Robust Response Transformations for Generalized Additive Models via Additivity and Variance Stabilisation

Marco Riani[1], Anthony C. Atkinson[2], and Aldo Corbellini[3]

[1] Dipartimento di Scienze Economiche e Aziendale and Interdepartmental Centre for Robust Statistics, Università di Parma, 43100 Parma
mriani@unipr.it,
WWW home page: http://www.riani.it
[2] The London School of Economics, London WC2A 2AE, UK,
[3] Dipartimento di Scienze Economiche e Aziendale and Interdepartmental Centre for Robust Statistics, Università di Parma, 43100 Parma
aldo.corbellini@unipr.it

Abstract. The AVAS (Additivity And Variance Stabilization) algorithm of Tibshirani provides a non-parametric transformation of the response in a linear model to approximately constant variance. It is thus a generalization of the much used Box-Cox transformation. However, AVAS is not robust. Outliers can have a major effect on the estimated transformations both of the response and of the transformed explanatory variables in the Generalized Additive Model (GAM). We describe and illustrate robust methods for the non-parametric transformation of the response and for estimation of the terms in the model and report the results of a simulation study comparing our robust procedure with AVAS. We illustrate the efficacy of our procedure through a simulation study and the analysis of real data.

Keywords: augmented star plot, AVAS, backfitting, forward search, heatmap, outlier detection, robust regression

1 Introduction

The nonlinear parametric transformations of response variables is a common practice in regression problems, for example logarithms of survival times. Tibshirani (1988) used smoothing techniques to provide non-parametric transformations of the response together with transformations of the explanatory variables, a procedure he called AVAS (additivity and variance stabilization). The resulting model is a generalized additive model (GAM) with a response transformed to approximate constant variance. Tibshirani's work can be seen as a non-parametric extension of the power transformation family of Box and Cox (1964) in which the goals are the stabilization of error variance and the approximate normalization of the error distribution, hopefully combined with an additive model. It also extends the parametric transformation of explanatory variables of Box and Tidwell (1962). A discussion of the relationship of AVAS to the Box-Cox transformation is in Hastie and Tibshirani (1990, Cap.7).

Tibshirani's AVAS is not robust with respect to outliers. Our main purpose is to provide a robust version of his work, which, for obvious reasons, we call RAVAS. In developing our procedure we made four important improvements to the original AVAS. Like robustness, these are available as options. Thus, RAVAS can be used for fitting a response transformed GAM when robustness is not an issue, or for fitting a GAM without response transformation.

Section 2 introduces the generalized additive model and the associated backfitting algorithm for estimation of the transformations of the explanatory variables, which uses a smoothing algorithm. The AVAS procedure and the associated numerical variance stabilization transformation are described in §§2.3 and 2.4. Section 3 outlines the various forms of robust regression that are available in our algorithm and describes the resulting outlier detection procedures. The purpose is to provide an outlier free subset of the data for transformation and smoothing. An outline of our improvements to AVAS is in §4. Appreciably more detail of these is provided in Riani *et al.* (2023) as well as further data analyses. Section 5 presents the results of a simulation study comparing some properties of AVAS and RAVAS in the presence of outliers: the mean squared error of parameter estimates, the power of detection of outliers, (just for RAVAS) and the number of numerical iterations of the two algorithms required for convergence. The performance of AVAS is severely degraded by the presence of outliers. The last two sections present a data analysis, which makes use of the augmented star plot as a guide to the choice of options in the estimation process and includes a comparison of the choices using a heatmap of correlations. The purpose of the paper is both to introduce the MATLAB program we have written for this form of robust data analysis and to illustrate some of its properties.

2 Generalized Additive Models and the Structure of AVAS

2.1 Introduction

The generalized additive model (GAM) has the form

$$g(Y_i) = \beta_0 + \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon. \tag{1}$$

The functions f_j are unknown and are, in general, found by the use of smoothing techniques. A monotonicity constraint can be applied. If the response transformation or link function g is unknown, it is restricted to be monotonic, but scaled to satisfy the technically necessary constraint that $\text{var}\{g(Y)\} = 1$. In the fitting algorithm the transformed responses are scaled to have mean zero; the constant β_0 can therefore be ignored. The observational errors are assumed to be independent and additive with constant variance. The performance of fitted models is compared by use of the coefficient of determination R^2. Since the f_j are estimated from the data, the traditional assumption of linearity in the explanatory variables is avoided. However, the GAM retains the assumption that explanatory variable effects are additive. Buja *et al.* (1989) describe the background and early development of this model.

2.2 Backfitting

For the moment we assume that the response transformation $g(Y)$ is known. The backfitting algorithm, described in Hastie and Tibshirani (1990, p.91), is used to fit a GAM. The algorithm proceeds iteratively using residuals when one explanatory variable in turn is dropped from the model.

With $g(y)$ the $n \times 1$ vector of transformed responses, let $e_{(j)}$ be the vector of residuals when $f_j(x_j)$ is removed from the model without any refitting. Then

$$e_{(j)} = g(y) - \sum_{k \neq j=1}^{p} f_k(x_k). \tag{2}$$

The new value of $f_j(.)$ depends on ordered values of $e_{(j)}$ and x_j. Let the ordered values of x_j be $x_{s,j}$. The residuals $e_{(j)}$ are sorted in the same way to give the new order $e_{s,(j)}$. Within each iteration each explanatory variable is dropped in turn; $j = 1, \ldots, p$. The iterations continue until the change in the value of R^2 is less than a specified tolerance.

For iteration l the vector of sorted residuals for x_j is $e^l_{(j)}$. The new estimate of $f_j^{(l+1)}$ is

$$f_{s,j}^{(l+1)} = S\left\{e^l_{s,(j)}, x_{s,j}\right\}. \tag{3}$$

The function S depends on the constraint imposed on the transformation of variable j. If the transformation can be non-monotonic, S denotes a smoothing procedure. As does Tibshirani (1988), we use the supersmoother (Friedman and Stuetzle, 1982), a nonparametric estimator based on local linear regression with adaptive bandwidths. Monotonic transformations using isotonic regression are also an optional possibility (Barlow *et al.*, 1972).

The backfitting algorithm is not invariant to the permutation of order of the variables inside matrix X, with high collinearity between the explanatory variables causing slow convergence of the algorithm: the residual sum of squares can change very little between iterations. Our option orderR2, §4.1, attempts a solution to this problem by reordering the variables in order of importance.

2.3 The AVAS Algorithm

In this section we present the structure of the AVAS algorithm of Tibshirani (1988). The variance stabilising transformation used to estimate the response transformation is outlined in §2.4

Our RAVAS algorithm has a similar structure to that given here, made more elaborate by the requirements of robustness and the presence of options. In this description of the algorithm ty and tX are transformed values of y and X.

1. *Initialise Data.* Standardize response y so that $\overline{ty} = 0$ and var$(ty) = 1$, where var is the maximum likelihood biased estimator of variance. Centre each column of the X matrix so that $\overline{tX}_j = 0, j = 1, \ldots, p)$.

2. *Initial call to 'Inner Loop'* to find initial GAM using ty and tX; calculates initial value of the coefficient of determination, R^2. Set convergence conditions on number of iterations and value of R^2.
3. *Main (Outer) Loop.* Given values of ty and tX at each iteration the outer loop finds numerically updated values of the transformed response. Given the newly transformed response, updated transformed explanatory variables are found through the call to the backfitting algorithm (*inner loop*). In our version iterations continue until a stopping condition on R^2 is verified or until a maximum number of iterations has been reached.

2.4 The Numerical Variance Stabilizing Transformation

We first consider the case of a random variable Y with known distribution for which $\mathrm{E}(Y) = \mu$ and $\mathrm{var}(Y) = V(\mu)$. We seek a transformation $ty = h(y)$ for which the variance is, at least approximately, independent of the mean. Then Taylor series expansion of $h(y)$ leads to $\mathrm{var}(Y) \approx V(\mu)\{h'(\mu)\}^2$. For a general distribution $h(y)$ is then a solution of the differential equation $\mathrm{d}g/\mathrm{d}\mu = C/\sqrt{V(\mu)}$. For random variables standardized, as are the values ty, to have unit variance, $C = 1$ the variance stabilizing transformation is

$$h(t) = \int^t 1/\sqrt{V(u)}du. \tag{4}$$

In the AVAS algorithm for data, $1/\sqrt{V(u)}$ is estimated by the vector of the reciprocals of the absolute values of the smoothed residuals sorted using the ordering based on fitted values of the model. There are n integrals, one for each observation. The range of integration for observation i goes from the smallest fitted value, to the old transformed value $\widehat{ty}_i, i = 1, \ldots, n$. The computation of the n integrals uses the trapezoidal rule and is outlined in subsection 4.2. Since the transformation is the sum of an increasing number of non-negative elements, monotonicity is assured. The logged residuals in the estimation of the variance function are smoothed using the running line smoother of Hastie and Tibshirani (1986).

3 Robustness and Outlier Detection

3.1 Robust Regression

We robustify our transformation method through the use of robust regression to replace least squares. The examples in this paper have been calculated using *Adaptive Hard Trimming.* In the Forward Search (FS), the observations are hard trimmed, the amount of trimming being determined by the choice of the trimming parameter h, the value of which is found adaptively by the search. Atkinson *et al.*, 2010 provide a general survey of the FS, with discussion. We have also implemented *Least Trimmed Squares,* Hampel, 1975, Rousseeuw, 1984 as well as *Soft trimming (downweighting).* Specifically we include S and MM estimation.

3.2 Robust Outlier Detection

Our algorithm works with k observations treated as outliers, providing the subset S_m of $m = n - k$ observations used in model fitting and parameter estimation. This section describes our outlier detection methods.

The default setting of the forward search uses the multivariate procedure of Riani *et al.* (2009) adapted for regression (Torti *et al.*, 2021) to detect outliers at a simultaneous level of approximately 1% for samples of size up to around 1,000. Optionally, a different level can be selected. For the other two methods of robust regression we apply a Bonferroni inequality to robust residuals to give a simultaneous test for outliers.

Since different response transformations can indicate different observations as outliers, the identification of outliers occurs repeatedly during our robust algorithm, once per iteration of the outer loop.

4 Improvements and Options

Our RAVAS procedure introduces five improvements to AVAS, programmed as options. These do not have a hierarchical structure, so that there are 2^5 possible choices of the options. The augmented star plot of §6 provides a method for assessing these choices. We discuss the motivation and implementation for each. The order in §4.1 is that in which the options are applied to the data when all five are used. We also give the names of the options, which are used as labels in the augmented star plot.

4.1 Initial Calculations

The structure of our algorithm is an elaboration of that of AVAS outlined in §2.3. Four of the five options can be invoked before the start of the outer loop.

Initialisation of Data: Option tyinitial Our numerical experience is that it is often beneficial to start from a parametric transformation of the response. This is optionally found using the automatic robust procedure for power transformations described by Riani *et al.* (2022). For $\min(y) > 0$ we use the Box-Cox transformation. For $\min(y) \leq 0$ the extended Yeo-Johnson transformation is used (Atkinson *et al.*, 2020). This family of transformations has separate Box and Cox transformations for positive and negative observations. In both cases the initial parametric transformations are only useful approximations, found by searching over a coarse grid of parameter values. The final non-parametric transformations sometimes suggest a generalization of the parametric ones.

Ordering Explanatory Variables in Backfitting: Option scail To avoid dependence of the fitted model on the order of the explanatory variables, one approach is to use an initial regression to remove the effect through scaling (Breiman, 1988). With b_j the coefficient of $f_j(x)$ in the multiple regression of $g(y)$ on all $f_j(x)$, the option scail provides new transformed values for the explanatory variables: $\widehat{tX}_j = b_j f_j(x)$, $j = 1, \ldots, p$. Option scail is used only in the initialisation of the data.

Robust Regression and Robust Outlier Detection: Option rob We robustify our method through the use of robust regression as described in §3. The subset S_m, changing at each iteration, defines the observations used in backfitting and in the calculation of the variance stabilising transformation.

Ordering Predictor Variables: Option orderR2 For complete elimination of dependence on the order of the variables, we include an option that, at each iteration, provides an ordering which is based on the variable which produces the highest increment of R^2. With this option the most relevant features are immediately transformed and those that are perhaps irrelevant will be transformed in the final positions. For robust estimation, this procedure is applied solely to the observations in the subset S_m. Option orderR2 is available at each call to the backfitting function.

4.2 Outer Loop

Numerical Variance Stabilising Transformation: Option trapezoid Plots of residuals against fitted values are widely used in regression analysis to check the assumption of constant variance. Here the observations have been transformed, so the fitted values are $\widehat{ty}_i$. To estimate the variance stabilizing transformation, the fitted values have to be sorted, giving a vector of ordered values $\widehat{ty}^s$. The residuals are ordered in the same way and, following the procedure of §2.4, provide estimates v_i of the integrand $V^{-0.5}(y)$ in (4). The v_i provide estimates at the ordered points $\widehat{ty}^s$. Calculation of the variance transformation (4) is however for sorted observed responses ty_i^s, rather than fitted, transformed responses $\widehat{ty}^s$. As did Tibshirani, we use the trapezoidal rule to approximate the integral. Linear interpolation and extrapolation are used in calculation of the v_i at the ty_i^s. We provide an option 'trapezoid' for the choice between two methods for the extrapolation of the variance function estimate, the interpolation method remaining unchanged. Our approach leads to trapezoidal summands in the approximation to the integral for the extrapolated elements, whereas Tibshirani's leads to rectangular elements. When we are concerned with robust inference, there are only $m = n - k$ members of $\widehat{ty}^s$ whereas there are n values of ty_i^s, so that robustness increases the effect of the difference between the two rules. The option trapezoid = false uses rectangular elements in extrapolation.

5 Simulations

We now use simulations to compare overall properties of AVAS with our robust version. The model was linear regression with data generated to have an average value of R^2 of 0.8. The responses were standardized to have zero mean and unit variance; 10% of the observations were contaminated by a shift δ and the responses were exponentiated. There were 1,000 simulations for $n = 200$ and $n = 1,000$ and 200 for $n = 10,000$. We encountered no numerical problems in the simulations. The figures compares the performances of RAVAS (with all options) and standard AVAS (with no options). Results for RAVAS use a dashed line.

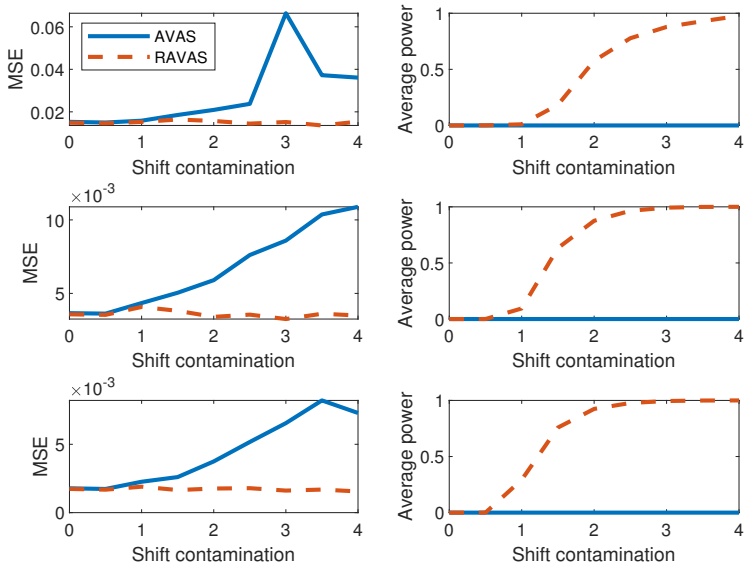

Fig. 1. Mean squared error (MSE) and average power. Top panels $n = 200$, $p = 5$, mid panels $n = 1,000$, $p = 10$, bottom panels $n = 10,000$, $p = 20$

The left-hand panels of Figure 1 show the mean squared error of the parameter estimates in the linear models. For RAVAS, those for $n = 200$ and 1,000 exhibit a slight increase for moderately small values of δ which then decreases to be close to zero as δ increases and the outliers become easier to detect. That for $n = 10,000$ is virtually constant. The results for AVAS rapidly become much larger. The right-hand column shows the average power, that is the proportion of generated outliers that are detected by RAVAS. This climbs, in all cases, steadily to one. Of course, AVAS does not detect outliers.

We also compared the number of iterations to convergence of the algorithms; the default maximum is 20. Figure 2 shows results for the same simulations as above. The three panels show that RAVAS converges in around 3 iterations, except for $n = 200$ when there is a peak around $\delta = 2$, that is when the outliers are large enough to have an effect, but are still difficult to detect. This behaviour is distinct from that of AVAS, where the number of iterations increases steadily both with δ and with the sample size.

6 The Generalized Star Plot

We have added five options to the original AVAS. There are therefore 32 combinations of options that could be chosen. It is not obvious that all will be necessary when analysing any particular set of data. Our program provides flexibility in the assessment of these options. One possibility is a list of options ordered by, for example, the value of R^2 or of the significance of the Durbin-Watson test. In this section we describe the

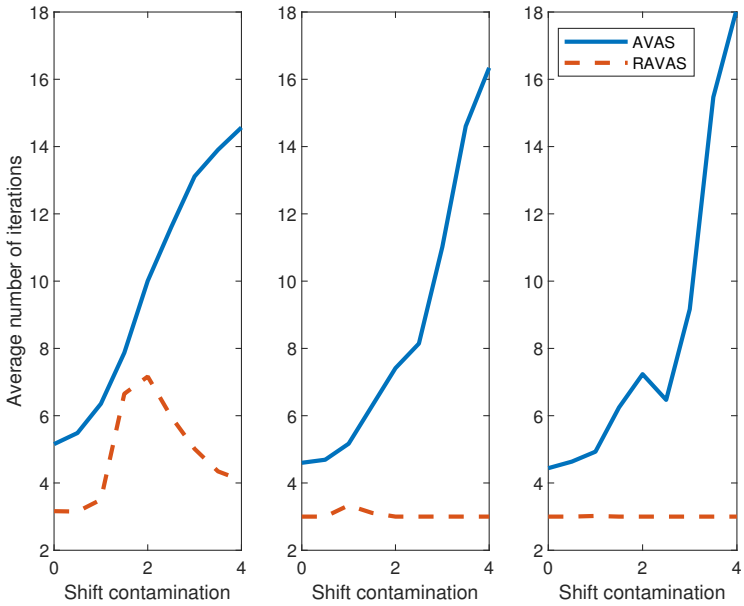

Fig. 2. Average number of iterations to convergence. Left-hand panel $n = 200$, $p = 5$, central panel $n = 1,000$, $p = 10$, right-hand panel $n = 10,000$, $p = 20$

augmented star plot, one graphical method for visualizing interesting combinations of options in a particular data analysis. An example is Figure 3.

We remove all analyses for which the residuals fail the Durbin-Watson test of independence and the Jarque-Bera normality test (Jarque and Bera, 1987), at the 10 per cent level (two-sided for Durbin-Watson). The threshold of 10% can be optionally changed. We order the remaining, admissible, solutions by the Durbin-Watson significance level multiplied by the value of R^2 and by the number of units not declared as outliers. Other options are available. The lengths of rays in individual panels of the plot are of equal length for those features used in an analysis. All rays are in identical places in each panel of the plot; the length of the rays for each analysis are proportional to p_{DW}, the significance level of the Durbin-Watson test.

The ordering in which the five options are displayed in the plot depends on the frequency of their presence in the set of admissible solutions. For example, if robustness is the one which has the highest frequency, its ray is shown on the right. The remaining options are displayed counterclockwise, in order of frequency.

7 Prediction of the Weight of Fish

Two websites, https://www.kaggle.com/aungpyaeap/fish-market and http://jse.amstat.org/datasets/fishcatch.txt present data on the weight of 159 fish caught in a lake near Tampere, Finland. Interest is in the relationship between weight and five measurements of dimensions of the fish. There are 7 species of fish including pike. These behave

rather differently from the other six species so we ignore them. We use the first three lengths for which the remaining fish seem homogenous. This assumption will be tested by our robust analysis if one or more species are identified as outliers. The variables are:

y Weight of the fish (in grams)
x_1 Length from the nose to the beginning of the tail (in cm)
x_2 Length from the nose to the notch of the tail (in cm)
x_3 Length from the nose to the end of the tail (in cm).

After the deletion of the data on pike, 142 observations remain. Scatter plots of the response against the three explanatory variables reveal that all three lengths are highly correlated with the response, as they are with each other. It is reasonable to assume that weight increases with each of the explanatory variables. We therefore impose a monotonicity constraint on the transformations of the x_j. However, multiple regression with highly correlated explanatory variables can lead to problems in interpretation, such as estimated effects having a physically incorrect sign.

The augmented star plot for these data is in Figure 3. There are six combinations of options that satisfy the constraints on the distribution of residuals. The first solution, with an R^2 of 0.991 uses all five options except trapezoid. Robustness is used in all, succeeding selections giving R^2 values of 0.988 or 0.983.

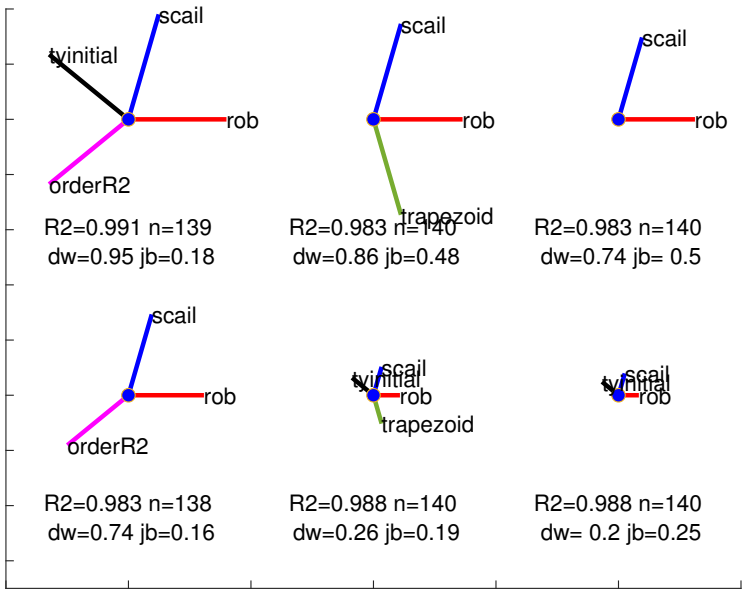

Fig. 3. Weight of fish. Augmented star plot of six options. Option 1 excludes trapezoid

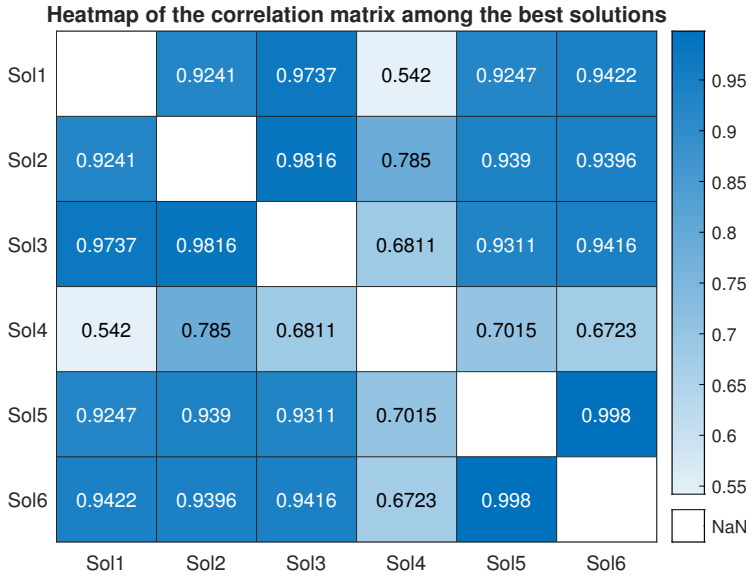

Fig. 4. Weight of fish. Heatmap of pairwise response correlations among the six solutions

The heatmap of the response correlations between the pairs of solutions is in Figure 4. This shows that the first three solutions are strongly correlated with each other, as they are with the fifth and sixth solutions, the fifth and sixth solutions themselves having a very high correlation of 0.998. The heatmap emphasizes that solution four is appreciably different from the other five.

We now consider the adaptive identification of outliers using the FS. The first solution identifies three outliers. The left-hand panel of Figure 5 shows that the response has been smoothly transformed. The plot of residuals against fitted values in the right-hand panel shows that there is only one remote outlier and that there is no remaining structure in the residuals. The plots of transformed explanatory variables (not given here) show that $f(x_1)$ is decreasing and slightly curved. The other two functions are increasing but only that for x_2 is almost straight, with slight curvature for the lowest values of the variable.

The interpretation of the results from fitting three explanatory variables is that the variables are too highly correlated to give individually meaningful results. In our final analysis of the data we used only x_1. The star plot showed that the best selection included all options, except orderR2, which option is not possible with a single explanatory variable. The value of R^2 for this fit is 0.980 with the deletion of a single outlier. The three acceptable solutions had mutual fitted response correlations of 0.9994 or 1 - the fitted model was stable to the choice of options.

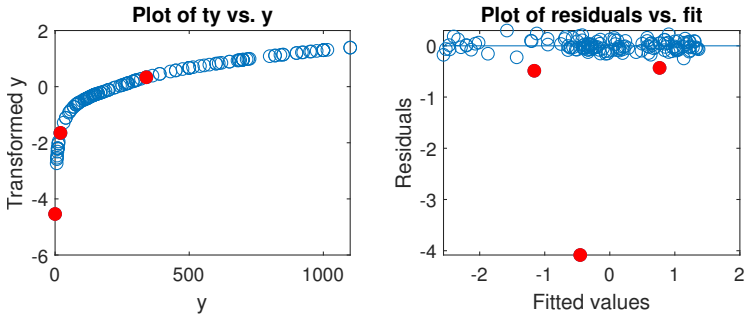

Fig. 5. Weight of fish. Left-hand panel, transformed y against y; right-hand panel, residuals against fitted values. Three outliers in red in the online version

In regressing volume on measurements of length, arguments from dimensional analysis suggest that volume should have a one third transformation. Our final plot, Figure 6, compares the transformed responses from the fits with three and one explanatory variables to $y^{1/3}$, for which transformation the value of $R^2 = 0.968$. The figure shows that both non-parametric transformations are indeed close to $y^{1/3}$ with a small systematic departure for the largest values of x. The fitted values from the single explanatory variable follow the power transformation slightly more closely than that when three variables are used. The transformation of x_1 is virtually straight with some curvature for large values. The flexibility of the non-parametric transformation provides an improved simple model compared with regression on untransformed x_1.

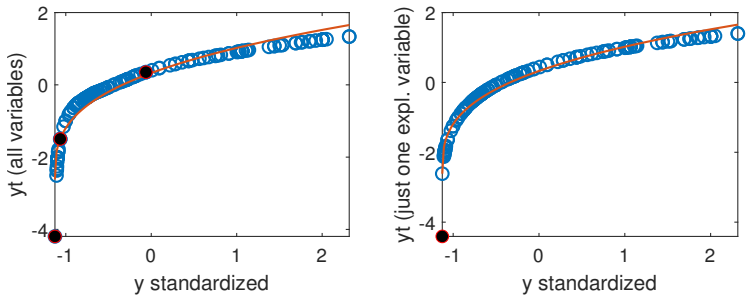

Fig. 6. Weight of fish. Non-parametric transformation of response y compared to $y^{1/3}$. Left-hand panel, three explanatory variables: right-hand panel, only x_1. Three and one outliers in black and red in the online version

Bibliography

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 analysis of transformations for profit-and-loss data. *Applied Statistics*, **69**, 251–275. DOI: https://doi.org/10.1111/rssc.12389.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. 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 Tidwell, P. W. (1962). Transformations of the independent variables. *Technometrics*, **4**, 531–550.

Breiman, L. (1988). Comment on "Monotone regression splines in action" (Ramsey, 1988). *Statistical Science*, **3**, 442–445.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *Annals of Statistics*, **17**, 453–510.

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

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

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, **1**, 297–318.

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

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, **52**, 163–172.

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., Atkinson, A. C., and Corbellini, A. (2022). Automatic robust Box-Cox and extended Yeo-Johnson transformations in regression. *Statistical Methods and Applications*. https://doi.org/10.1007/s10260-022-00640-7.

Riani, M., Atkinson, A. C., and Corbellini, A. (2023). Robust transformations for multiple regression via additivity and variance stabilization. *Journal of Computational and Graphical Statistics*. (Revision under review).

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

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

Torti, F., Corbellini, A., and Atkinson, A. C. (2021). fsdaSAS: a package for robust regression for very large datasets including the Batch Forward Search. *Stats*, **4**, 327–347.