing transformation calculations. For least squares regression the required quantities (sums of squares and products of y and X) can be sequentially updated. The procedure can be extended to the Box-Cox transformation to include storing sums of products of X and $z(\lambda)$ for selected values of λ. Zhang and Yang (2017) choose a grid of 41 values.

Box and Cox (1964) extended their transformation to the shifted power transformation of $(y + \mu)$, where both μ and the transformation parameter λ are to be estimated. A diff culty is that the range of the observations now depends on μ, so that the inferential problem is non-regular. Atkinson *et al.* (1991) suggest a grouped likelihood approach to parameter estimation, but the estimates may depend on the size of the grouping interval.

In §10 we analyse data from John and Draper (1980). The normal plot of the residuals (Atkinson, 1985, Figure 9.17) shows a long tailed symmetrical distribution, which structure led John and Draper to develop the modulus transformation with

$$y(\lambda) = \left\{ \frac{(|y| + 1)^{\lambda} - 1}{\lambda} \right\} \text{sign}(y),$$

for $y \neq 0$. This symmetric transformation family applies the same transformation to the positive and negative tails of the distribution. The Yeo-Johnson transformation, which can also be applied to observations that can be negative or positive, is either convex or concave over the whole range of y, whereas the extended Yeo-Johnson transformation of §8.2 can be convex or concave in either tail as the data dictate.

The two transformations of Aranda-Ordaz (1981) provide invertible transformations for binary data. In the symmetrical transformation in which "successes" and "failures" are interchangeable, the value $\lambda = 0$ yields the logistic model. In the asymmetrical transformation the limits are the complementary log log and logistic models.

These methods are for independent univariate responses. The Box-Cox transformation was generalized to multivariate data by Andrews *et al.* (1971) and Gnanadesikan (1977). Velilla (1995) considers robust and diagnostic aspects of multivariate transformations. Atkinson *et al.* (2004) provide examples of the analysis of transformed multivariate data using the forward search.

A more general point is inference for transformed data on the original scale. The properties of predictions on the back-transformed scale are considered by many, including Taylor (1986) and Carroll and Ruppert (1988). A second point is that Box and Cox also develop a Bayesian procedure for transformation, which leads to a data-dependent prior. Pericchi (1981) suggested a prior that avoided data-dependence, which was modif ed by Sweeting (1984). Gottardo and Raftery (2009) combine Bayesian transformations with model selection.

Transformation of Explanatory Variables

Box and Tidwell (1962) explore power transformations of the explanatory variables in regression. Since the response is not transformed, residual sums of squares can be compared directly for different transformations.

Transformation of the response in ARIMA models results in transformation of any lagged responses in the model. The constructed variables of Atkinson *et al.* (1997) for Box-Cox transformation of ARIMA models were used by Riani (2009) and Proietti and Riani (2009) in fan plots for the transformation of time series.

Transformation of Parameters

The lower left-hand panel of Figure 8.2 of McCullagh and Nelder (1989) shows a virtually parabolic loglikelihood for a single gamma observation when the plot is against $\mu^{-1/3}$. For multiparameter problems approximate orthogonality and a nearly quadratic form of the log likelihood will usually speed the convergence of iterative methods of estimation. This is a matter of numerical analysis, but approximate independence of the components of parameters combined with approximate normality is also desirable for statistical reasons, including ease of interpretation in multiparameter problems. Ross (1990) includes many examples.

5 Robustness and Graphics

The data analyses so far are based on aggregate statistics. They do not allow for the presence of dispersed or grouped outliers, or for inf uential observations, one or a few of which may appreciably change the estimate of the transformation parameter and so the interpretation of the data. Several robust statistical methods address this problem, at least for many statistical models, such as regression, if not for data transformation. A diff culty in the intelligent application of robust methods is that many require the specif cation of a parameter dependent on the amount of contamination expected in the data or the required eff ciency of estimation.

There are three general classes of approaches to robust regression: (i). *Soft Trimming* (downweighting). M estimation and derived methods (Huber, 1973). Observations near the centre of the distribution retain their value, but observations far from the centre have a weight that decreases with distance from the centre; (ii). *Hard Trimming.* In Least Trimmed Squares (LTS: Hampel, 1975, Rousseeuw, 1984) the amount of trimming is determined by the choice of the trimming parameter h, which is specif ed in advance. The LTS estimate is intended to minimize the sum of squares of the residuals of h observations and (iii). *Adaptive Hard Trimming.* In the Forward Search (FS), the observations are again hard trimmed, but the value of h is determined adaptively by the data. Starting from a small initial subset of data, the number of observations used in f tting then increases until all are included and outliers identif ed. Atkinson *et al.* (2010) provide a general survey of the FS with discussion.

Two properties of these robust regression estimators are important when selecting a method: (i). *Breakdown Point, bdp* ; the asymptotic proportion of observations that can go to ∞ without affecting the parameter estimates. This def nition requires both that $n \to \infty$ and also that the distance between the contaminated and uncontaminated observations increases with n; (ii). *Eff ciency of Estimation,* of the parameters relative to least squares for uncontaminated data.

Ideally a robust estimator would have both a high breakdown point and a high eff ciency. Unfortunately this is not possible. For hard trimming, once one of the values, for example the breakdown point bdp, has been selected, the other is determined. Riani *et al.* (2014) extended these results to S-estimation. To illuminate the non-asymptotic properties of robust estimators Riani *et al.* (2014) monitor the behaviour of several extensions of M estimation, including MM estimation (Yohai, 1987), by analysing data over a grid of values of the eff ciency of estimation of the parameters of the linear model. They observe that there is often a point at which the f t switches from being robust to non-robust least squares. This important property, which at present cannot be determined analytically, depends both on the nominal properties of the estimator and on the particular data set being analysed.

The examples in Riani *et al.* (2014) indicate that the FS, combined with a suitable stopping rule to avoid the inclusion of outliers, provides a robust procedure with good properties which avoids any *a priori* specif cation of quantities indicating the required degree of robustness. We therefore use it as the method for robust estimation of transformations. Details of the method are in §6.2.

The FS by its nature provides a series of f ts to subsets of the data of increasing size. Forward plots of residuals, that is of residuals as the subset size m increases, are informative about the presence of outliers. They are used both as a tool to determine outliers and as a means of understanding the structure of the data. The left-hand panel of Figure 6 illustrates the outlier detection procedure. The panels of Figure 8 show the information gained by linking plots, making clear the effect of individual observations on the estimated transformation parameter, the test for outliers, the trajectory of residuals over the FS and the position of the observations on scatter plots. A different use of dynamic graphics in the determination of robust transformations is in Seo (2019)

6 A Robust Approximate Score Test for the Transformation Parameter

6.1 Constructed Variables

For inference Box and Cox (1964) rely on complete-sample likelihood inference through the likelihood ratio statistic (8). A disadvantage of this likelihood ratio test is that a numerical maximization is required to f nd the value of $\hat{\lambda}$. In our robust procedure using the FS, we calculate almost n test statistics for the hypothesis that $\lambda = \lambda_0$, typically for f ve values of λ_0. There is an appreciable literature on methods that avoid such maximizations of the likelihood: score tests (Cook and Weisberg, 1982; Atkinson, 1985) and Lagrange multiplier tests (Breusch and Pagan, 1979).

We use the approximate score statistic $T_p(\lambda)$, (Atkinson, 1973) derived by Taylor series expansion of $z(\lambda)$ (4) about λ_0. This leads to the approximate regression model

$$\begin{aligned} z(\lambda_0) &= x^T\beta - (\lambda - \lambda_0)w(\lambda_0) + \epsilon \\ &= x^T\beta + \gamma\, w(\lambda_0) + \epsilon, \end{aligned} \tag{11}$$

where $\gamma = -(\lambda - \lambda_0)$ and the constructed variable $w(\lambda_0) = \partial z(\lambda)/\partial\lambda|_{\lambda=\lambda_0}$, which only requires calculations at the hypothesized value λ_0.

The approximate score statistic for testing the transformation is the t statistic for regression on $-w(\lambda_0)$, that is the test for $\gamma = 0$ in the presence of all components of β. Because $T_p(\lambda_0)$ is the t test for regression on $-w(\lambda_0)$, large positive values of the statistic mean that λ_0 is too low and that a higher value should be considered.

A different approximate score statistic for the Box-Cox transformation is found by Lawrance (1987) through an approximation to the variance of the score statistic, leading to an improved null distribution for the statistic. Some numerical comparisons of the two procedures are in Atkinson and Lawrance (1989). A similar procedure for testing the value of the parameter in the Yeo-Johnson transformation is shown in §9.

6.2 The Fan Plot

The robust transformation of regression data is complicated by the dependence of outliers on the value of λ. In the data of Wiens mentioned in §3, very small values, arbitrarily allocated to observations below the detection limit, have an appreciable effect when the data are log transformed. Atkinson and Riani (2000) show how different observations appear outlying for various transformations of the Poison data of Box and Cox (1964).

We use the Forward Search to provide a robust plot of the approximate score statistic $T_p(\lambda)$. We start with a fit to $m_0 = p + 1$ observations and then successively fit to larger subsets. For the subset of size m we order all observations by closeness to the fitted model; the residuals determine closeness. The subset size is increased by one to consist of the subset with the $m+1$ smallest squared residuals and the model is refitted to this new subset. The process continues with increasing subset sizes until, finally, all the data are fitted. The process moves from a very robust fit to non-robust least squares. Any outliers will enter the subset towards the end of the search. We thus obtain a series of fits of the model to subsets of the data of size $m, m_0 \leq m \leq n$ for each of which we refit the model and calculate the value of the score statistics for selected values of λ_0. These are then plotted against the number of observations m used for estimation to give the "fan plot". As Figure 4 shows, the ordering of the observations in a fan plot may depend on the value of λ_0.

Since the constructed variables are functions of the response, the statistics cannot exactly follow the t distribution. Atkinson and Riani (2002) provide some numerical results on the distribution in the fan plot of the score statistic for the Box-Cox transformation. They find that departures from the null distribution are most extreme towards the end of the search, where the statistic has too large a variance; increasingly strong regression relationships lead to null distributions that are closer to t.

7 Mental Illness Data: A Robust Analysis

Section 3 introduced data on 53 patients and provided an analysis based on aggregate statistics, which indicated the log transformation of the response. Here we compare the Box-Cox transformation with three nonparametric transformations, both on the original data and on a version contaminated with outliers, and illustrate the use of the fan plot in the detection of outliers influential for the parametric data transformation.

Original data. The left-hand panel of Figure 3 shows the fan plot for five values of λ_0, fanning out as the search progresses. The trajectory of values of the score test for the log transformation ($\lambda_0 = 0$) remains within the 99% limits (± 2.58) throughout the search. Other values for λ_0 are rejected, -1 and +1 more strongly than ± 0.5. There are no abrupt changes in the trajectories which might indicate the inclusion of an influential observation in the subset of observations used in fitting. The log transformation is further confirmed by a comparison of the QQ plots of residuals which is straightened by the transformation; approximate normality has been achieved.

The right-hand panel of Figure 3 shows the monotonic transformation for these

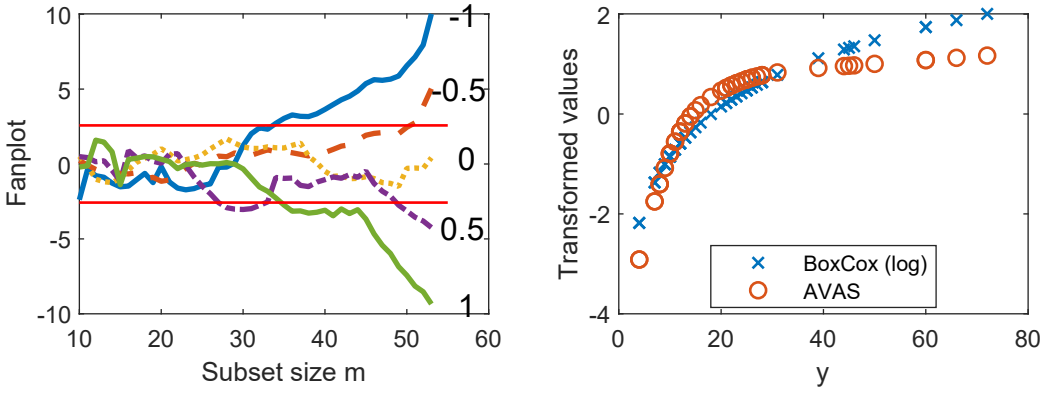

Figure 3: Mental illness data. Left-hand panel, fan plot indicating the log transformation; right-hand panel, transformed and original response values for two transformations

data found by AVAS (we are not transforming the explanatory variables). The figure also includes the Box-Cox logarithmic transformation. In order to compare these two transformations we have scaled the transformed values of the observations to have mean zero and variance one and plotted these values against the untransformed observations. The transformation found by AVAS is the more highly curved in the figure, closely corresponding to the inverse transformation. The effect of the inverse transformation is that the fitted model exhibits two outliers, which are not present for the log transformed data.

Because transformations need to be invertible, the ACE transformation is constrained to be monotonic. The result is almost a straight line, that is no transformation. For the transformation with AREG, the calculation for $k = 3$, yields a virtually straight line transformation, but convex rather than concave. For $k = 4$ and 5 the transformation is decreasing. These four transformations, maximizing R^2, give virtually the same value of R^2 as that of the original data. The Box-Cox transformation and AVAS yield transformations with reduced values of R^2, but residuals with improved normality, desirable for inferential purposes.

Contaminated data. We now study the effect of outliers by modifying three of the smallest observations (17, 30 and 53) to have the value 1. The intention is to produce large outliers on the reciprocal scale, which have little effect on the untransformed data and so influence $\hat{\lambda}$ towards one. The fan plot for these contaminated data is in Figure 4. The effect is dramatic. For four values of λ_0, the three outliers enter at the end of the search, causing the trajectories for $\lambda_0 = -1, -0.5$ and 0 to move appreciably 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 just on a fit to all the data would be 0.5. However, this value is rejected earlier in the search.

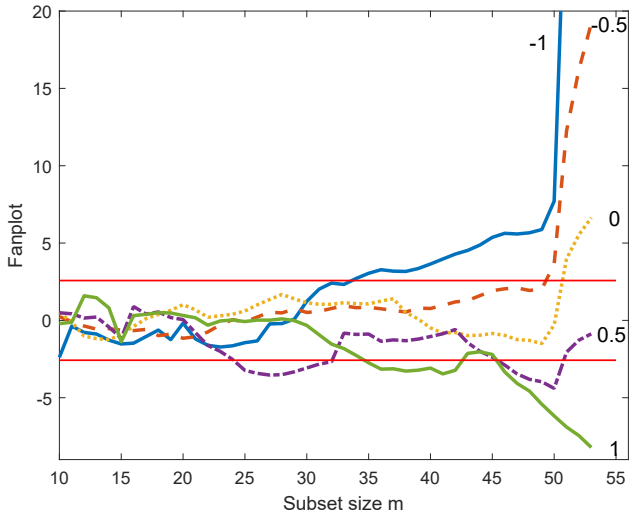

Figure 4: Contaminated mental illness data: fan plot, showing the effect of three changed observations

A plot similar to the right-hand panel of Figure 3 for the contaminated data shows that the curvature in the plot of transformed against original y from AVAS is reduced; the transformation is close to the log transformation. Regression with the log transformation on the contaminated data then clearly shows that the three altered observations have large residuals. The transformation from ACE is, as previously, virtually a straight line. The transformations indicated by AREG for k up to 5 are similar to those in the absence of outliers. In this example the value of k is not clearly determined by the data. Regression using the transformation suggested by AVAS indicates the three outliers for the contaminated data and the logarithmic transformation, but suggests an inadequate transformation for the uncontaminated data.

Section 2 of the supplementary material for this paper includes an analysis of the motivating data set from Chen *et al.* (2002) on gasoline consumption. The data appear to need transformation with λ around 1.5. However, application of the fan plot reveals that all evidence for this transformation comes from one observation, with by far the lowest values of both x and y. The second example in the supplementary material is an analysis of the Poison data from Box and Cox (1964). The fan plot, like that in Figure 3, is exemplary, with no indication of influential observations.

8 The Transformation of Responses that can be Negative as well as Positive

8.1 The Yeo-Johnson Transformation

Yeo and Johnson (2000) extended the Box-Cox transformation to observations that can be positive or negative by using different Box-Cox transformations for the two classes of response. For $y \geq 0$ this is the generalized Box-Cox power transformation of $y + 1$. For negative y the transformation is of $-y + 1$ to the power $2-\lambda$. They used a smoothness condition to combine the transformations for positive and negative observations, so obtaining a one-parameter transformation family. Atkinson *et al.* (2020) extended the transformation to allow for distinct transformation parameters for the two response classes. They further provided constructed variables for this extended transformation and an extended fan plot which permits checking the correctness of the two transformations. This section brief y summarizes their results.

As with the Box-Cox transformation, analysis of data from this transformation needs to include the Jacobian J of the transformation to allow for changes of scale as λ varies. We continue to work with a normalized transformation $z(\lambda) = y(\lambda)/J^{1/n}$ in which the Jacobian is spread over all observations. If, to extend the notation of Box and Cox, $\dot{y}_{YJ}$ is the nth root of J, it follows from equation (3.1) of Yeo and Johnson (2000) that

$$\dot{y}_{YJ} = \exp\left[\sum\{\text{sgn}(y_i)\log(|y_i|+1)\}/n\right]. \tag{12}$$

The normalized version of the transformation is then

$$\begin{array}{lll} y \geq 0: & \dfrac{(y+1)^\lambda - 1}{\lambda \dot{y}_{YJ}^{\lambda-1}} & (\lambda \neq 0); \quad \dot{y}_{YJ}\log(y+1) \quad (\lambda = 0) \\ y < 0: & -\dfrac{\{(-y+1)^{2-\lambda} - 1\}}{(2-\lambda)\dot{y}_{YJ}^{\lambda-1}} & (\lambda \neq 2); \quad -\log(-y+1)/\dot{y}_{YJ} \quad (\lambda = 2). \end{array} \tag{13}$$

8.2 The Extended Yeo-Johnson Transformation and Homogeneity of Transformation

Some authors, for example Weisberg (2005), query the physical interpretability of the constraint that positive and negative observations should be transformed by the same value of λ, which is indeed violated by the data analysed in §10. Accordingly, Atkinson *et al.* (2020) extended the score test to testing for the equality of the value of λ in the two regions of y. The procedure takes the transformation

parameter as λ for one part and $\lambda+\alpha$ for the other and uses the score testing procedure for $\alpha = 0$, leading to tests that positive, or negative y_i need a transformation different from λ.

There are now separate Jacobians, $\dot{y}_P$ and $\dot{y}_N$, for positive and negative y from breaking the summation in (12) into parts for positive and negative observations. To test for positive observations having a distinct transformation let $v = y + 1$. The general model for $y \geq 0$ is

$$z(\alpha, \lambda) = \frac{v^{\lambda+\alpha} - 1}{(\lambda + \alpha)\dot{y}_N^{\lambda-1}\dot{y}_P^{\lambda+\alpha-1}}, \tag{14}$$

which reduces to the standard transformation when $\alpha = 0$ since $\dot{y} = \dot{y}_N\dot{y}_P$.

When $y < 0$ let $v_N = -y + 1$. Keeping the parameter for positive y as $\lambda + \alpha$ the model for the negative y only depends on α through the Jacobian. Then

$$z(\alpha, \lambda) = -\frac{v_N^{2-\lambda} - 1}{(2 - \lambda)\dot{y}_N^{\lambda-1}\dot{y}_P^{\lambda+\alpha-1}}. \tag{15}$$

Similar expressions when the parameter for negative y is $\lambda + \alpha$ are given by Atkinson *et al.* (2020).

8.3 The Extended Fan Plot; Checking Postulated Transformations

The original Yeo-Johnson transformation of §8.1 yields a score test and so a fan plot for a set of values of λ_0. The extended Yeo-Johnson transformation of §8.2 provides constructed variables for testing whether the positive and negative observations also require the transformation λ_0. The extended fan plot accordingly contains three trajectories for each value of λ_0. If the same transformation is applied to both positive and negative responses, agreement of the three trajectories indicates that only one transformation is needed.

An important feature of the extended fan plot is that it provides a method of testing a proposed transformation with different parameters, λ_P and λ_N for transformation of the positive and negative observations. Once the data have been correctly transformed, the extended fan plot testing $\lambda_0 = 1$ for the transformed data should lie within the bounds for all values of m. We use this method in §§9 and 10 to conf rm transformations of the data which have distinct values of λ_P and λ_N.

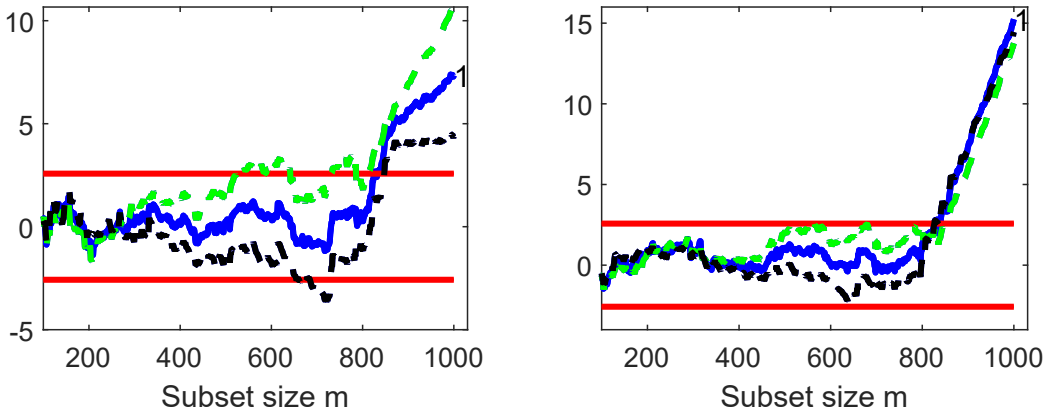

Figure 5: Simulated data. Extended fan plots with green (mostly upper) curve for positive observations: left-hand panel, $\lambda_0 = 1$, indicating inf uential observations and the need for two transformation parameters; right-hand panel, checking $\lambda_N = 0.25$ and $\lambda_P = 1.5$ (test of $\lambda_0 = 1$ for transformed data), conf rming transformation parameter values and suggesting the presence of potential outliers.

9 A Simulated Example

We simulated 1,000 regression observations, with three explanatory variables and independent normal errors, which were heavily contaminated with outliers. We f rst demonstrate the use of the augmented fan plot in indicating a transformation and then use the forward search on the transformed data to detect the outliers. We compare this procedure with the information from using nonparametric transformations on the contaminated data.

Contaminated data. Different transformations were applied to the positive and negative observations and 200 of the observations were shifted to provide outliers. This is a challengingly high contamination proportion unless the outliers are distinct. However, they are not evident on the scatterplots of y against the individual x vectors (Figure 1 of the supplementary material). Linear regression gave an R^2 value of 0.31. The fan plot for these data for $\lambda_0 = 0.5, 0.75, 1, 1.25$ and 1.5 indicated a value of 1.25 for the overall transformation parameter, with all score statistic values lying within the 99% interval. For other parameter values a large number lie outside the interval; around 200 for $\lambda_0 = 1$.

To investigate whether an overall transformation with $\lambda = 1.25$ is satisfactory we calculated extended fan plots for a few values of λ. That for $\lambda_0 = 1$ in the left-hand panel of Figure 5 shows clear evidence, from $m = 400$ or less, that different values are needed for λ_N and λ_P. The plot also shows the effect of a set of inf uential observations entering in the last 160 steps.

The extended fan plot is used to find pairs of parameter values. The data are transformed with pairs of values λ_P and λ_N and the extended fan plot for $\lambda_0 = 1$ for the transformed data inspected. The right-hand panel of Figure 5 shows the plot for $\lambda_N = 0.25$ and $\lambda_P = 1.5$. The trajectories for the positive and negative observations are close together and close to the trajectory for a single value of λ. In addition the influential observations are clearly articulated.

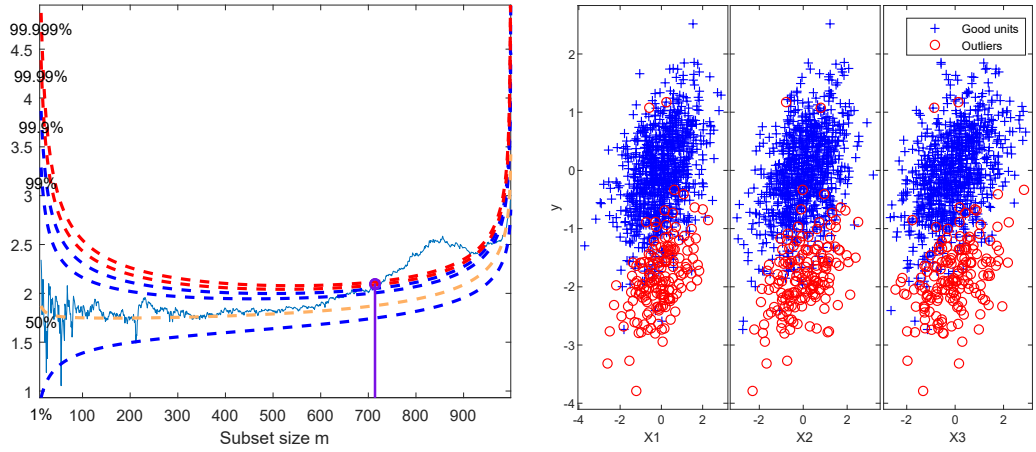

Figure 6: Simulated data, detection of outliers using the forward search on data transformed with $\lambda_N = 0.25$ and $\lambda_P = 1.5$. Left-hand panel, forward plot of minimum deletion residuals. Right-hand panel, scatter plots of the transformed data against explanatory variables; the 164 observation identified as outliers are plotted as open (red) circles.

To check for outliers, we perform a robust analysis on transformed data with $\lambda_N = 0.25$ and $\lambda_P = 1.5$. The left-hand panel of Figure 6 shows a plot of (absolute) minimum deletion residuals from the forward search on the transformed data. For each m these are the outlier tests for the next observation to enter the subset, which is the one closest to the already fitted model. The envelopes in the plot provide guidance as to whether the observation is outlying. The procedure for outlier detection is that of Riani *et al.* (2009) adapted to regression. As a result 164 outliers were identified. The scatter plots in the right-hand panel of the figure show outliers as circles. The R^2 for this regression is 0.518, an appreciable improvement over the original 0.31.

Nonparametric transformations. We now consider nonparametric transformations of the response in regression. AVAS yields a virtually linear relationship between the transformed and original y, that is effectively no transformation at all. The value of R^2 is 0.303, slightly worse than untransformed regression. Results for ACE are in the two panels of Figure 7. Now the plot of transformed against residual y is virtually a straight line for positive y, but there is a bend in

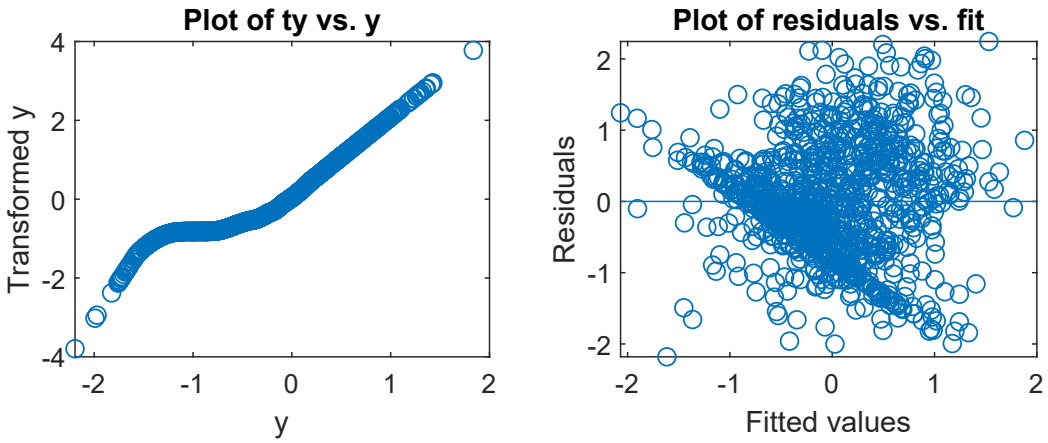

Figure 7: Simulated data. Properties of simulated data transformed with ACE: left-hand panel, transformed y against y; right-hand panel, residuals against f tted values

the curve near $y = 0$. The transformed y are constant for many negative values of y, the downward slope resuming for the most extreme values. The plot of residuals against f tted values shows a cluster, but also a diagonal band of points, indicating some structure in the data that has not been explained by the regression model. The value of R^2 is 0.369, an improvement on the value of 0.31 for the untransformed data. For $k = 5$, AREG produces a transformation similar to that of ACE with an R^2 value of 0.355. For $k = 6$, the value of R^2 is larger, but the transformation is non-monotonic.

The data were, in fact, generated with three explanatory variables drawn independently from standard normal distributions. All regression coeff cients were 0.3 and the error standard deviation was 0.5. Once the data had been generated, a value of 1.9 was subtracted from 200 responses. The negative observations were then transformed so that the value of λ_N was 0. For the positive observations $\lambda_P = 1.5$. Subtraction of 1.9 from 200 observations, some of which were originally negative and some not, has meant that some uncontaminated positive observations have become negative when contaminated. In the data generating process such observations have been transformed to require transformation with λ_N rather than λ_P. We have recovered the transformation parameters for both positive and negative observations. Regression on the uncontaminated generated data, before they had been transformed, gave an R^2 value of 0.538, slightly greater than the 0.518 we found without knowing the number of outliers. Our simulation procedure was designed to increase the complexity of the data. The combination of the extended Yeo-Johnson transformation and robust outlier detection has recovered the transformation used in this complicated data generating process.

10 John and Draper Difference Data

Section 4 of the supplementary material contains part of an analysis of 1,405 observations on the profit or loss of individual firms of which 407 make a loss. Further analysis is in Atkinson *et al.* (2020) which uses the procedure exemplified in §9. Here we robustly find a transformation for data from John and Draper (1980), delete outliers on that scale and then compare parametric and nonparametric transformations on the "cleaned" data. The readings are on the subjective assessment of the thickness of pipe. Five inspectors assessed wall thickness at four different locations on the pipe. The experiment was repeated three times. The sixty responses are a multiple of the difference between the inspector's assessment and the 'true' value determined by an ultrasonic reader. If both readings were available, the Box-Cox transformation could be applied to all 120 readings and the differences analysed in the transformed scale. But the ultrasonic readings are no longer known.

The fan plot of the data when the Yeo-Johnson transformation is applied shows large increases in the values of the score statistics when the last six or seven observations are included in the subset used for fitting; possible values of λ up to this point are between 0.75 and 1.5. Since $\lambda = 1$ is a possible value, we use the untransformed data to check for outliers.

The top left-hand panel of Figure 8 is the fan plot for $\lambda_0 = 1$. The last six observations to enter the subset, marked by red dots in the online version, produce changes in the value of the score statistic, all in the same direction. The top right-hand panel shows the forward plot of minimum deletion residuals. The red line shows that the six influential observations are also outlying, as are many other observations. The linked plot in the bottom-left hand panel, a forward plot of scaled residuals, shows that the six observations (lowest, red, lines) have large negative residuals at least from $m = 30$. The last panel of the figure is a scatter plot of y against the first three of the explanatory variables, with the outlying observations again marked in red.

We "cleaned" the data by removing these six observations. Extended fan plots for the remaining 54 observations show that positive and negative observations require distinct transformations; 1.5 for the negative observations and 0 for the positive ones.

Table 2 gives the values of R^2 for regression with four transformations of the response. Unconstrained ACE gives a value of 0.364, compared with 0.355 for the constrained version. The plots of transformed against untransformed y for these two transformations are similar. The value from the extended Yeo-Johnson transformation is slightly less at 0.336. It is surprising that AVAS performs so poorly, producing an R^2 of 0.275. It is to be expected that a nonparametric transformation with its flexibility of shape will produce a better transformation than one with

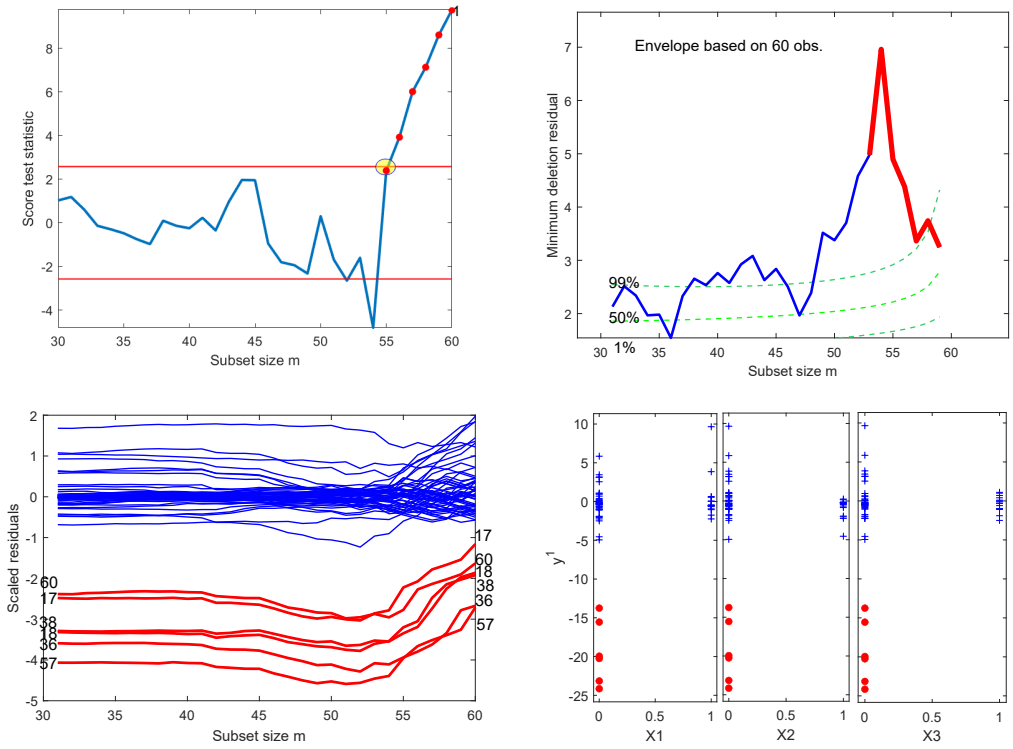

Figure 8: Cleaned difference data; brushing linked plots from the forward search when $\lambda = 1$. Upper left-hand panel, fan plot for $\lambda_0 = 1$; six inf uential observations are highlighted. Upper right-hand panel, forward plot of deletion residuals with the six inf uential observations highlighted. Lower left-hand panel, forward plot of scaled residuals; the six outliers have the six lowest trajectories. Lower right-hand panel, scatter plot of y against $x_1 - x_3$ with outliers plotted as red dots

only a few adjustable parameters. However there are only 54 observations in the cleaned data, which may be small for smoothing methods.

The top left-hand panel of Figure 9 gives the plot of transformed against original observations for the extended Yeo-Johnson transformation with $\lambda_N = -1.5$ and $\lambda_P = 0$. The curve is convex for negative y and concave for positive y, resulting in an overall sigmoid shape. The centre panel shows the results of the unconstrained transformation with ACE. The most negative observations are transformed to a convex curve; there is then a series of virtually constant transformed values before a point of inf ection at $y = 0$, above which the transformation is almost a straight line, that is no transformation. The transformation when ACE is constrained to be monotonic is found by isotonic regression on the transformation in the f gure; it is similar in shape but with a horizontal central section.

Table 2: Values of R^2 for four transformations of the John and Draper difference data with six outliers deleted

Transformation	R^2
ACE unconstrained	0.364
ACE with monotonicity constraint	0.355
Extended Yeo-Johnson	0.336
AVAS	0.275

Both procedures give a plot of residuals against f tted values similar to that in the lower centre panel, with a strong diagonal band and a scatter of points. The transformation from AVAS in the right-hand panel is gently convex up to $y = 0$ and more sharply concave thereafter. The residuals plotted in the lower left- and right-hand panels, have a more random scatter than those from ACE, lacking the diagonal band. Finally, the transformations found by AREG for k from 3 to 6 are non-monotonic. They are shown in Figure 11 of the supplementary material.

The plot of the ACE transformed data in Figure 9 indicates that a smooth power transformation such as the extended Yeo-Johnson, does not adequately catch the structure of the data. The panels of Figure 8 show that there are many outliers, both positive and negative in the untransformed scale and a band of observation in the lower-left panel with small residuals. Although the ACE transformation hardly transforms these observations and brings in both tails of the distribution, the plot of residuals against f tted values in Figure 9 suggests that there is some structure in the data that may need the addition of further terms to the linear model.

11 Conclusions: nonparametric Transformations

The data analyses in this paper show that the nonparametric transformations can provide guidance in the choice of a parametric transformation, as well as indications of its inadequacy. AVAS indicates the inverse transformation for the mental illness data of §7, whereas the Box-Cox transformation is logarithmic. For the contaminated data AVAS f nds the log transformation. For these data constrained ACE provides transformations which are almost linear. On the other hand, for the simulated data of §9, ACE on the contaminated data produces a transformation with an inf ection near $y = 0$ and another near $y = -1$. The plot of residuals from ACE suggests that the data may contain some further structure that needs modelling. The results of this analysis with ACE are similar to that for the difference data in §10 once the data have been "cleaned"; again there is a transformation with

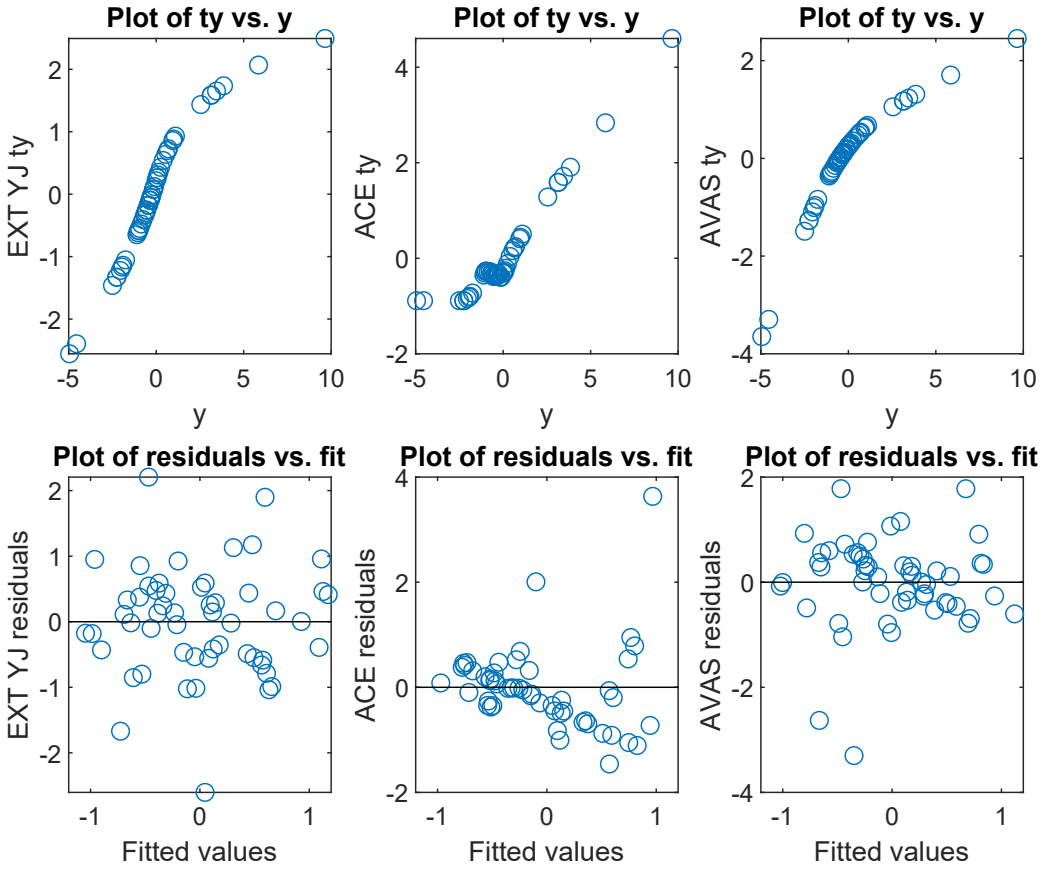

Figure 9: Cleaned difference data; properties of data transformed by the extended Yeo-Johnson transformation, ACE and AVAS. Left-hand column, extended Yeo-Johnson; central column, ACE; right-hand column, AVAS. Upper row transformed y against y; lower row, residuals against f tted values

perhaps three zones and a non-random residual plot. For these examples AVAS produces respectively a virtually straight line transformation and a smooth concave curve. The transformations indicated by AREG for these examples are either non-informative, that is no transformation is indicated, or are non-monotonic.

Unlike AREG, but like the forward search, ACE and AVAS have the advantage that the methods do not require the advance specif cation of parameters. Further results in the supplementary material indicate that ACE and AVAS may behave well for non-negative data. However, for the balance sheet data, both ACE and AVAS are inf uenced by the outliers in the data. A promising strategy is that of §10 in which the fan plot, a robust method, is used to indicate a parametric transformation and a scale in which the data can be cleaned. Parametric and nonparametric transformations can then be compared on data without outliers.

The calculations in this paper used Matlab routines from the FSDA toolbox, available as a Matlab add-on from the Mathworks f le exchange `https://www.mathworks.com/matlabcentral/fileexchange/`. The data, the code used to reproduce all results including plots, and links to FSDA routines are available at `http://www.riani.it/ARC2019`.

Acknowledgements

This research benef ts from the HPC (High Performance Computing) facility of the University of Parma. We acknowledge f nancial support from the University of Parma project "Statistics for fraud detection, with applications to trade data and f nancial statements" and from the Department of Statistics, London School of Economics.

References

Andrews, D. F., Gnanadesikan, R., and Warner, J. L. (1971). Transformations of multivariate data. *Biometrics*, **27**, 825–840.

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response models. *Biometrika*, **68**, 357–363.

Atkinson, A. C. (1973). Testing transformations to normality. *Journal of the Royal Statistical Society, Series B*, **35**, 473–479.

Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables (with discussion). *Journal of the Royal Statistical Society, Series B*, **44**, 1–36.

Atkinson, A. C. (1985). *Plots, Transformations, and Regression*. Oxford University Press, Oxford.

Atkinson, A. C. and Cox, D. R. (1988). Transformations I. *Wiley Statsref*.

Atkinson, A. C. and Lawrance, A. J. (1989). A comparison of asymptotically equivalent tests of regression transformation. *Biometrika*, **76**, 223–229.

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

Atkinson, A. C. and Riani, M. (2002). Tests in the fan plot for robust, diagnostic transformations in regression. *Chemometrics and Intelligent Laboratory Systems*, **60**, 87–100.

Atkinson, A. C., Pericchi, L. R., and Smith, R. L. (1991). Grouped likelihood for the shifted power transformation. *Journal of the Royal Statistical Society, Series B*, **53**, 473–482.

Atkinson, A. C., Koopman, S. J., and Shephard, N. (1997). Detecting shocks: Outliers and breaks in time series. *Journal of Econometrics*, **80**, 387–422.

Atkinson, A. C., Riani, M., and Cerioli, A. (2004). *Exploring Multivariate Data with the Forward Search*. Springer–Verlag, New York.

Atkinson, A. C., Riani, M., and Cerioli, A. (2010). The forward search: theory and data analysis (with discussion). *Journal of the Korean Statistical Society*, **39**, 117–134. doi:10.1016/j.jkss.2010.02.007.

Atkinson, A. C., Riani, M., and Corbellini, A. (2020). The transformation of prof t 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.

Bliss, C. I. (1934). The method of probits. *Science*, **79**(2037), 38–39.

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.

Box, G. E. P. and Tidwell, P. W. (1962). Transformations of the independent variables. *Technometrics*, **4**, 531–550.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and transformation (with discussion). *Journal of the American Statistical Association*, **80**, 580–619.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coeff cient variation. *Econometrica*, **47**, 1287–1294.

Brown, B. W. and Hollander, M. (1977). *Statistics: A Biomedical Introduction*. Wiley, New York.

Buja, A. and Kass, R. E. (1985). Comment on "Estimating optimal transformations for multiple regression and transformation" by Breiman and Friedman. *Journal of the American Statistical Association*, **80**, 602–607.

Cai, T., Tian, L., and Wei, L. J. (2005). Semiparametric Box-Cox power transformation models for censored survival observations. *Biometrika*, **92**, 619–632.

Carroll, R. J. (1982). Prediction and power transformations when the choice of power is restricted to a finite set. *Journal of the American Statistical Association*, **77**, 908–915.

Carroll, R. J. and Ruppert, D. (1985). Transformations in regression: a robust analysis. *Technometrics*, **27**, 1–12.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York.

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.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, London.

Cox, D. R. (1977). Nonlinear models, residuals and transformations. *Mathematische Operationsforschung und Statistik, Serie Statistik*, **8**, 3–22.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, Series B*, **49**, 1–39.

Firth, D. (1988). Multiplicative errors: Log-normal or gamma? *Journal of the Royal Statistical Society, Series B*, **50**, 266–268.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, **10**, 507–521.

Foster, A. M., Tian, L., , and Wei, L. W. (2001). Estimation for the Box-Cox transformation model without assuming parametric error distribution. *Journal of the American Statistical Association*, **96**, 1097–1101.

Friedman, J. and Stuetzle, W. (1982). Smoothing of scatterplots. Technical report, Department of Statistics, Stanford University, Technical Report ORION 003.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York.

Gottardo, R. and Raftery, A. (2009). Bayesian robust transformation and variable selection: a unif ed approach. *Canadian Journal of Statistics*, **37**, 361–380.

Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods. *Bulletin of the International Statistical Institute*, **46**, 375–382.

Harrell Jr, F. E. (2019). *Hmisc*. R package version 4.2-0.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data. *Journal of the American Statistical Association*, **79**, 302–309.

Hoyle, M. H. (1973). Transformations - an introduction and a bibliography. *International Statistical Review*, **41**, 203–223.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, **1**, 799–821.

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

Kleinbaum, D. G. and Kupper, L. (1978). *Applied Regression Analysis and Other Multivariable Methods*. Duxbury, Boston, Mass.

Lawrance, A. J. (1987). The score statistic for regression transformation. *Biometrika*, **74**, 275–279.

Marazzi, A., Villar, A. J., and Yohai, V. J. (2009). Robust response transformations based on optimal prediction. *Journal of the American Statistical Association*, **104**, 360–370. DOI: 10.1198/jasa.2009.0109.

McCullagh, P. (2002a). Comment on "Box-Cox transformations in linear models: large sample theory and tests of normality" by Chen, Lockhart and Stephens. *The Canadian Journal of Statistics*, **30**, 212–213.

McCullagh, P. (2002b). What is a statistical model? *The Annals of Statistics*, **30**, 1225—-1310.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (2nd Ed.)*. Chapman and Hall, London.

Nychka, D. and Ruppert, D. (1995). Nonparametric transformations for both sides of a regression model. *Journal of the Royal Statistical Society, Series B*, **57**, 519–532.

Pericchi, L. R. (1981). A Bayesian approach to transformations to normality. *Biometrika*, **68**, 35–43.

Perrotta, D., Riani, M., and Torti, F. (2009). New robust dynamic plots for regression mixture detection. *Advances in Data Analysis and Classif cation*, **3**, 263–279. doi:10.1007/s11634-009-0050-y.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics*, **29**, 15–23,14.

Proietti, T. and Riani, M. (2009). Transformations and seasonal adjustment. *Journal of Time Series Analysis*, **30**(1), 47–69.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, **3**, 425–461.

Reid, N. (2002). Comment on "Box-Cox transformations in linear models: large sample theory and tests of normality" by Chen, Lockhart and Stephens. *The Canadian Journal of Statistics*, **30**, 211.

Riani, M. (2009). Robust transformations in univariate and multivariate time series. *Econometric Reviews*, **28**, 262 – 278.

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.

Riani, M., Cerioli, A., Atkinson, A. C., and Perrotta, D. (2014). Monitoring robust regression. *Electronic Journal of Statistics*, **8**, 642–673.

Ross, G. J. S. (1990). *Nonlinear Estimation*. Springer–Verlag, New York.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871–880.

Royston, P. J. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics*, **43**, 429–467.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The Statistician*, **41**, 169–178.

Seo, H. S. (2019). A robust method for response variable transformations using dynamic plots. *Communications for Statistical Applications and Methods*, **26**, 463–471. DOI: https://doi.org/10.29220/CSAM.2019.26.5.463.

Song, X.-Y. and Lu, Z.-H. (2012). Semiparametric transformation models with Bayesian P-splines. *Statistics and Computing*, **22**, 1085–1098.

Spector, P., Friedman, J., Tibshirani, R., Lumley, T., Garbett, S., and Baron, J. (2016). *acepack: ACE and AVAS for Selecting Multiple Regression Transformations*. R package version 1.4.1.

Sweeting, T. J. (1984). On the choice of prior distribution for the Box-Cox transformed linear model. *Biometrika*, **71**, 127–134.

Taylor, J. M. G. (1986). The retransformed mean after a f tted power transformation. *Journal of the American Statistical Association*, **81**, 114–118.

Taylor, J. M. G. (2004). Transformations II. *Wiley Statsref*.

Taylor, J. M. G. and Liu, N. (2007). Statistical issues involved with extending standard models. In V. Nair, editor, *Advances in Statistical Modeling and Inference*, pages 299–311. World Scientif c, Singapore.

Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, **83**, 394–405.

Velilla, S. (1995). Diagnostics and robust estimation in multivariate data transformations. *Journal of the American Statistical Association*, **90**, 945–951.

Wang, D. and Murphy, M. (2005). Identifying nonlinear relationships in regression using the ACE algorithm. *Journal of Applied Statistics*, **32**, 243 – 258.

Wang, N. and Ruppert, D. (1995). Nonparametric estimation of the transformation in the transform-both-sides regression model. *Journal of the American Statistical Association*, **90**, 522–534.

Weisberg, S. (2005). Yeo-Johnson power transformations. https://www.stat.umn.edu/arc/yjpower.pdf.

Wiens, B. L. (1999). When log-normal and gamma models give different results: a case study. *The American Statistician*, **53**, 89–93.

Yang, Z. (2006). A modif ed family of power transformations. *Economics Letters*, **92**, 14–19. https://ink.library.smu.edu.sg/soe-research/179.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959.

Yohai, V. J. (1987). High breakdown-point and high eff ciency estimates for regression. *The Annals of Statistics*, **15**, 642–656.

Zhang, T. and Yang, B. (2017). Box-Cox transformation in big data. *Technometrics*, **59**(2), 189–201.