stats

MDPI

Article

The Use of Modern Robust Regression Analysis with Graphics: An Example from Marketing

Marco Riani [1,*], Anthony C. Atkinson [2], Gianluca Morelli [1] and Aldo Corbellini [1]

[1] Dipartimento di Scienze Economiche e Aziendali and Interdepartmental Centre for Robust Statistics, Università di Parma, 43100 Parma, Italy; gianluca.morelli@unipr.it (G.M.); aldo.corbellini@unipr.it (A.C.)
[2] The London School of Economics, London WC2A 2AE, UK; a.c.atkinson@lse.ac.uk
[*] Correspondence: mriani@unipr.it; Tel.: +39-0521-902473

Abstract: Routine least squares regression analyses may sometimes miss important aspects of data. To exemplify this point we analyse a set of 1171 observations from a questionnaire intended to illuminate the relationship between customer loyalty and perceptions of such factors as price and community outreach. Our analysis makes much use of graphics and data monitoring to provide a paradigmatic example of the use of modern robust statistical tools based on graphical interaction with data. We start with regression. We perform such an analysis and find significant regression on all factors. However, a variety of plots show that there are some unexplained features, which are not eliminated by response transformation. Accordingly, we turn to robust analyses, intended to give answers unaffected by the presence of data contamination. A robust analysis using a non-parametric model leads to the increased significance of transformations of the explanatory variables. These transformations provide improved insight into consumer behaviour. We provide suggestions for a structured approach to modern robust regression and give links to the software used for our data analyses.

Keywords: AVAS; Box–Cox transformation; brushing; forward search; generalized additive model (GAM); linked plots

check for updates

Academic Editor: Wei Zhu

Received: 22 November 2024
Revised: 25 December 2024
Accepted: 31 December 2024
Published: 8 January 2025

Citation: Riani, M.; Atkinson, A.C.; Morelli, G.; Corbellini, A. The Use of Modern Robust Regression Analysis with Graphics: An Example from Marketing. *Stats* **2025**, *8*, 6. https://doi.org/10.3390/stats8010006

1. Introduction

Routine regression analysis resulting in analysis of variance tables as a summary of the data may sometimes miss important aspects of data. We illustrate this assertion through the analysis of a set of 1711 observations on the determinants inspiring loyalty in consumers. The data are suitable for regression analysis. We start with a conventional least squares analysis in which all explanatory variables are highly significant and could lead straightforwardly to building a regression model. However we do not stop there, but provide a series of more advanced analyses based on the dynamic methods of the forward search introduced in Section 2. These show a failure of the simple fitted regression model to describe all features of the data. We demonstrate a series of analyses using generalizations of the linear model of regression which lead to an informative model for the data. Many of these extensions illustrate the use of robustness in data analysis. These methods are enhanced by the use of a variety of graphical methods, mostly derived from standard statistical estimates and tests. An important extension of the models considered is to non-parametric regression. The purpose is to provide a paradigmatic example of regression analysis using these powerful tools which can outperform the standard methods of analysis. We suggest that they be used as routine in any regression analysis, providing a checklist to ensure adequate data analyses. We want to move beyond the tables of p-values of the significance of variables that, all too often, are presented, without graphical output, as a satisfactory analysis of regression data.

We begin with some history of regression, from Legendre and Gauss and least squares to dynamic robust regression. The loyalty data are introduced in Section 3 and subjected to a least squares analysis. Plots of the residuals indicate departures from normality. The dynamic plot of residuals in the original scale, which monitors the behaviour of residuals with an increase in the size of the subset of observations S_m used to fit the least squares model to the data, indicates some potential outliers. In Section 4, we monitor the added variable plot of the t-statistics for parameter estimates, which shows that the outliers have a negligible effect on the estimated parameters. In Section 5, we explore the transformation of the response variable using the fan plot, an extension of the monitored added variable plot, to monitor the transformation of the data. The procedure indicates the square root transformation for the data and indicates 41 outliers. The residuals from fitting the square root transformed response continue to show departures from normality. Section 6 introduces the AVAS procedure of Tibshirani [1] for transforming both the explanatory variables and the response in a regression model. Explanatory variables are transformed using smoothing algorithms and the response by a numerical procedure, using functions of residuals, which seeks a response transformation to constant error variance. The response transformation is thus a non-parametric extension of the Box–Cox transformation. The results of this fitting procedure are encouraging, giving an improved value of 0.790 for the adjusted R^2. However, there is still some evidence of outliers. We conclude our analysis in Section 7, analysing the data with RAVAS [2], a robust extension of AVAS, to provide robustness against outliers. This robust analysis leads to an even better fitting model, with 14 clearly defined residuals and sharp conclusions about the significance of the variables. The paper continues in Section 8 with a marketing-oriented interpretation of the forms of the transformed explanatory variables. Section 9 uses the evidence of our data analyses to suggest a structured approach to robust regression analysis, with recommendations for the extension of our methods to other problems in Section 10. Links to the publicly available software used in this paper are given. In the data analytical sections of the paper we first introduce the technique employed and then apply it to the loyalty data. Further details of all robust methods and examples of other data analyses are in Atkinson et al. [3].

2. Some History

A little history may be helpful. The first clear and concise exposition of the method of least squares was published by Legendre in 1805. In 1809, Carl Friedrich Gauss published his method (meine Methode) of calculating the orbits of celestial bodies. He claimed to have been in possession of the method of least squares since 1795. This naturally led to a priority dispute with Legendre. However, Gauss went beyond Legendre and succeeded in connecting the method of least squares with the principles of probability and to the normal distribution. The technique is described as an algebraic procedure for fitting linear equations to data.

The first statistical advance was in Gosset's derivation of Student's t distribution for testing the mean of a normal sample [4]. Although Gosset's derivation was correct, he was unable to provide a proof of his result. In 1915, Fisher sent Gosset a rigorous proof of his derivation, but the extension to testing regression coefficients was not clarified for another ten years. The result was first publicized in Fisher's famous *Statistical Methods for Research Workers* [5]. The history of this interaction between Gosset and Fisher is given by Lehmann [6]. The ability to test the significance of the individual regression coefficients is a useful first step in model building. However, especially with correlated explanatory variables, the deletion of a single variable may cause large changes in the significance of the remaining variables; strategies for the choice of which variables to include are required. Several strategies, including forward selection, backwards elimination, their combination in stagewise methods and the consideration of all possible regressions are described, for example, in Draper and Smith [7] (Chapter 6).

Given some numbers, the procedure of the previous paragraph for model choice will select a "best" model. However, the data may contain outliers and the model may be inadequate. It is therefore imperative to examine the agreement between the fitted model and the data. One option is a set of procedures often called "regression diagnostics". These grew from the examination and analysis of plots of residuals [8]. It is noteworthy that Anscombe's paper contains only two such plots. With the development of computer power and graphics, it became possible to fit many models to the same dataset, to assess the effect of deleting groups of observations and to represent the outcome in a variety of plots, some informative. The books in this area include Belsley et al. [9], Cook and Weisberg [10], Atkinson [11], and Cook and Weisberg [12]. Because these analyses start from a least squares (LS) fit to all the data, there is a combinatorial explosion of cases to consider, even if only triplets of observations are considered for deletion. In our analyses we use some of the graphical techniques developed in this literature.

The methods of robust regression complement and extend the diagnostic procedures of observational deletion and plotting. Unlike diagnostic methods, robustness starts from the realization that data rarely follow the simple models of mathematical statistics. In particular, there are often dispersed or grouped outliers which may be informative about the need for a more complex model. The distance between mathematical theory and data reality has led, over the last sixty years, to the development of a large body of work on robust statistics, intended to provide fitted models unaffected by the presence of outliers and other model failings. By the time of [13] (the Princeton Robustness Study), according to [14], it was expected that in the near future "any author of an applied article who did *not* use the robust alternative would be asked by the referee for an explanation". This "grand plan" failed, in part because there are not one but many robust alternatives, the properties of which are only proven asymptotically and so may provide a poor guide to the analysis of small samples of data. Here "small" may sometimes be several thousand. Now, a further fifty years on, a consensus seems to be emerging as to the correct route to an appropriate robust data analysis.

Robust regression analysis is typically calibrated against a normal distribution of errors with contamination. The asymptotic breakdown point bdp of an estimator is the proportion of observations that can be replaced by *any* outliers with the estimator remaining in a bounded set. For LS, $bdp = 0$. For problems in which it is assumed that the data are sampled from a single population, with added contamination, the maximum asymptotic value of bdp is 0.5, giving rise to what is colloquially known as "very robust regression". Values greater than 0.5 are not possible, since then the outliers become the majority distribution. However, for data arising from several populations, as is the case in clustering, values of bdp greater than 0.5 are meaningful [15]. Huber [16] introduced M-estimators, that is maximum likelihood-type estimators, in which the linear estimating equation for LS is replaced by a function that downweights extreme observations, with very extreme observations being effectively deleted. The parameters of the curve determine the value of bdp. Related estimators, such as S and MM, vary in the way in which the error variance σ^2 is estimated. See Maronna et al. [17] for details. Two other methods of robust regression are based directly on least squares. In least trimmed squares (LTS; Rousseeuw and Leroy [18] (Section 3.4)), LS estimation is applied to a subset of the observations of prespecified size h, which determines bdp. Numerical search is necessary to find the subset minimizing the residual sum of squares for all observations with a consistency correction [19] for the estimate of σ^2, since the fitting procedure selects a central subset of observations. In the forward search [20], LS is used to fit models to subsets S_m of size m, with m going from m_0 to n. The initial small subset S_{m0} is chosen robustly and the FS moves forward increasing the size of S_m one observation at a time. Estimation of σ^2 again requires a consistency correction. Since the model is refitted for each m, if more than one observation enters the subset in going from S_m to S_{m+1}, some observations must leave the subset. The effect is that the inclusion of several outliers can cause an abrupt change in the fitted model. During the search, the values of quantities of interest, such as parameter estimates and residuals, are monitored and outliers detected. Brushing of computer graphics provides links, for example, between patterns of residuals and scatterplots of the data.

The first two robust methods above each provide a single snapshot of the data at a specified value of *bdp*. It has long been advocated that a very robust fit be compared with a non-robust fit [18] (p. 111). Such comparisons are for two extreme forms of regression: the most and least robust. As it moves forward, the FS produces a series of such snapshots, conceptually forming a film or movie. The consensus in robust regression (and robust methods in general) comes from replacing the single snapshots from M and LTS estimation by a movie formed from calculations at many values of *bdp*. By monitoring robust regression in this way, we obtain information on the important changes in conclusions that come from differing assumptions about the degree of contamination in the data. In particular, as *bdp* decreases, there often comes a point in contaminated data when the residuals switch to being those from LS since they are calculated from biased parameter estimates. For higher values of *bdp*, the outliers are clearly differentiated from the remaining data. The lowest value of *bdp* for which the outliers are clear provides an empirical *bdp* for the particular analysis. Using a higher value of *bdp* produces unbiased parameter estimates, but of unnecessarily reduced efficiency, since some non-outlying observations will have been deleted. Monitoring regression models is described by Riani et al. [21]; several examples illustrate the estimation of the empirical *bdp*. Monitoring for multivariate analysis is in Cerioli et al. [22]. In the analyses in this paper, we mainly use the FS because empirical evidence indicates that it provides estimators with a higher value of empirical *bdp* than does monitoring other methods of robust regression.

3. Multiple Regression

3.1. Motivation

It is typical of retail customers to make purchases from multiple outlets. There are many reasons for this behaviour. A major goal of retailers is to reduce such fragmentation of spending through loyalty. Thus, loyalty is a marketing objective aimed at capturing customers for a considerable part of their expenditure. The attempt to build loyalty uses many marketing levers, such as the development of programs which stimulate customers with discounts, free goods, prizes, or special services [23]. Developing customer loyalty is seen as a form of protection from competition and gives more control over marketing planning [24]. From a theoretical point of view, the only suitable tool for a retailer to provide an objective measurement of customers' loyalty is the share of their wallet, that is the measurement of how customers divide their purchases among various competitors. It is self-evident that, for the retailer, the share of the wallet cannot be measured directly. Therefore loyalty must be estimated through other determinants. Two complications emerge. The first is an ambiguous definition, in the marketing literature, of the determinants that can potentially affect loyalty. The second is linked to the nature of some of the variables that can contribute to loyalty; very often, it is necessary to include variables expressed as ratings of customer perceptions. To investigate some determinants of loyalty, we analyse data from a questionnaire intended to illuminate the relationship between loyalty and perceptions of such factors as price and community outreach.

3.2. The Data

The data, representing the results of a consumer-loyalty questionnaire, come from the website https://data.world/cesarpolo/consumer-loyalty-in-retail (accessed on 22 November 2024). The dataset used was chosen because the information concerning it is limited solely to the names of the variables. On the web, there is no description of the scale of the variables, of any data pre-processing nor any other information. We took up this challenge because we are confident that our successful blind analysis of data of this type gives us hope for good results in the case of datasets whose structure is known. The dataset contains 1711 observations, all complete. The variables for a regression model are as follows:

y (response) Loyalty
x_1 Price
x_2 Quality
x_3 Community Outreach
x_4 Trust
x_5 Customer Satisfaction
x_6 Negative Publicity.

If regression data come from measuring a physical system, usually it is to be expected that the data are generated by a smooth underlying model. The Taylor series expansion of this unknown model would lead to a useful polynomial model. However, here, all variables are taken from the subjective responses to the questionnaire. Consequently, the variables may be highly nonlinear and may benefit from transformations to produce a simple regression model. The purpose of the analysis is to determine which of the variables above most affect loyalty and what is the sign of the effect. As the analysis progresses, we explore nonlinear transformations of both the response and the explanatory variables.

As in all data analyses, we should begin by plotting the data. Figure 1 is a yX plot of the data, that is, a side-by-side presentation of the scatterplots of the response against each explanatory variable. This is rather condensed. Figure 2 shows the scatterplot of loyalty against trust; there is a clear indication of some relationship between the two variables, as there also is in some of the other panels of Figure 1. Details of the correlation structure of the variables are in Table 1.

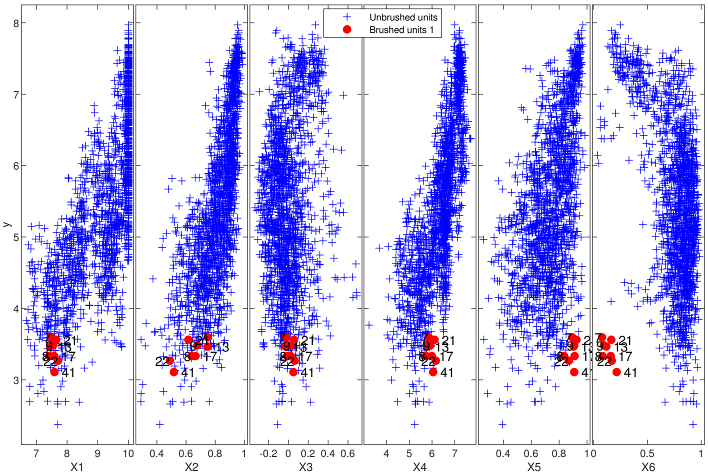

Figure 1. Original data: yX plot with the 8 brushed units from Figure 5 highlighted. The labelling of the observation numbers is produced automatically.

Table 1. Correlation index between loyalty and explanatory variables.

	Loyalty	**Price**	**Quality**	**Community Outreach**	**Trust**	**Customer Satisfaction**	**Negative Publicity**
Loyalty	1	0.727	0.713	0.182	0.755	0.524	−0.45
Price	0.727	1	0.686	−0.117	0.838	0.28	−0.193
Quality	0.713	0.686	1	0.055	0.617	0.41	−0.229
Community Outreach	0.182	−0.117	0.055	1	0.018	0.327	−0.288
Trust	0.755	0.838	0.617	0.018	1	0.384	−0.337
Customer Satisfaction	0.524	0.28	0.41	0.327	0.384	1	−0.488
Negative Publicity	−0.45	−0.193	−0.229	−0.288	−0.337	−0.488	1

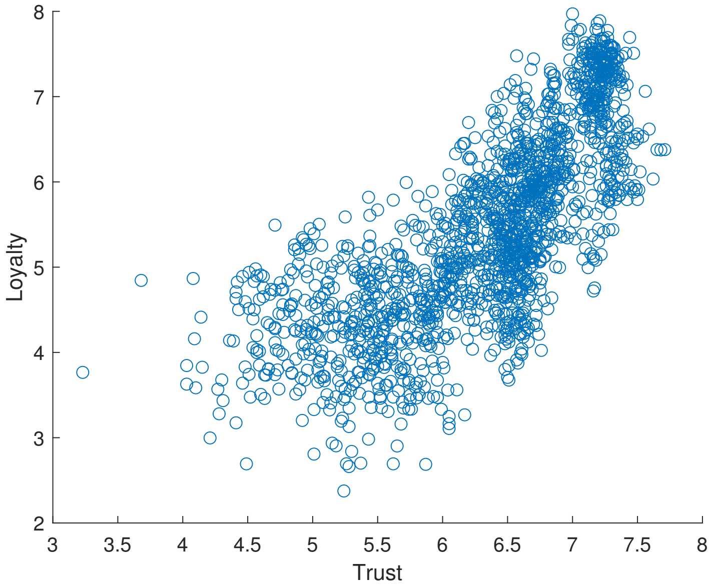

Figure 2. Original data: scatterplot of loyalty (y) against trust (x_4).

Accordingly, we start with fitting a linear regression model using LS.

3.3. The Model: Parametric Regression

In linear regression models there are n observations on a continuous response y. The expected value of the response $\mathrm{E}(Y)$ is related to the values of p known constants by the relationship $\mathrm{E}(Y) = X\beta$, with Y the $n \times 1$ vector of responses, X an $n \times p$ matrix of known constants and β a vector of p unknown parameters. The model for the ith of the n observations is $y_i = x_i^T \beta + \epsilon_i$. It is customary to assume that the errors ϵ_i have zero mean, constant variance σ^2 and are uncorrelated. Like any assumption in the analysis of data, this one should be checked; one indication of the need for a response transformation is a relationship between the variance of the errors and the magnitude of the response.

The least squares estimate $\hat{\beta}$ of the parameter vector β is $\hat{\beta} = (X^T X)^{-1} X^T y$, a linear combination of the observations, which will be normally distributed if the observations are. The vector of n predictions from the fitted model is $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$. H is often called the "hat" matrix. The ith least squares residual is $e_i = y_i - \hat{y}_i$. The vector of *least squares residuals* is $e = y - \hat{y} = y - X\hat{\beta} = (I - H)y$, whence the variance of an individual residual is var $e_i = (1 - h_i)\sigma^2$, with h_i the ith diagonal element of H. The value of h_i, the "leverage", is a measure of the remoteness of x_i in the space of X.

The residual sum of squares $S(\hat{\beta}) = \sum_{i=1}^n e_i^2 = y^T (I - H) y$ provides the residual mean square estimate of σ^2: $s^2 = S(\hat{\beta})/(n - p)$. Division of the least squares residuals by s gives the scaled residuals $r_i = e_i / s$. The values of these residuals do not depend on the value of

σ^2. But, like the least squares residuals, they do not in general have the same variance. The *Studentized residuals*, which do have equal variance, are defined as

$$\tilde{r}_i = e_i / s(1 - h_i)^{0.5} = y_i - \hat{y}_i / \{s(1 - h_i)^{0.5}\}. \tag{1}$$

The Studentized residuals, which are widely used in model checking, are not independent, nor do they follow a Student's t distribution. Cook and Weisberg [10] (p. 19) show that $r_i^2/(n-p)$ has a scaled beta distribution. But with a value of n as large as that for these data, the differences between the various residuals are slight and all can be taken as having a normal distribution with constant variance.

To test the individual terms of the model, we use t tests. The overall effect of the model can be assessed by the coefficient of determination R^2, with the results summarized in an analysis of variance table. If the total corrected sum of squares of the observations is $\sum_{i=1}^{n}(y_i - \bar{y})^2$, where $\bar{y} = \sum y_i/n$, the coefficient of determination is defined as $R^2 = \{S_0 - S(\hat{\beta})\}/S_0$. The *adjusted* R^2, written R^2_{adj}, allows for the degrees of freedom of the two sums of squares: $R^2_{\text{adj}} = 1 - (1 - R^2)\{(n-1)/(n-p)\}$. A value near one indicates that a large proportion of the total sum of squares of the observations has been explained by the regression. However, a large value, while encouraging, says nothing about the contribution of particular groups of observations to various aspects of the fit, such as the importance of specific parameters.

Calculation of the t tests requires the variances of the elements of $\hat{\beta}$. Since $\hat{\beta}$ is a linear function of the observations, var $\hat{\beta} = \sigma^2(X^T X)^{-1}$. Let the kth diagonal element of $(X^T X)^{-1}$ be v_k, when the t test for testing that $\beta_k = 0$ is

$$t_k = \hat{\beta}_k / (s_\nu^2 v_k)^{0.5}, \qquad k = 1, \ldots, p, \tag{2}$$

where $s_\nu^2 = S(\hat{\beta})/(n-p)$ is an estimate of σ^2 on $\nu = n - p$ degrees of freedom. If $\beta_k = 0,\ t_k$ has a t distribution on ν degrees of freedom. If the explanatory variables are correlated, dropping x_k from the model may cause appreciable change in the significance and even the signs of the remaining t statistics.

3.4. The Fitted Model and Its Residuals

The analysis of variance for regression on all variables is in Table 2. All variables are highly significant, with R^2_{adj} = 0.741. It is natural to ask whether, and if so how, we might do better. One possibility, given the curved nature of the relationships in Figure 1, is to try a model with all interactions. After using a stepwise procedure for model selection, a model is obtained with 17 terms and a value of adjusted R^2 value of 0.783. This is a small improvement in fit for such a cumbersome model. Instead, we look at plots of residuals and the fitted model to assess how well the data and fitted model agree.

The plot of residuals against fitted values is often a powerful diagnostic of omitted structure. However, in this case, it is uninformative. But other plots of residuals are highly revealing. The upper panel of Figure 3 shows a QQ (quantile–quantile) plot of Studentized residuals against normal quantiles, with a 95% pointwise envelope to provide a judgment of the straightness of the plot [11]. Although the extreme residuals are not extreme, smaller positive and negative residuals do lie outside the envelope. In particular, there are too many negative residuals that are too large to have come from a normal distribution. The lower panel of the figure is a symmetry plot of residuals about their median. For n even, let the ordered residuals be $e_{[i]}$ with the median residual $e_m = e_{[n/2]}$. Working out from the median, the points plot $e_{[m-i]}$ against $e_{[m+i]}$. (The procedure is the same for n odd, but the notation is, unhelpfully, more complicated.) The upper right-hand part of the plot exhibits, in an accentuated form, the large negative residuals lying outside the envelope in the plot of the upper panel.

Table 2. ANOVA in original scale for y.

	Estimate	SE	tStat	p Value
(Intercept)	-2.0312	0.18383	-11.05	1.8399×10^{-27}
Price	0.32801	0.031277	10.487	5.5983×10^{-25}
Quality	2.5894	0.16721	15.486	1.043×10^{-50}
Community Outreach	0.75773	0.095933	7.8986	5.016×10^{-15}
Customer Satisfaction	0.99442	0.12458	7.9822	2.6177×10^{-15}
Negative Publicity	-0.95476	0.089834	-10.628	1.3692×10^{-25}
Number of observations: 1711, error degrees of freedom: 1704				
Root Mean Squared Error: 0.578				
R-squared: 0.742, Adjusted R-Squared: 0.741				
F-statistic vs. constant model: 815, p-value = 0				

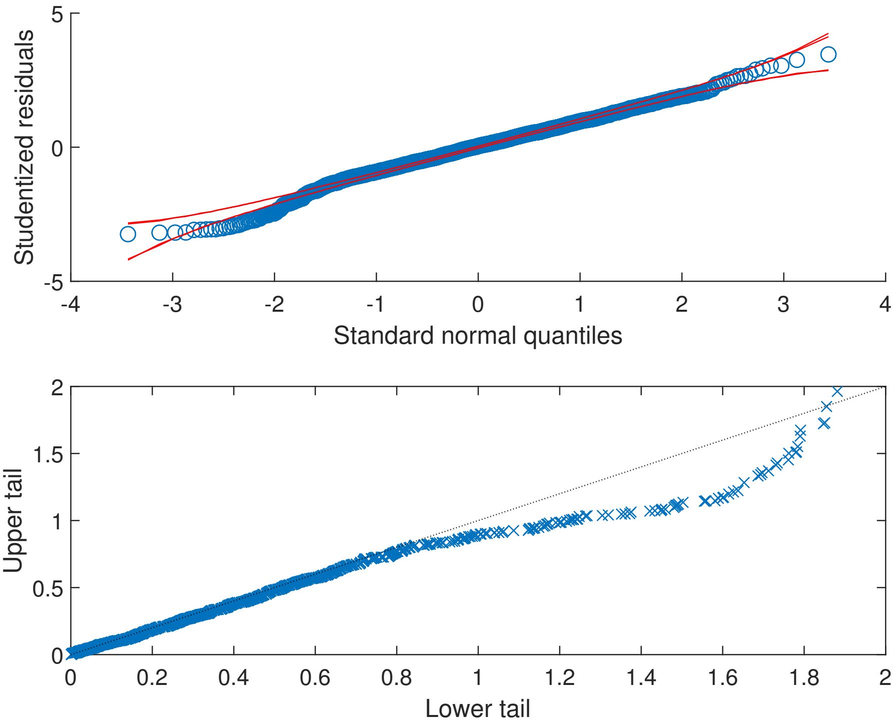

Figure 3. Original data: cumulative plots of residuals. **Upper** panel: QQ plots of Studentized residuals with 95% simulation envelope; **lower** panel: symmetry plot of Studentized residuals.

Figure 4 is a histogram of the residuals, which is slightly asymmetric. The superimposed normal distribution in the left-hand panel shows that there are too few residuals with values around minus one for the distribution to be normal. The kernel density estimate, in the right-hand panel, emphasizes the non-normal shape of this tail of the distribution.

Figures 3 and 4 show the anomalous values of some residuals from the fit to all n observations. Further information about the structure of the data can be found by looking at the monitoring plot of all residuals. Figure 5 shows the plot of residuals from the FS.

There are several interesting features. One is that there is a change in the fitted model around $m = 1000$, leading to an abrupt change in the structure of the residuals. For $m > 1000$, the pattern is remarkably constant so that the distribution of residuals is throughout close to that shown in Figure 4; the upper tail is not far from normal, but the lower tail is too thin.

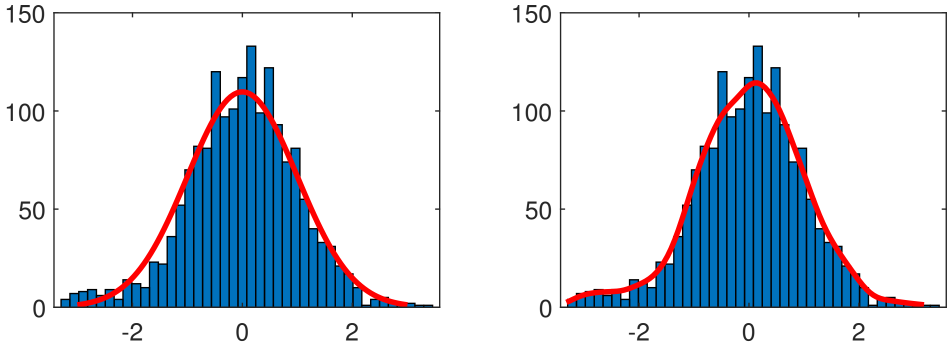

Figure 4. Original data: histogram of Studentized residuals with superimposed curves. **Left**-hand panel: superimposed normal curve; **right**-hand panel: superimposed kernel density estimate.

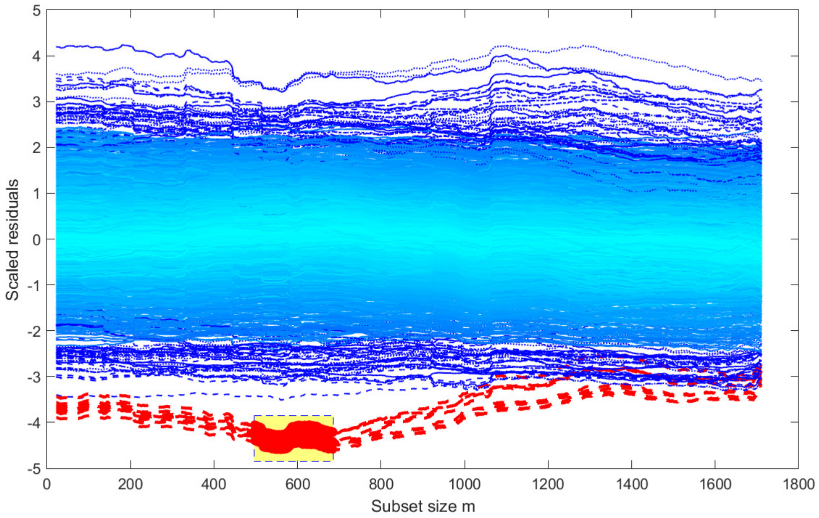

Figure 5. Original data: monitoring residuals from the FS with brushing. Eight observations are highlighted.

A second feature is the band of large negative residuals for m around 200. We use this to illustrate the use of brushing and linking plots. The yellow rectangle in the bottom left-hand corner of Figure 5 shows the brush selected with the cursor. This has highlighted, in red, the trajectories of the eight most negative residuals (for observations 7, 8, 9, 13, 17, 21, 22 and 41). Only four of these (7, 9, 13 and 17) are among the most negative by the end of the search. The question as to which observations these are is answered in Figure 1. In this version of the yX plot of the data, the positions of the observations giving these large residuals are shown as red dots. All these observations have low values of y. The plot of y against x_6, in particular, shows that, in this view, they form a distinct outlying cluster. Since these observations are close together in index number, it would be interesting to explore these observations further, if such fine detail of the data were available.

4. The Added Variable Plot and Extensions

4.1. The Added Variable Plot

It is clear, especially from the residual plots of Figure 3, that there is some unexplained structure in the data, perhaps caused by outliers. In addition, the fitted model may be incorrect. The FS procedure for outlier detection on the original data detects 28 outliers. One question is how the outlying observations affect the fitted model. A second is whether the fit can be improved by transforming the response, perhaps by modelling it as a power of y, such as the square root or the log. Since outliers may be present, these questions are most powerfully answered by monitoring aspects of the robustly fitted model during the FS.

A monitoring plot of the t-statistics for the explanatory variables is a useful tool for discovering the effect of outliers and model failures on the values of the estimated parameters. Although the plots from the FS can be used directly, statistics which disentangle the effect of the FS from estimation of the parameter of interest can have the correct null distribution. To achieve this, we extend the regression model to include an extra explanatory variable, the added variable w, so that the regression model becomes $\mathrm{E}(Y) = X\beta + w\gamma$, where γ is a scalar. Interest is in monitoring the estimates of the coefficient γ of w. In the application to monitoring the FS, the variables in X are used to drive the FS, the values of w being ignored.

We start with static regression with n observations. If the model without γ can be fitted, $(X^T X)^{-1}$ exists and the estimate

$$\hat{\gamma} = w^T(I - H)y / w^T(I - H)w = w^T Ay / w^T Aw, \tag{3}$$

where $A = (I - H)$.

Calculation of the test statistic also requires s_w^2, the residual mean square estimate of σ^2 from regression on X and w, given by

$$(n - p - 1)s_w^2 = y^T y - \hat{\beta}^T X^T y - \hat{\gamma} w^T y = y^T Ay - (y^T Aw)^2 / (w^T Aw). \tag{4}$$

The t statistic for testing that $\gamma = 0$ is then

$$t_w = \hat{\gamma} / (s_w^2 / w^T Aw)^{0.5}. \tag{5}$$

4.2. Monitoring Tests for Regression Coefficients: Extended Added Variable Plots

The expressions for $\hat{\gamma}$, var $\hat{\gamma}$ and s_w^2 are all functions of the quadratic forms $w^T Ay, w^T Aw$ and $y^T Ay$. These can be written as sums of squares and products of the residuals from regression of w on X. In order to obtain monitoring plots of t statistics for all variables with the correct null distribution, Atkinson and Riani [25] write the regression model for all n observations in the added variable form as $y = Q\theta + \epsilon = X_{(-j)}\beta_j + w_j\gamma_j + \epsilon, j = 2, \ldots, p$ where $X_{(-j)}$ is $(n - 1) \times p$ and γ_j is a scalar. They in turn take each of the columns of Q as the vector w_j (except the column corresponding to the constant term in the model). There are then $p - 1$ independent forward searches providing $p - 1$ statistics at each step of each search.

The argument that the null distribution of the added variable test is not affected by the ordering of the data depends on the properties of residuals. The added variable test for γ_j is a function solely of residuals from regression on $X_{(-j)}$. Since these residuals are in a space orthogonal to $X_{(-j)}$, the ordering of the observations using $X_{(-j)}$ does not affect the null distribution of the test statistic. Then, for normally distributed errors, the estimates $\hat{\gamma}$ and s^2 are independent, so it follows that the null distribution of the statistic is Student's t.

4.3. The Monitoring Plot of Added Variable t-Statistics for the Loyalty Data

Figure 6 shows the monitoring plots of all six t-statistics from the six forward searches. There is a steady increase in the significance of all variables as the search progresses. If the outliers were important in changing the values of the tests, we would expect an appreciable change in the plotted values, especially when the outliers start to enter the subset. The entry of the 28 outliers found by the overall FS into each individual search is shown in red in the figure. The behaviour for x_1, x_3, x_4 and x_5 is similar. The outliers enter near the end of the search and cause a noticeable, but not practically significant, change in the values of the t-statistics. The behaviour for x_2 and x_6 does not show this sharp change. If all outliers enter at the end, they will have no effect until $m = 1684$. Inspection of a zoom of the last part of the FS shows that a few of the outliers in all searches enter well before this value. The search with x_6 as the added variable is most extreme in this way, with the first outlier entering at $m = 1595$. In the fan plot for response transformation in Figure 7, there is an individual FS for each of five different power transformations of the response. We again plot outliers, this time highlighting the outliers found for the individual searches.

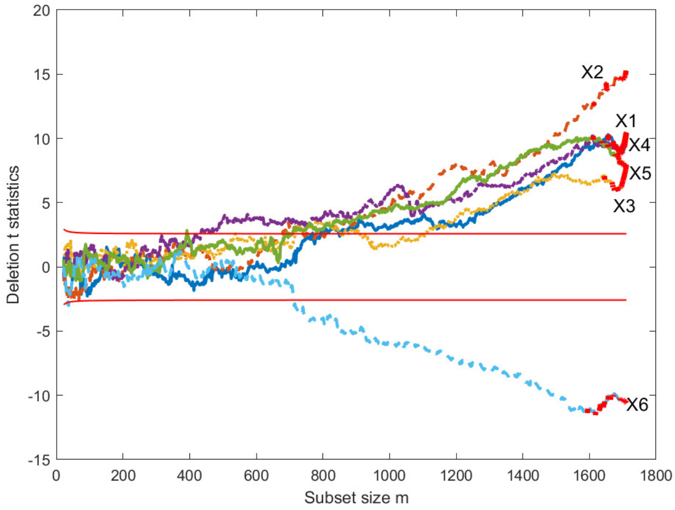

Figure 6. Original data: monitoring of the six added variable t-statistics. Observations highlighted in red are the 28 outliers found from a single FS.

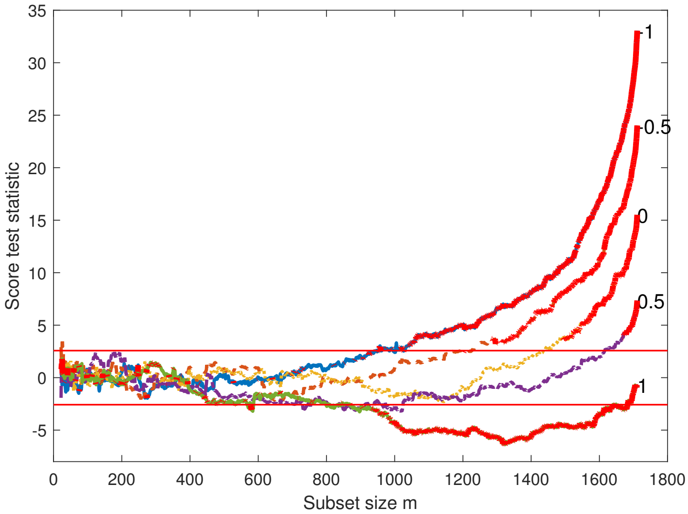

Figure 7. Original data: "fan plot"; monitoring plot of score statistics for five values of λ_0 for the Box–Cox transformation. Observations highlighted in red are the outliers from the individual searches.

5. Response Transformation

5.1. Introduction: The Box–Cox Transformation

There is little evidence from the monitoring plot of t-statistics in Figure 6 of any effect of outlying observations on the fit to the data. It is, however, clear from the residual plots in Figure 3 that there is some systematic misfit between the data and the linear regression model with constant error variance. This may well be indicating that the response requires transformation to provide errors of constant variance and, hopefully, residual plots without such structures as those of Figure 3. In this section, we systematically develop methods for the robust fitting of linear models when the transformation of the response is to be explored.

The use of transformations in data analysis has a long history, as statistical histories go [26]. The parametric family of power transformations described by Box and Cox [27] introduced constructive methods of finding response transformations leading to the approximate normality of the response in the linear model $\mathrm{E}(Y) = X\beta$. The aims of their transformation are the simplicity of structure for $\mathrm{E}(Y)$ (in a regression model the hope is that interaction and second-order terms are not needed), constancy of error variance, and approximate normality of the error distribution.

5.2. Maximum Likelihood Estimation for the Box–Cox Transformation

The Box–Cox transformation for *positive responses* is a function of the parameter λ. The transformed response is

$$y(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda & (\lambda \neq 0) \\ \log y & (\lambda = 0). \end{cases} \tag{6}$$

The value $\lambda = 1$ corresponds to no transformation, $\lambda = 1/2$ to the square root transformation and $\lambda = -1$ to the reciprocal transformation. The use of l'Hôpital's rule as $\lambda \to 0$ shows that the logarithmic transformation is obtained for $\lambda = 0$.

For a given λ, the model is just transformed regression, $y(\lambda) = X\beta(\lambda) + \epsilon$, but now the variance of ϵ, $\sigma^2(\lambda)$, depends on λ. For given λ, the parameters are found by minimization of the sum of squares $S(\lambda) = \{y(\lambda) - X\beta\}^T\{y(\lambda) - X\beta\}$. The least squares estimates of the parameters are $\hat{\beta}(\lambda) = (X^TX)^{-1}X^Ty(\lambda)$, with the mean square estimate of σ^2 being $s^2(\lambda) = S(\lambda)/(n-p)$.

To estimate λ, it is necessary to allow for the change of scale of $y(\lambda)$ with λ. The Jacobian of the transformation J_{BC} [27] is given by $\log J_{\mathrm{BC}} = (\lambda - 1)\sum \log y_i = n(\lambda - 1)\log \dot{y}$, where $\dot{y}$ is the geometric mean of the observations. The normalized transformation, allowing for the change of scale, is

$$z(\lambda) = y(\lambda)/J_{\mathrm{BC}}^{1/n} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} & (\lambda \neq 0) \\ \dot{y}\log y & (\lambda = 0). \end{cases} \tag{7}$$

For fixed λ, the likelihood is maximized by the least squares estimates $\hat{\beta}(\lambda) = (X^TX)^{-1}X^Tz(\lambda)$, with $R(\lambda)$, the residual sum of squares of the $z(\lambda)$, given by $R(\lambda) = z(\lambda)^T(I - H)z(\lambda)$; the maximum likelihood estimator of σ^2 is $\hat{\sigma}^2(\lambda) = R(\lambda)/n$. Then, $\hat{\lambda}$, the maximum likelihood estimate of λ, minimizes $R(\lambda)$.

Box and Cox do not use $\hat{\lambda}$ directly but, where possible, find a value for λ with a physical interpretation that is acceptable and lies within the confidence interval derived from $R(\lambda)$. Physical laws often suggest that such values will be ratios of small integers. Both examples in Box and Cox [27] are from designed experiments and their analyses suggest, respectively, the reciprocal and logarithmic transformations. In this paper, we are analysing data which are not generated by a simple mechanistic model and so for which there may be no straightforward interpretation of any transformations.

Unlike the estimate of λ based on minimization of $R(\lambda)$, minimization of $S(\lambda)$ leads to estimates of β that are highly correlated with the value of λ. References to the confusion this has caused about appropriate inferences about the value of λ are in Section 2 of Atkinson et al. [28]. The practical procedure suggested by Box and Cox is analysis in terms of $z(\lambda)$ leading to $\hat{\lambda}$ and hence to an interpretable estimate $\tilde{\lambda}$ chosen from a grid of plausible values. Carroll [29] argues that the grid needs to become denser as n increases. In conclusion, we stress that our robust data analytic procedures select the value of λ from often coarse grids, the value of λ used in data analyses being guided by the data. It is not the value of $\hat{\lambda}$.

5.3. An Approximate Score Test

For inference about plausible values of the transformation parameter λ, Box and Cox suggest the likelihood ratio test $T_{LR} = n\log\{R(\lambda_0)/R(\hat{\lambda})\}$, where λ_0 is the null transformation to be tested. This test requires the iterative numerical calculation of the estimate $\hat{\lambda}$, which can be avoided by the use of a score test. Atkinson [30] introduced an approximate score test $T_{\text{BC}}(\lambda)$, which is the t-test for an added variable (5), where the added variable w_{BC}, often called a "constructed variable", is found by Taylor series expansion of $z(\lambda)$:

$$z(\lambda) \doteq z(\lambda_0) + (\lambda - \lambda_0)\left.\frac{\partial z(\lambda)}{\partial \lambda}\right|_{\lambda=\lambda_0} = z(\lambda_0) + (\lambda - \lambda_0)w_{\text{BC}}(\lambda_0), \tag{8}$$

which only requires calculations at the hypothesized value λ_0. Straightforward algebra shows that $T_{\text{BC}}(\lambda_0)$ is the t test for regression on $-w_{\text{BC}}(\lambda_0)$, so that large positive values of the statistic mean that λ_0 is too low and that a higher value should be considered.

We now use the algebra of this section, combined with the FS, to develop robust methods for data transformation that enable monitoring of the effects of individual observations on the estimated transformation parameters.

5.4. The Fan Plot

The robust transformation of regression data is complicated by the inter-relationship of outliers and the value of λ since any relationship between mean and variance is a strong indication of the need for response transformation. Examples in Atkinson and Riani [31] [Chapter 4] show the effect of modifying observations on the estimated transformation parameter. We use the forward search to provide a monitoring plot, the "fan plot", of the approximate score statistic $T_{\text{BC}}(\lambda_0)$ over a grid $\mathcal{G}$ of values of λ_0. There is a different search for each value of λ. The ordering of the observations in a fan plot, which reflects the presence of outliers, may depend on the value of λ_0.

Since w_{A} is a constructed variable that includes functions of the observations, $T_{\text{BC}}(\lambda)$, unlike the statistics calculated from monitoring plots of added variables plotted in Figure 6, cannot exactly have a t distribution. Atkinson and Riani [32] provide some numerical results on the distribution in the fan plot of the score statistic for the Box–Cox transformation; increasingly strong regression relationships lead to null distributions that are closer to Student's t.

5.5. Initial Robust Transformation of the Loyalty Data

Figure 7 is the fan plot for the approximate score statistics for five standard values of the Box–Cox parameter λ, together with 99% intervals for testing the value of the parameter. The outliers from the individual searches are highlighted. The lowest trajectory is for the most positive value, $\lambda = 1$, that is, no transformation. Although this transformation is acceptable at the end of the FS, the score statistics lie steadily outside the envelope until $m = 1682$. Working upwards in the plot, the curve for $\lambda = 0.5$ is within the envelope until $m = 1615$. Thereafter, the trajectory rises rapidly. The trajectories for the more negative values of $\lambda = 0, -0.5$ and -1 have a similar shape, but move outside the envelope for smaller values of m. The indication is that these smaller values of λ will not provide a satisfactory transformation for all the data. Although this plot indicates taking $\lambda = 0.5$, there is no obvious discontinuity in the slope of any of the curves to provide a clear indication of the presence of outliers and so to provide a guide to an unambiguous selection of the transformation. The figure shows that the number of outliers detected also increases smoothly as λ decreases from 0.5. The pattern of the outliers for $\lambda = 1$ is different, but again does not show any sharp change as outliers enter the subset for the FS. These trajectories are rather more an indication of a systematic failing in the fitted model.

In the Box–Cox transformation, it is required that the response be positive. Ref. [28] provides monitoring plots for transformations of responses that can be positive or negative, including a generalization of the transformation of Yeo and Johnson [33].

5.6. An Automatic Procedure for the Box–Cox Transformation

In using robust methods to choose the best Box–Cox transformation for a set of data, there may be a trade-off between a seemingly appropriate transformation and the number of observations deleted as outliers. In general, statistical choice between models with the same number of observations, but with differing numbers of parameters, is often made through the use of the Bayesian Information Criterion (BIC), introduced by Schwarz [34]. To allow for models that differ in the number of observations, due to the deletion of outliers, Riani et al. [35] introduced an extended form of BIC. This section provides an introduction to their theory. The method was included by Riani et al. [36] in an automatic method of determining a Box–Cox transformation which is briefly described below and used in the next subsection to analyse the loyalty data.

The inclusion of extra terms in a regression model reduces the residual sum of squares $S(\hat{\beta}_p)$, where p is the number of parameters in the model. The BIC uses a term increasing with p to penalize this increase and so indicate the best number of parameters to include. For regression, $BIC = -n \log\{S(\hat{\beta}_p)/n\} - p \log(n)$. Written in this form, large values of the index are to be preferred, corresponding to small values of $S(\hat{\beta}_p)$, that is, to maximizing the BIC by choice of p.

As our main tool in assessing numerical information from score statistics calculated during the FS, we use an extended BIC. The original version [34] can be extended to choose the value of λ for complete data with n observations:

$$\mathrm{BIC}(\lambda) = -n \log\{S(\lambda)/n\} + 2 \log J_{\mathrm{BC}}^{n} - (p + n_\lambda) \log(n), \tag{9}$$

where J_{BC}^{n} is the Jacobian for all observations and, for the Box–Cox transformation, $n_\lambda = 1$. Use of the forward search to provide robustness against outliers and incorrect transformations leads to the comparison of models fitted to differing numbers of observations. We render outlier deletion compatible with BIC through use of the mean shift outlier model in which deleted observations are each fitted with an individual parameter, so having a zero residual. Let the forward search terminate with m^* observations. Then, $n - m^*$ observations will have been deleted. This can be expressed by writing the regression model as $z(\lambda) = X\beta + D\phi + \epsilon$, where D is an $n \times (n - m^*)$ matrix with a single one in each of its columns and $n - m^*$ rows, all other entries being zero. These entries specify the observations that are deleted [10] (p. 21).

To incorporate deletion of observations in BIC (9) for values of $\lambda \in \mathcal{G}$, let the value of $S(\lambda)$ when $n - m^*(\lambda)$ observations are deleted be $S(\lambda, m^*)$. Then, $\mathrm{BIC}(\lambda)$ is replaced by

$$\mathrm{BIC}(\lambda, m^*) = -n \log\{S(\lambda, m^*)/m^*\} + 2 \log J_{\mathrm{BC}}^{n} - \{p + n_\lambda + n - m^*(\lambda)\} \log(n), \tag{10}$$

in which $S(\lambda, m^*)$ is divided by m^*. A full treatment is in Riani et al. [35].

In addition to the value of BIC, we introduce an empirical diagnostic quantity, the agreement index, AGI. This is a function of the reciprocal of the sum of the absolute values of the score statistics $T_{\mathrm{BC}}(\lambda_0, m)$. Taking the sum over a range of values of m provides information on the stability of the transformation, avoiding reliance solely on its value at m^*. Taking the reciprocal means that large values are desirable.

5.7. Automatic Robust Transformation of the Loyalty Data

The results from the automatic analysis are summarized in the single plot with three panels of Figure 8.

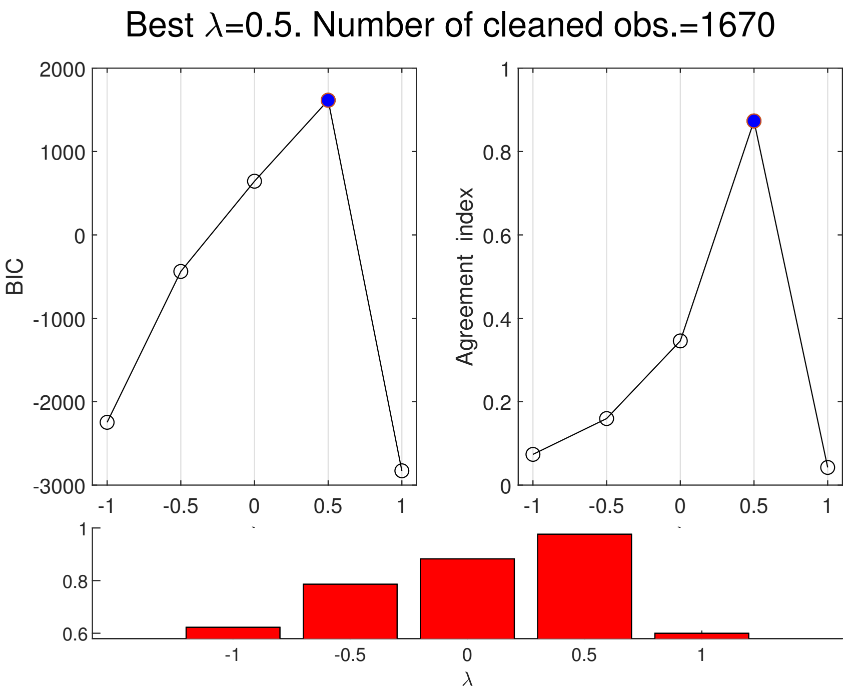

Figure 8. Automatic analysis of Box–Cox transformation. **Upper-left** panel, extended BIC (10); **upper-right** panel, agreement index AGI; **lower** panel, $m^*(\lambda)/n$, the proportion of observations used in fitting the model. Calculations for five values of λ.

The automatic analysis has faithfully extracted all features of Figure 7. The upper-left panel of the figure shows the plot of extended BIC for the five values of λ. There is a clear peak at $\lambda = 0.5$ and the square root transformation is indicated. None of the trajectories of $T_{\mathrm{BC}}(\lambda_0)$ remains inside the 99% bounds over the whole search. However, that for $\lambda = 0.5$ is within for the greatest part of the search, with the trajectories for $\lambda_0 = 0$, -0.5 and -1 having a similar shape, but respectively moving outside the band at smaller values of m. Figure 7 shows that the trajectory of $T_{\mathrm{BC}}(1)$ leaves the band at a slightly lower m than does that for $T_{\mathrm{BC}}(-1)$ and, indeed, $\lambda_0 = 1$ has the smallest, that is worst, value of the extended BIC. In the right-hand panel of the figure, the agreement index AGI shows strong agreement with the plot of the BIC values. The lower panel shows the proportion of observations included in the final transformed analyses. For the square root transformation, this value is close to one (0.976).

The FS on the first-order model with $\sqrt{y}$ as response identifies 41 outliers. The analysis of variance table for the 1670 observations remaining after outlier deletion has a value for adjusted R^2 of 0.779, compared to 0.741 for the original analysis (Table 2). In the new table, four of the variables (price, quality, customer satisfaction and negative publicity) all have appreciably higher levels of significance, with the remaining two levels of significance being slightly reduced. To check the fit of this new model, we start with standard plots of the residuals in Figure 9. The left-hand panel shows a histogram of the Studentized residuals with the superimposed kernel density curve. This is appreciably more symmetrical than the plot of the residuals from regression without transformation in the right-hand panel of Figure 4. The right-hand panel of Figure 9 shows residuals against fitted values, which is completely uninformative about any departures from the model. However, the plots of

Studentized residuals, not given here, show only a slight reduction in the patterns seen in Figure 3. In particular, the symmetry plot is close in form to the lower panel of Figure 3.

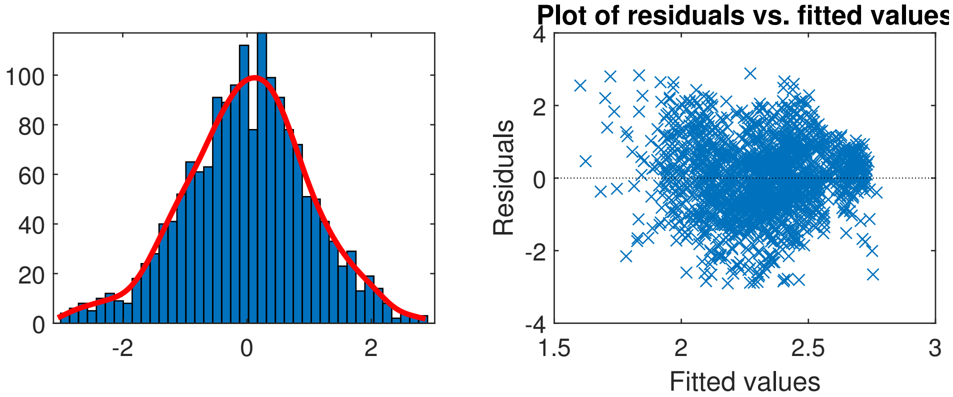

Figure 9. Square root transformation. **Left**-hand panel: histogram of Studentized residuals with superimposed kernel density estimate; **right**-hand panel: Studentized residuals against fitted values.

It is clear that there is still some unexplained structure in the data that is not illuminated by this analysis. The monitoring plot of FS residuals from the square root fit in Figure 10 is helpful. Comparison with the monitoring plot of FS residuals for the untransformed data shows that the change in structure around $m = 1000$ has virtually vanished. A second feature is that the pattern of residuals is very stable over the whole range of m. The trajectories of the 41 outliers are highlighted in red in the figure. These show that the effect of the transformation, as in the left-hand panel of Figure 9, has been to symmetrize the distribution of residuals. There is appreciably more identification of negative residuals as coming from outlying observations than there is of positive residuals. (The outliers are highlighted in the yX plot of Figure 11). The group of eight outliers forming a clear low cluster in the panel for x_6 in Figure 1 is still identified as outlying, as are several observations with the lowest value of y. In addition, some intermediate values now show as outlying in some plots, for example of y against x_2.

It is clear that the parametric Box–Cox transformation cannot provide a sufficiently flexible family of transformations to explain these data. Accordingly, in the next sections, we explore the use of non-parametric transformations of both the response and of the explanatory variables.

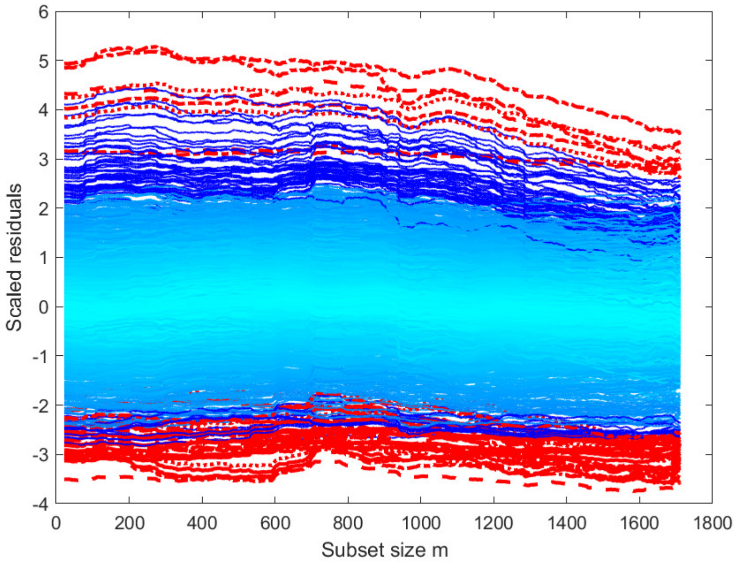

Figure 10. Square root transformation: monitoring residuals with 41 outliers highlighted.

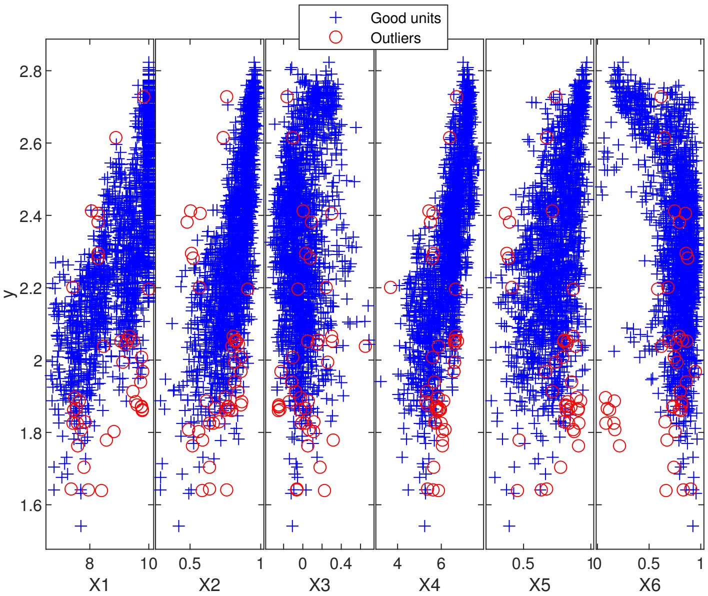

Figure 11. Square root transformation: yX plot with the 41 outliers highlighted.

6. Robust Non-Parametric Regression with Transformation of the Response and Explanatory Variables

6.1. AVAS: Additivity and Variance Stabilization

The parametric Box–Cox transformation of the response has failed to produce a satisfactory model for the loyalty data. We now move to non-parametric transformations of both the response and explanatory variables. Such transformations allow the data to determine the form of the transformations, rather than yielding a parameter estimate in a specified function. If any of these transformations suggest a parametric form, the parameter can be estimated and the effect of the parametric transformation compared with that of the non-parametric one.

Our starting point is Tibshirani [1] who 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 approximately constant variance. Tibshirani's work can be seen as a non-parametric extension of the power transformation family of Box and Cox [27]. A discussion of the relationship of AVAS to the Box–Cox transformation is in Hastie and Tibshirani [37] (Cap. 7).

We use a statistically improved version of AVAS to analyse the loyalty data in Section 6.4. However, the procedure is not robust with respect to outliers. In Section 7, we introduce a robust version of his work, which, for obvious reasons, we call RAVAS. In developing our procedure, we made several important improvements to the original AVAS. Like robustness, these have been programmed to be 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. Riani et al. [38] provide an introduction to RAVAS, including data analysis. Further details of the algorithm are in Riani et al. [2].

6.2. Generalized Additive Models and the Structure of AVAS

We now introduce the generalized additive model and the associated backfitting algorithm for the 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 Section 6.3.

The generalized additive model (GAM) has the form

$$g(Y_i) = \beta_0 + \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i. \tag{11}$$

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 adjusted coefficient of determination R^2_{adj}. 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. [39] describe the background and early development of this model.

The backfitting algorithm, described in Hastie and Tibshirani [37] (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. For the moment, we assume that the response transformation $g(Y)$ is known.

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. The new value of $f_j(.)$ depends on ordered values of x_j and $e_{(j)}$. Let the ordered (sorted) values of x_j be $x_{s,j}$. The residuals $e_{(j)}$ are sorted in the same way to give $e_{s,(j)}$. 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

$$f_{s,j}^{(l+1)} = S\left\{e^l_{s,(j)}, x_{s,j}\right\}. \tag{12}$$

The function S depends on the constraint imposed on the transformation of variable j. If the transformation can be non-monotonic, S denotes a univariate smoothing procedure. As does Tibshirani [1], we use the supersmoother [40], a non-parametric estimator based on local linear regression with adaptive bandwidths. Monotonic transformations using isotonic regression are also an optional possibility [41].

6.3. The Numerical Variance Stabilizing Transformation and the AVAS Algorithm

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

$$g(t) = \int^t 1/\sqrt{V(u)}du. \tag{13}$$

In the AVAS algorithm for data, $1/\sqrt{V(u)}$ is estimated by the vector v 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, computed with the trapezoidal rule. 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 [42].

Evaluation of the integral (13) requires some extrapolation of the values of residuals. Since the purpose is to find a transformation for which $g(y)$ is not a function of the residuals, we extrapolate using the average value of the v. Tibshirani [1] uses another extrapolation. Simulations in Riani et al. [2] (Section 5.2) show the importance of our modification. A careful exposition of the steps of the numerical integration leading to the new transformation of the response is in Atkinson et al. [3] (Section 7.4.9).

The algorithm starts with estimation of the GAM for an initial response transformation $g_0(y)$. At iteration l, the algorithm uses the residuals from the GAM with transformation $g_{l-1}(y)$ to calculate $g_l(y)$. Estimation of the GAM uses the new transformation. The procedure is iterated to convergence.

6.4. Non-Robust Analysis of the Loyalty Data with AVAS

Figure 12 shows that the transformation of the response is slight. The plot of residuals versus fitted values is not a straight band, but is of a more elliptical shape, indicating that there may be several outliers as, for a similar reason, does the plot of loyalty against fit. However, there is no trend of the residuals against fitted values, the presence of which would be indicative of a systematic failure in the model. The more interesting plot is Figure 13 which shows the highly nonlinear transformations of the explanatory variables found by AVAS. These are similar to those found after the deletion of outliers, so we discuss them following our robust analysis. The analysis of variance shows that this non-robust form of model fitting produces a value of $R^2_{\text{adj}} = 0.790$, the highest of any analysis so far, achieved without any deletion of outliers. For clarity, we confirm that this analysis was performed using the non-robust version of our RAVAS algorithm, including the modification in the calculation of the variance stabilizing transformation noted in Section 6.3.

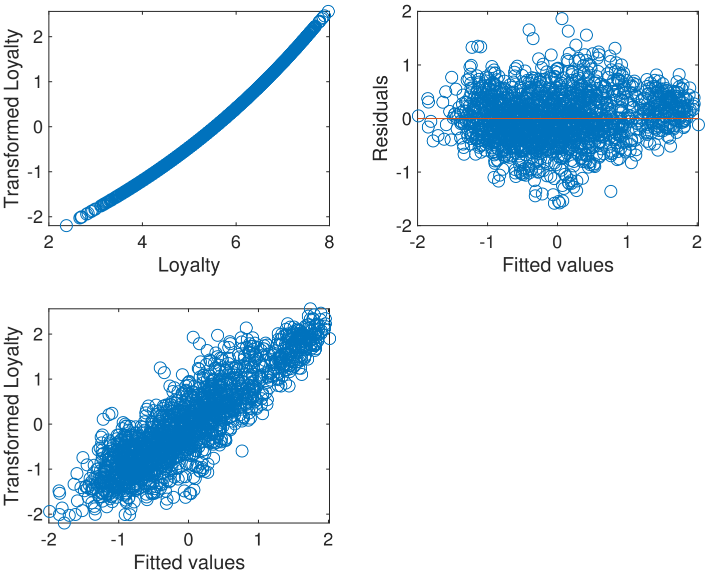

Figure 12. Non-robust AVAS: transformed response $g(y)$. **Upper left**-hand panel, $g(y)$ against y; **upper right**-hand panel, residuals against fitted values; **lower** panel, $g(y)$ against fitted values.

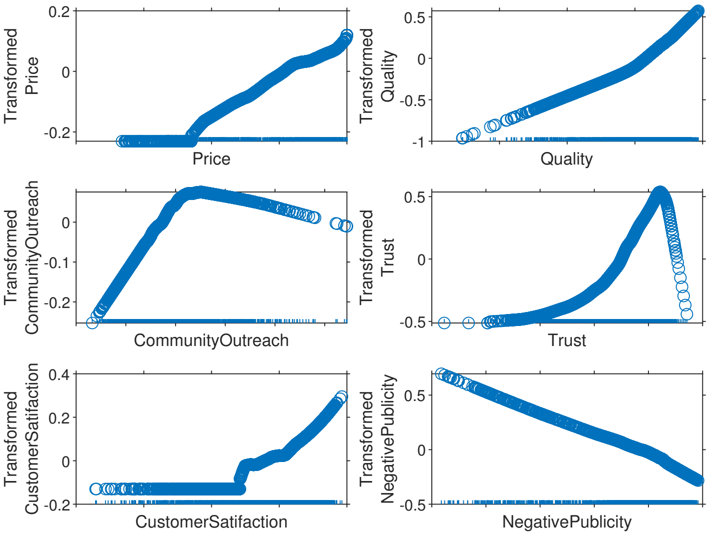

Figure 13. Non-robust AVAS: transformations $f_j(x_j)$ of the explanatory variables against original x_j.

7. Robust Non-Parametric Regression with Response and Explanatory Variable Transformations: RAVAS

7.1. Improvements and Options

Since different response transformations can indicate different observations as outliers, the identification of outliers occurs repeatedly during our robust algorithm, once for each estimated transformation $g_l(y)$.

Our RAVAS procedure introduces five improvements to AVAS, programmed as options. These do not have a hierarchical structure. The overall structure of the algorithm is unchanged from that for AVAS.

Robustness. The performance of AVAS is severely degraded by the presence of outliers. In the example, we provide robustness through use of the FS. The subset S_m, changing at each iteration, defines the observations used in backfitting and in the calculation of the variance stabilizing transformation.

Extrapolation of Residuals. As we mentioned above, in our data analyses, we use the average of the v for extrapolation for the calculation of residuals in the variance stabilizing transformation.

Statistical Improvements to Computational Methods. Our numerical experience is that it is often beneficial to start from a parametric transformation of the response. We use the automatic robust procedure for power transformations described above. In addition, the backfitting algorithm is not invariant to the permutation of order of the explanatory variables. The effect for the first iteration can be reduced through scaling the explanatory variables [43]. In each iteration, the variables are ordered based on those that produce the highest increment of R^2. With this option, the most relevant features are immediately transformed. For robust estimation, this procedure is applied solely to the observations in the subset S_m.

7.2. Robust Analysis with RAVAS

We conclude with the results of the robust analysis using RAVAS with all automatic options. With the robustness of RAVAS we are able to detect outliers—our procedure finds 16 which are highlighted in Figure 14. The plots of residuals against fitted values and of loyalty against fitted values in Figure 14 are pretty much what would be expected from the non-robust analysis except that, with the robust analysis, the residuals have moved further from the main cloud of points. The plot of the transformed response against the original values shows that, although many of the outliers occur among the higher values of y, nothing like all do so. The plots of the transformed explanatory variables, in Figure 15, are similar to those of the non-robust transformations in Figure 13 except for the transformation of trust. We leave the discussion of these curves until we have considered the statistical properties of this fitted model.

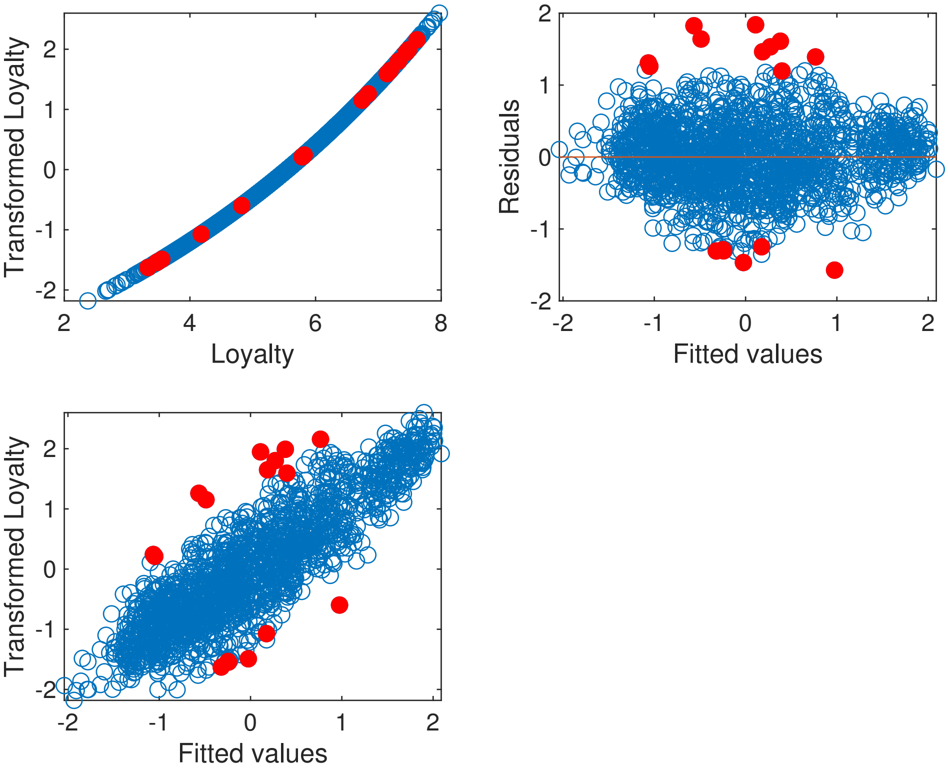

Figure 14. RAVAS: transformed response $g(y)$. **Upper left**-hand panel, $g(y)$ against y; **upper right**-hand panel, residuals against fitted values; **lower** panel, $g(y)$ against fitted values. Sixteen outliers highlighted.

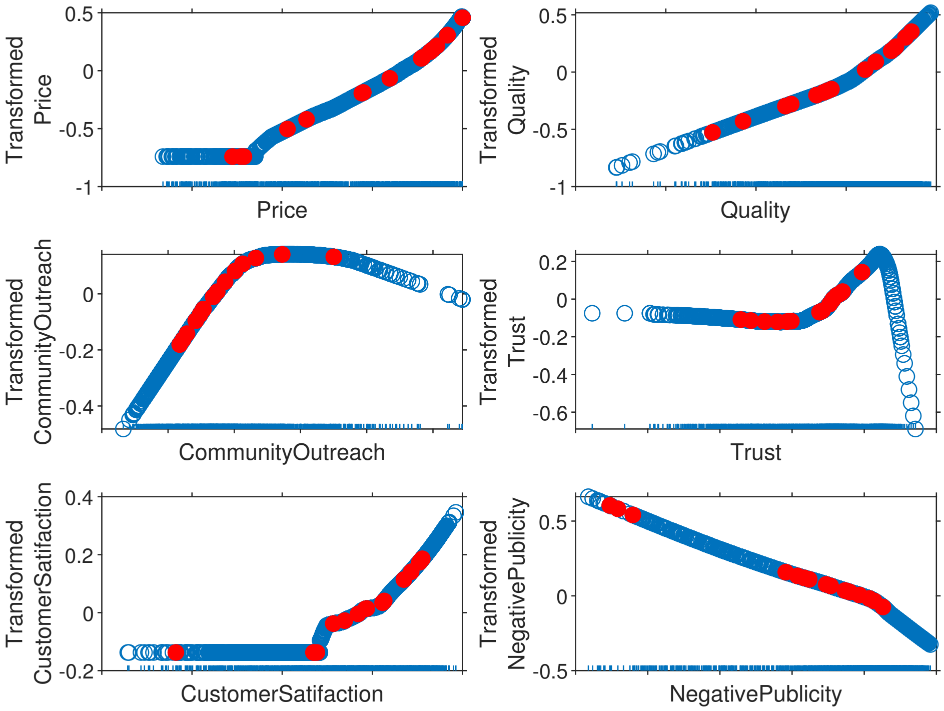

Figure 15. RAVAS: transformations $f_j(x_j)$ of the explanatory variables against original x_j.

The analysis of variance table, Table 3, gives a value of 0.806 for R^2_{adj}, the highest value from all of our analyses, despite the modest number of outliers detected. It is very important that the *t*-statistic for price has the value 18.26 in this table, as opposed to 10.49 in Table 2 for the original data and a similar value of 9.68 for the AVAS analysis. As a check on this model, we look at the plots that, for the untransformed data, gave the strongest warning of nonstandard behaviour of the data. The first were the QQ plot and the symmetry plot of residuals given in Figure 3 for the untransformed data. They are in Figure 16 for the RAVAS analysis. The comparison between the two figures shows that the use of RAVAS and the deletion of 16 observations (less than 1% of the total) have led to an acceptable pattern of residuals, especially in comparison with those of Figure 3. The envelope for the large negative residuals in the upper panel of the plot shows that the negative residuals are now within the envelopes, which they were not in Figure 3. Also, the symmetry plot of the residuals in the lower panel is now appreciably straighter than that in Figure 3.

Table 3. ANOVA: RAVAS (16 observations deleted).

	Estimate	SE	tStat	*p* Value
(Intercept)	-6.5067×10^{-16}	0.010712	-6.0742×10^{-14}	1
Price	0.79922	0.043762	18.263	3.8539×10^{-68}
Quality	1.0932	0.064689	16.9	2.5635×10^{-59}
Community Outreach	0.94503	0.086075	10.979	3.8876×10^{-27}
Trust	1.8194	0.12527	14.523	4.0187×10^{-45}
Customer Satisfaction	0.88925	0.11701	7.5998	4.8883×10^{-14}
Negative Publicity	0.9537	0.071283	13.379	7.1151×10^{-39}
Number of observations: 1695, error degrees of freedom: 1688				
Root Mean Squared Error: 0.441				
R-squared: 0.806, Adjusted R-Squared: 0.806				
F-statistic vs. constant model: 1.17×10^3, *p*-value = 0				

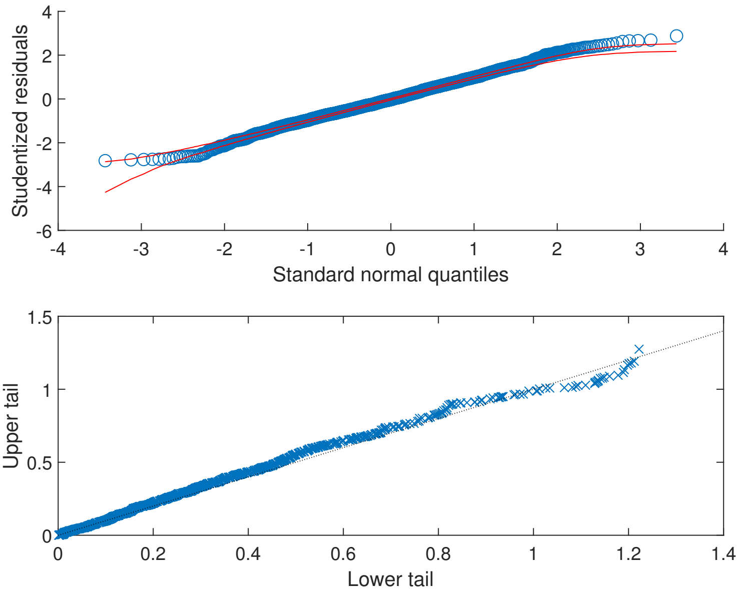

Figure 16. RAVAS (16 observations deleted): cumulative plots of residuals. **Upper** panel: QQ plots of residuals with 95% simulation envelope; **lower** panel: symmetry plot of residuals.

The histograms of the residuals for the untransformed data in Figure 4 showed appreciable skewness. These plots are repeated in Figure 17 for the analysis with RAVAS with 16 outliers excluded. The similarity of the two panels in the figure indicates that a close approximation to normality has been achieved. These 16 outliers are those found in the process of estimating the transformations. As a final analysis, we take the transformation of the response in Figure 14 and of the explanatory variables in Figure 15, and apply the outlier detection procedure of the FS, which identifies nine outliers. The effect of deleting fewer observations is slightly to reduce the value of R^2_{adj} from 0.806 to 0.801 and to slightly increase the value of the t-statistic for price from 18.263 to 18.323, with similarly small increases or decreases in the other t-statistics. The general conclusions of the data analysis are unchanged. The monitoring plot of the residuals when nine observations are deleted is in Figure 18. This shows that the pattern of residuals, including those identified as outliers, is virtually constant. This is an improvement, both in constancy and symmetry, on the monitoring plot of residuals from the square root transformation in Figure 10, which was obtained after 41 observations had been identified as outlying. It is interesting that the majority of the observations deleted for the square root transformation had negative residuals, whereas, for the RAVAS transformation, throughout the search, the majority of observations deleted had positive residuals.

One conclusion is that transformation of the response alone, as judged by these monitoring plots, is not sufficient to provide a good model; the joint transformation of the response and the explanatory variables is needed to produce a well-behaved model.

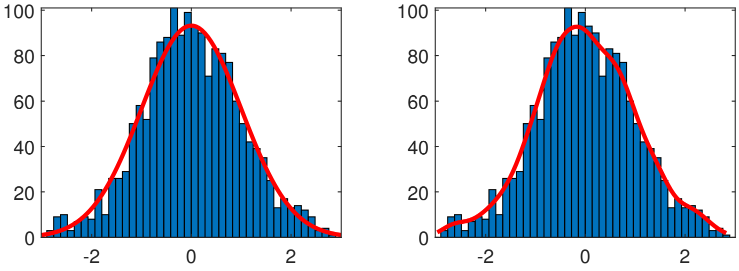

Figure 17. RAVAS (16 observations deleted): histogram of residuals with superimposed curves. **Left**-hand panel: superimposed normal curve; **right**-hand panel: superimposed kernel density.

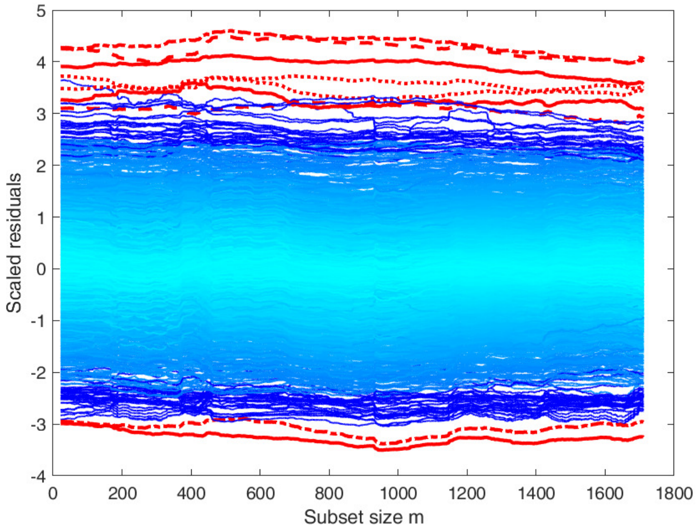

Figure 18. The very stable monitoring plot of FS residuals using transformed variables from RAVAS; the trajectories of the resulting nine outliers are highlighted.

8. Interpretation

The evidence from comparing the plots of transformed x's against their untransformed versions in Figure 15 shows some interesting and unexpected shapes that agree with and confirm some marketing theories based on the distinction between traditional goods—"mass" levers such as range and price—and relational goods—"targeted" levers based on individual customer information [44]. This distinction, the correlations between the original variables (Table 1) and well-known marketing assumptions, helps us to explain the results obtained:

1. x_1—Price. Our RAVAS analysis leads to a t value for this variable that is almost twice the value found from regression on the original data given in Table 2. The primary importance of price is indicated. The first part of the curve shows low loyalty for cheap items. Price is important, but the brand is not. These products are highly fungible without any particular characteristics. Choice between them is often strongly conditioned by promotions and rewards programs [45]. But, increasingly, for more expensive items, loyalty becomes higher as specific characteristics are felt to be important. The purchase of products with higher price positioning and with particular characteristics is the result of a careful selection process by the consumer; an extreme example is represented by the luxury market [46]. When consumers find the product that meets their needs, they tend to build loyalty and are often unwilling to change the product. Furthermore, the increase in loyalty for high-priced goods is strongly conditioned by the need to show off. Loyalty to a certain brand is hugely influenced by emotions given by the achievement of a social value which, according to the consumer, derives from the product or service.
2. x_2—Quality. These are data on what people believe they are doing. In fact, quality is not directly measurable by the consumer but is the result of perception. What emerges from the results is that loyalty increases with perceived quality, in a surprisingly linear way, although with proportionally increased loyalty for products of high perceived quality. As is the case for price, a higher level of quality also includes a psychological component which increases loyalty so that higher prices are assumed to be a guarantee of higher quality. This is supported by the cross-correlation indices between the price, quality and loyalty variables in Table 1. The result agrees and confirms studies in which it was found that price promotions on brands with the lowest loyalty rate must be more aggressive to steal loyal consumers through quality. Furthermore, the possible switch is more probable between fungible assets [47].
3. x_3—Community Outreach. This strategy usually improves corporate image and brand reputation and can work very well in customer relationships. Several studies recorded an increase in consumer loyalty towards companies sensitive to ethical sponsorships, environmental protection, transparency and social responsibility [48]. These strategies present a roughly quadratic relationship, although the decreasing upper part is relatively sparse. This means that community outreach promotes loyalty up to a certain level, after which such strategies may make customers feel that the firm is more interested in its image than in serving them. This seems to mirror the findings that are emerging, for example, from some ongoing studies of green companies, where consumers are starting to suspect that most awareness initiatives are a mere facade strategy [49].
4. x_4—Trust. Trust is unimportant for loyalty at low levels; then, loyalty increases with trust, up to a point, and finally decreases. To explain this trend in the relationship, we have to recall once again the concept of "mass" and "targeted" products mentioned above. Since trust, price and loyalty are strongly correlated, we can say that, for low-priced products, the choice is based on "If I see a bargain I go for it", pointing out the unimportance of trust perception. For medium-priced products, on the other hand, there is an increase in loyalty as a function of trust and value [50]. When the value of the goods increases, they fall into the category of "targeted" goods, those for which a selection process based on emotions prevails and this leads to neglecting loyalty despite great trust. Another interpretation for the final decrease in loyalty is suggested by the result of a study that, in summary, found that customers with high

levels of satisfaction may not especially trust another company that tries to raise their already high level of satisfaction [51].

5. x_5—Customer Satisfaction. The evidence for this relationship appears obvious. Customer satisfaction is a feeling that derives from a mix of different factors such as the quality of service, the quality of the products and the price level [52]. It is therefore natural to expect a result such as the one obtained, in which low levels of customer satisfaction are accompanied by low levels of loyalty which then increase steadily with increases in customer satisfaction.
6. x_6—Negative Publicity. Negative publicity has a strong impact on reputation, which is a loyalty driver. Indeed, when the quality of the product or service is not easy to measure, loyalty is supported more by reputation than by customer satisfaction [53]. Given this, it is quite natural to expect, as our analysis has shown, that reputational collapse will manifest itself as a decrease in loyalty in a surprisingly linear way.

In this paper, we have taken a seemingly innocuous set of marketing data about the subjective determinants of customer loyalty. The analysis of variance table of the original observations (Table 2) gave a value of 0.741 for R^2_{adj} and highly significant regression on all six explanatory variables. It would have been natural to have stopped the analysis at this point. Our further analysis shows that the data do not satisfy the conditions for a regression analysis to be efficient. Our final, robust analysis with non-parametric transformations of the response and explanatory variables yields a fitted generalized additive model for which statistical assumptions are satisfied and for which the proportion of variation in the data explained by the model is appreciably increased in comparison with the original regression model. More important for the understanding of consumer behaviour are the transformations of the explanatory variables shown in Figure 15. These provide appreciably more insight into human behaviour than do models using the first-order terms of the initial regression analysis. These transformations should also be helpful in the planning of marketing campaigns.

We intend that these steps provide a checklist for the analysis of regression data. Some mention of the results from these or similar checks should ideally be included in any published regression analysis.

9. A Procedure for a Structured Approach to Modern Robust Regression Analysis

The analysis of a set of data of an unfamiliar type is rarely straightforward. Cox and Donnelly [54] present principles for applied statistics that cover a wide range of topics, including problem formulation and design of investigations as well as the analysis of data. Wolstenholme et al. [55] is less wide in scope, describing a program for the semi-automatic analysis of data of many types. These references are not concerned about single or multiple sets of observations that may seriously affect the fitted model and its interpretation. If robustness is included, the choices during analysis are multiplied. As our analyses in this paper show, robust regression methods do not follow a simple path. However, we find that there is sufficient structure to indicate the helpfulness of the procedures summarized in Table 4, which suggests five major steps. We comment briefly on them, particularly to provide both bibliographic references and references to sections of the paper.

Table 4. Modern procedure for robust and efficient data analysis.

	Description	**Tools to Use**
STEP 1	Variable transformation.	Parametric (fan plot + automatic procedure for finding best value of λ). Non-parametric (RAVAS + automatic option selection).
Output: Best value of transformation parameter for the response. Find observations influential for transformation. Find best transformation to work with. Compare parametric and non-parametric approach.		
STEP 2	Robust variable selection.	Monitoring of added value t-statistics, candlestick plot or, if the number of variables is very large, robust LASSO.
Output: Find the effect of influential subsets of units on t-statistics. Find a set of relevant explanatory variables.		
STEP 3	Monitoring of scaled residuals. Analysis of FS, S and MM residuals.	Brushing in the monitoring plots to understand the position of outlying residuals and the eventual presence of subgroups of units.
Output: Analysis of correctness of the model and detection of the optimal value of *bdp* or efficiency to use.		
STEP 4	Outlier detection and removal of outlying observations.	Routines for automatic outlier detection. Analysis of the position of the outlying units in the yX plot.
Output: Find a subset of clean units.		
STEP 5	Check residuals on the subset of clean units: heteroskedasticity, normality and serial correlation. Comparison of parametric and non-parametric approach. Find whether the linear approach is reasonable.	QQ plots with envelopes, normality plots, autocorrelation tests.
Output: If some test fails, go to step 2 and restart with another model (e.g., heteroskedastic approach).		

9.1. Step 1: Variable Transformation

The fan plot for transformation of the response is introduced in Section 5; the robust non-parametric transformation RAVAS is in Section 7.

9.2. Step 2: Robust Variable Selection

The added variable plot for t-statistics is introduced in Section 4. The candlestick plot [56] provides a structured plot of the values of the C_p criterion for model choice [57] in the later stages of the FS. For the robust LASSO, see Freue et al. [58], Kepplinger [59]. Note that, since we have removed the outliers and found the appropriate scale, robust variable selection enables us to address the potential problem of (approximate) collinearity among the variables, because we restrict attention to a subset of variables.

9.3. Step 3: Monitoring of Scaled Residuals

Monitoring of the forward plot of residuals has been used frequently. It was introduced in Section 3.4. Alternatively, it is possible to monitor the S residuals or MM residuals in order to determine an optimal value of the empirical *bdp* or efficiency to use in the automatic outlier detection procedure based on these estimators.

9.4. Step 4: Outlier Detection and Removal of Outlying Observations

The FS routine for automatic outlier detection is illustrated by Riani et al. [21]. Analysis of the position of the outlying units in the yX plot is demonstrated in Figure 1.

9.5. Step 5: Check Residuals on the Subset of Clean Units: Heteroskedasticity, Normality and Serial Correlation

QQ plots with envelopes are introduced in Section 3.4. Histograms for assessing normality are first used in Figure 4. We use the test of Durbin and Watson [60] to test for the correlations of regression residuals.

10. Other Problems

In this section we briefly describe other applications of our method of robust monitoring. We start with large datasets and analysis for causality. The assumption of independent and identically distributed observations is unlikely to hold for the large datasets used in machine learning. The potential presence of heterogeneity in measurement error is an additional incentive to use the monitoring approach, leading to the identification of subgroups of data which may have distinct properties. In addition, the identification of stable or unstable patterns or substructures in the data may be of interest. Such structures may well be obscured by outliers, individually or in groups.

An interesting example is causality. Historically, Neyman, see Rubin [61], and Cox [62] concluded that the establishment of causality requires data from a designed experiment, for example, a clinical trial with appropriate randomization and blinding. More recent approaches consider the establishment of causality from observational data. Bühlmann [63] provides a review. The analysis considers sets of data from various known "environments", for example, sources or time periods. An initial aim is to find a set of predictor variables that have the same regression coefficients over the environments. It is clear that robust methods are needed to detect outliers, which may be generating spurious differences between environments. The outcomes of Bühlmann's paper are new prediction methods which are robust against new potentially adversarial environments, which have either not been seen in the data or only been partially seen.

In this approach to causality, the environments are known. In the more general problem of heterogeneity in data for machine learning, the problem is one of the determination of outlier free clusters. Robust clustering of regression models is described in Torti et al. [64].

There are several aspects of linear regression modelling that we have not considered. One is the effect of errors in the explanatory variables model. Kukush and Mandel [65] give a recent non-robust discussion. We have concentrated in this paper on linear regression models. Application of the FS to nonlinear regression models, such as those that occur in chemical kinetics, are illustrated in Chapter 5 of Atkinson and Riani [31]. Generalized linear models are the subject of Chapter 6. The monitoring approach can be extended to any complex model, linear or nonlinear.

Robust methods, which not only detect outliers but, as a result of monitoring, use as many good observations as possible, enable us to understand the degree of stability of the model; we learn the extent of the model's invariance. Tukey [66] wrote "One of the major arguments for regression instead of correlation is potential stability... [it is possible] that the regression coefficients [may] remain the same over a wide range of situations. We are seeking stability of our coefficients so that we can hope to give them theoretical significance". Seeking stability of coefficients is in full agreement with the monitoring approach. We argue that monitoring is, in Bühlmann's phrase, "geared towards causality".

11. Acknowledgements and Software

All the calculations in this paper have used the Flexible Statistics and Data Analysis (FSDA) MATLAB toolbox, which is freely downloadable from the MathWorks File Exchange website https://www.mathworks.com/matlabcentral/fileexchange/72999-fsda (accessed

on 22 November 2024) or from GitHub at the web address https://uniprjrc.github.io/FSDA/ (accessed on 22 November 2024).

We acknowledge financial support from the following:

- The European Union NextGenerationEU/NRRP, Mission 4 Component 2 Investment 1.5, Call 3277 (30 December 2021), Award 0001052 (23 June 2022), under the project ECS00000033 "Ecosystem for Sustainable Transition in Emilia-Romagna", Spoke 6 "Ecological Transition Based on HPC and Data Technology".
- The University of Parma project "Robust statistical methods for the detection of frauds and anomalies in complex and heterogeneous data".
- The Ministry of Education, University and Research project "Innovative statistical tools for the analysis of large and heterogeneous customs data" (2022LANNKC).

Our research has also benefited from the High Performance Computing (HPC) facilities of the University of Parma. We are grateful to two anonymous referees for helpful comments which led to the clarification of the current paper.

Author Contributions: Methodology, M.R. and A.C.A.; Software, A.C.; Investigation, G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in FSDA, at https://github.com/UniprJRC/FSDA, accessed on 22 November 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tibshirani, R. Estimating transformations for regression via additivity and variance stabilization. *J. Am. Stat. Assoc.* **1988**, *83*, 394–405. [CrossRef]
2. Riani, M.; Atkinson, A.C.; Corbellini, A. Robust transformations for multiple regression via additivity and variance stabilization. *J. Comput. Graph. Stat.* **2023**, *33* , 85–100. [CrossRef]
3. Atkinson, A.C.; Riani, M.; Corbellini, A.; Perrotta, D.; Todorov, V. *Robust Statistics Through the Monitoring Approach: Applications in Regression*; Springer: Berlin/Heidelberg, Germany, 2025; In press.
4. Student. The probable error of a mean. *Biometrika* **1908**, *6*, 1–25. [CrossRef]
5. Fisher, R.A. *Statistical Methods for Research Workers*; Oliver and Boyd: Edinburgh, UK, 1925.
6. Lehmann, E.L. *Fisher, Neyman, and the Creation of Classical Statistics*; Springer: New York, NY, USA, 2011.
7. Draper, N.R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, NY, USA, 1966.
8. Anscombe, F.J. Examination of Residuals. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1961; Volume 1, pp. 1–36.
9. Belsley, D.A.; Kuh, E.; Welsch, R.E. *Regression Diagnostics*; Wiley: New York, NY, USA, 1980.
10. Cook, R.D.; Weisberg, S. *Residuals and Influence in Regression*; Chapman and Hall: London, UK, 1982.
11. Atkinson, A.C. *Plots, Transformations, and Regression*; Oxford University Press: Oxford, UK, 1985.
12. Cook, R.D.; Weisberg, S. *Applied Regression Including Computing and Graphics*; Wiley: New York, NY, USA, 1999.
13. Andrews, D.F.; Bickel, P.J.; Hampel, F.R.; Tukey, W.J.; Huber, P.J. *Robust Estimates of Location: Survey and Advances*; Princeton University Press: Princeton, NJ, USA, 1972.
14. Stigler, S.M. The Changing History of Robustness. *Am. Stat.* **2010**, *64*, 277–281. [CrossRef]
15. Cerioli, A.; Farcomeni, A.; Riani, M. Wild adaptive trimming for robust estimation and cluster analysis. *Scand. J. Stat.* **2019**, *46*, 235–256. [CrossRef]
16. Huber, P.J. *Robust Statistics*; Wiley: New York, NY, USA, 1981.
17. Maronna, R.A.; Martin, R.D.; Yohai, V.J. *Robust Statistics: Theory and Methods (with R)*, 2nd ed.; Wiley: Chichester, UK, 2019.
18. Rousseeuw, P.J.; Leroy, A.M. *Robust Regression and Outlier Detection*; Wiley: New York, NY, USA, 1987.
19. Tallis, G.M. Elliptical and Radial Truncation in Normal Samples. *Ann. Math. Stat.* **1963**, *34*, 940–944. [CrossRef]
20. Atkinson, A.C.; Riani, M.; Cerioli, A. The Forward Search: Theory and data analysis (with discussion). *J. Korean Stat. Soc.* **2010**, *39*, 117–134. [CrossRef]
21. Riani, M.; Cerioli, A.; Atkinson, A.C.; Perrotta, D. Monitoring Robust Regression. *Electron. J. Stat.* **2014**, *8*, 642–673. [CrossRef]

22. Cerioli, A.; Riani, M.; Atkinson, A.C.; Corbellini, A. The power of monitoring: How to make the most of a contaminated multivariate sample (with discussion). *Stat. Methods Appl.* **2018**, *27*, 559–666. [CrossRef]
23. Berman, B. Developing an effective customer loyalty program. *Calif. Manag. Rev.* **2006**, *49*, 123–148. [CrossRef]
24. Mascarenhas, O.A.; Kesavan, R.; Bernacchi, M. Lasting customer loyalty: A total customer experience approach. *J. Consum. Mark.* **2006**, *23*, 397–405. [CrossRef]
25. Atkinson, A.C.; Riani, M. Forward search added-variable *t* tests and the effect of masked outliers on model selection. *Biometrika* **2002**, *89*, 939–946. [CrossRef]
26. Cox, D. Nonlinear models, residuals and transformations. *Math. Operationsforsch. U Statist.* **1977**, *8*, 3–22.
27. Box, G.E.P.; Cox, D.R. An analysis of transformations (with discussion). *J. R. Stat. Soc. Ser. B* **1964**, *26*, 211–252. [CrossRef]
28. Atkinson, A.C.; Riani, M.; Corbellini, A. The Box-Cox transformation: Review and extensions. *Stat. Sci.* **2021**, *36*, 239–255. [CrossRef]
29. Carroll, R.J. Prediction and Power Transformations when the Choice of Power is Restricted to a Finite Set. *J. Am. Stat. Assoc.* **1982**, *77*, 908–915. [CrossRef]
30. Atkinson, A.C. Testing transformations to normality. *J. R. Stat. Soc. Ser. B* **1973**, *35*, 473–479. [CrossRef]
31. Atkinson, A.C.; Riani, M. *Robust Diagnostic Regression Analysis*; Springer: New York, NY, USA, 2000.
32. Atkinson, A.C.; Riani, M. Tests in the fan plot for robust, diagnostic transformations in regression. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 87–100. [CrossRef]
33. Yeo, I.K.; Johnson, R.A. A new family of power transformations to improve normality or symmetry. *Biometrika* **2000**, *87*, 954–959. [CrossRef]
34. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
35. Riani, M.; Atkinson, A.C.; Corbellini, A.; Farcomeni, A.; Laurini, F. Information Criteria for Outlier Detection Avoiding Arbitrary Significance Levels. *Econom. Stat.* **2022**. [CrossRef]
36. Riani, M.; Atkinson, A.C.; Corbellini, A. Automatic robust Box-Cox and extended Yeo-Johnson transformations in regression. *Stat. Methods Appl.* **2022**, *32*, 75–102. [CrossRef]
37. Hastie, T.J.; Tibshirani, R.J. *Generalized Additive Models*; Chapman and Hall: London, UK, 1990.
38. Riani, M.; Atkinson, A.C.; Corbellini, A. Robust response transformations for generalized additive models via additivity and variance stabilisation. In *Selected Papers of 13th Scientific Meeting of Classification and Data Analysis Group—CLADAG 2021*; Grilli, L., Lupparelli, M., RampichinI, C., Rocco, E., Vichi, M., Eds.; Springer: Cham, Switzerland, 2023.
39. Buja, A.; Hastie, T.; Tibshirani, R. Linear Smoothers and Additive Models. *Ann. Stat.* **1989**, *17*, 453–510. [CrossRef]
40. Friedman, J.; Stuetzle, W. Smoothing of scatterplots. In *Technical Report ORION 003*; Technical Report; Department of Statistics, Stanford University: Stanford, CA, USA, 1982.
41. Barlow, R.E.; Bartholomew, D.J.; Bremner, J.M.; Brunk, H.D. *Statistical Inference under Order Restrictions*; Wiley: Chichester, UK, 1972.
42. Hastie, T.; Tibshirani, R. Generalized Additive Models. *Stat. Sci.* **1986**, *1*, 297–318. [CrossRef]
43. Breiman, L. Comment on "Monotone regression splines in action" (Ramsey, 1988). *Stat. Sci.* **1988**, *3*, 442–445. [CrossRef]
44. Bellini, S.; Cardinali, M.G.; Ziliani, C. Building customer loyalty in retailing: Not all levers are created equal. *Int. Rev. Retail. Distrib. Consum. Res.* **2011**, *21*, 461–481. [CrossRef]
45. Lal, R.; Bell, D.E. The impact of frequent shopper programs in grocery retailing. *Quant. Mark. Econ.* **2003**, *1*, 179–202. [CrossRef]
46. Yoo, J.; Park, M. The effects of e-mass customization on consumer perceived value, satisfaction, and loyalty toward luxury brands. *J. Bus. Res.* **2016**, *69*, 5775–5784. [CrossRef]
47. Allender, W.J.; Richards, T.J. Brand loyalty and price promotion strategies: An empirical analysis. *J. Retail.* **2012**, *88*, 323–342. [CrossRef]
48. Singh, J.J.; Iglesias, O.; Batista-Foguet, J.M. Does having an ethical brand matter? The influence of consumer perceived ethicality on trust, affect and loyalty. *J. Bus. Ethics* **2012**, *111*, 541–549. [CrossRef]
49. Hameed, I.; Hyder, Z.; Imran, M.; Shafiq, K. Greenwash and green purchase behavior: An environmentally sustainable perspective. *Environ. Dev. Sustain.* **2021**, *23*, 13113–13134. [CrossRef]
50. Agustin, C.; Singh, J. Curvilinear effects of consumer loyalty determinants in relational exchanges. *J. Mark. Res.* **2005**, *42*, 96–108. [CrossRef]
51. Vlachos, P.A.; Vrechopoulos, A.P.; Pramatari, K. Too much of a good thing: Curvilinear effects in the evaluation of services and the mediating role of trust. *J. Serv. Mark.* **2011**, *25*, 440–450. [CrossRef]
52. Sivadas, E.; Baker-Prewitt, J.L. An examination of the relationship between service quality, customer satisfaction, and store loyalty. *Int. J. Retail. Distrib. Manag.* **2000**, *28*, 73–82. [CrossRef]
53. Selnes, F. An examination of the effect of product performance on brand reputation, satisfaction and loyalty. *Eur. J. Mark.* **1993**, *27*, 19–35. [CrossRef]
54. Cox, D.R.; Donnelly, C.A. *Principles of Applied Statistics*; Cambridge University Press: Cambridge, UK, 2011.

55. Wolstenholme, D.E.; O'Brien, C.M.; Nelder, J.A. GLIMPSE: A knowledge-based front end for statistical analysis. *Knowl.-Based Syst.* **1988**, *1*, 173–178. [CrossRef]
56. Riani, M.; Atkinson, A.C. Robust model selection with flexible trimming. *Comput. Stat. Data Anal.* **2010**, *54*, 3300–3312. [CrossRef]
57. Mallows, C.L. Some comments on C_p. *Technometrics* **1973**, *15*, 661–675.
58. Freue, G.V.C.; Kepplinger, D.; Salibian-Barrera, M.; Smucler, E. Robust elastic net estimators for variable selection and identification of proteomic biomarkers. *Ann. Appl. Stat.* **2019**, *13*, 2065–2090.
59. Kepplinger, D. Robust variable selection and estimation via adaptive elastic net S-estimators for linear regression. *Comput. Stat. Data Anal.* **2023**, *183*, 107730. [CrossRef]
60. Durbin, J.; Watson, G.S. Testing for Serial Correlation in Least Squares Regression: I. *Biometrika* **1950**, *37*, 409–428. [PubMed]
61. Rubin, D.B. Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Stat. Sci.* **1990**, *5*, 472–480. [CrossRef]
62. Cox, D.R. Causality: Some Statistical Aspects. *J. R. Stat. Soc. Ser. A* **1992**, *155*, 291–301. [CrossRef]
63. Bühlmann, P. Invariance, Causality and Robustness. *Stat. Sci.* **2020**, *35*, 404–426. [CrossRef]
64. Torti, F.; Riani, M.; Morelli, G. Semiautomatic robust regression clustering of international trade data. *Stat. Methods Appl.* **2021**, *30*, 863–894. [CrossRef] [PubMed]
65. Kukush, A.; Mandel, I. A validity test for a multivariate linear measurement error model. *Model Assist. Stat. Appl.* **2024**, *19*, 97–115. [CrossRef]
66. Tukey, J. Causation, regression, and path analysis. In *Statistics and Mathematics in Biology*; Kempthorne, O., Ed.; Iowa State College Press: Ames, IA, USA, 1954; pp. 35–66.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.