TRANSFORMATIONS IN REGRESSION: A ROBUST ANALYSIS

by

R. J. Carroll[1]

and

David Ruppert[2]

Running Title: ROBUST TRANSFORMATIONS

Keywords: Power transformations, Box-Cox model, robust estimation, influence functions, bounded influence, likelihood-ratio-type tests.

AMS 1970 Subject Classifications: Primary 62G35, 62J05.

Abstract

We consider two approaches to robust estimation for the Box-Cox power transformation model. One approach maximizes weighted, modified likelihoods. A second approach bounds the self-standardized gross-error sensitivity, a measure of the potential influence of outliers pioneered by Krasker and Welsch (JASA, 1982).

Among our primary concerns is the performance of these estimators on actual data. In examples that we study, there seem to be only minor differences between these three estimators, but they behave rather differently than the maximum likelihood estimator or estimators that bound only the influence of the residuals.

Confidence limits for the transformation parameter can be obtained by using a large-sample normal approximation or by modified likelihood-ratio testing.

These examples show that model selection, determination of the transformation parameter, and outlier identification are fundamentally interconnected.

[1]Department of Statistics, University of North Carolina at Chapel Hill. Supported by the Air Force Office of Scientific Research Contract No. AFOSR F49620 82 C 0009.

[2]Department of Statistics, University of North Carolina at Chapel Hill. Supported by the National Science Foundation Grant MCS 8100748.

1. Introduction

Transformations have long been used to bring data into a form which satisfies, at least approximately, the assumptions of a convenient parametric model. The choice of a transformation has often been made in an ad hoc trial-and-error fashion. In a seminal paper, Box and Cox (1964) suggested enlarging the parametric model to include transformation parameters, and then estimating these parameters simultaneously with the parameters in the original model. Specifically, they considered the model

$$y^{(\lambda)} = x^t\beta + \varepsilon \,, \tag{1.1}$$

where $y^{(\lambda)}$ is a transformed response variable, λ is a transformation parameter, x is a p-dimensional vector of explanatory variables, β is a vector of unknown regression coefficients, and ε is $N(0,\sigma^2)$ distributed. They focused on the power transformation family, modified slightly to include the log-transformation:

$$\begin{aligned} y^{(\lambda)} &= (y^{\lambda}-1)/\lambda \qquad \text{if } \lambda \neq 0 \\ &= \log(y) \qquad\quad \text{if } \lambda = 0. \end{aligned} \tag{1.2}$$

They used either maximum likelihood or Bayes estimation to simultaneously estimate β, σ, and λ. Unfortunately, the maximum likelihood estimator (MLE) is very sensitive to outliers (Carroll, 1982), and it can be highly inefficient if the distribution of ε has heavier tails than the normal distribution. For the classical regression model with normally distributed errors, the MLE is just the least squares estimator and methods have been proposed for handling the non-robustness. These methods include regression diagnostics (Belsley, Kuh, and Welsch, 1980; Cook and Weisberg, 1982) and robust estimation (Krasker and Welsch, 1982).

Useful diagnostic methods for transformation problems have been introduced by Atkinson (1981, 1982, 1983) and Cook and Wang (1982).

In a review of recent advances in regression methodology, Hocking (1983) and the discussants to his paper comment upon the role of robust estimation and regression diagnostics in data analysis. They contrast two tasks before the statistician, the identification of influential points and the accommodation of outliers.

At least some of these authors appear to implicitly assume that case-deletion diagnostics are the most suitable method for the identification of influential points. They all agree that the automatic accommodation of outliers by the unthinking use of a robust estimator is undesirable. With some exceptions, the overall impression is that the usefulness of robust methods is in doubt.

We agree that robust estimators should not be used blindly. However, when analyzing the examples presented later in this paper, we found that robust fits to the data provided quite a bit to think about. Both the weights given to the cases and the residuals from the robust fits can provide useful diagnostic information. Also, the score function evaluated at each case (with the parameter set at the robust estimate) can aid in the identification of influential points.

The most sensible view would seem to be that case-deletion diagnostics should be used in conjunction with robust methods and their associated diagnostics, and vice-versa. We believe our examples make a strong case for this compromise approach. Many but not all find that case-deletion diagnostics are often easier to devise for specific purposes and implement than are robust methods, especially in linear models. While case-deletion diagnostics have proved to be useful and the arguments listed above are powerful, they do have some potential drawbacks that need to be kept in

mind and which we hope will be the focus of future discussion. The first difficulty is a common form of masking, i.e., after deletion of one or two influential points a hitherto unsuspected observation emerges as extremely influential; see our first example for just such a circumstance. This leads to the little discussed problem of when to stop the sequential deletion of influential points; see Dempster and Gasko-Green (1981). In larger data sets it is likely to be groups of observations that have influence and not merely a single point; the identification of such a group by single case-deletion is likely to be tedious, and group-deletion methods have not been well-discussed and seem difficult to implement.

Not all robust methods are immune to the masking difficulty. However, robust estimates and their associated diagnostics do give us a different look at the data. Even in the small data sets we examine below, the robust methods single out groups of observations for further study, groups which are not identified by one pass through a single case-deletion diagnostic. It is because of this experience that we believe an interaction between case-deletion and robustness will be especially fruitful, for complex, moderate-sized data sets as well as larger data sets.

Robust estimators for the Box-Cox transformation parameter have been proposed by Carroll (1980, 1982) and Bickel and Doksum (1981). These estimators are designed to handle only outliers in the errors (ε_i), not outliers in the explanatory variables (x_i). For regression problems, the so-called bounded-influence regression estimators (Hampel, 1978; Krasker, 1980; Krasker and Welsch, 1982; Huber, 1983) are designed to handle outliers in both ε and x. These estimators are obtained by generalizing to multiparameter problems the work of Hampel (1968, 1974), which for a univariate parametric family finds the estimator with minimum asymptotic variance, subject to a bound on the gross-error-sensitivity and Fisher-consistency at the parametric model.

In this paper, we generalize bounded-influence regression estimators to the Box-Cox transformation problem and obtain estimators bounding the influence of outliers in both ε and x. We also introduce a Hampel-type estimator which bounds the influence function for β, σ and λ. The latter estimator is constructed to handle observations which greatly affect the derivative with respect to λ of the log-likelihood, as well as observations which are outliers in ε or x.

Since our estimators are M-estimators, their asymptotic properties can be established by known techniques. Hence, rather than establishing these technical details, we study the behavior of these estimators on actual data sets.

In general, we find that the estimators we introduce do handle outliers rather well, and they can be used to gain new insights into the structure of a data set. The selection of a regression model, the estimation of λ, and the identification of outliers are intrinsically connected. In particular, we often find that an observation outlying in both ε and x can be handled equally well by either removing it, or else by slightly enlarging the model and possibly also changing the value of λ. When an outlier is identified it is well worthwhile to consider whether it is an indication of an inadequate model. The examples suggest that in some cases the primary task of a robust estimator will be to identify observations to be set aside or inadequacies in the model. After these corrections, a robust estimator may not be needed; maximum likelihood estimation may suffice. In the context of regression, Welsch (1983) states "...I feel that if one does not have high confidence in the X data, it is essential to compare a least squares estimate with a bounded-influence estimate". We agree and add that this comparison should also be made when one does not have complete confidence in the model. Snee (1983) concurs when he states that "[Robust methods] can be

useful if we use them to fit the bulk of the data and then focus our attention on the remaining aberrant points or set of points seeking to understand why they are different".

2. The Estimators

We will introduce two classes of estimators, those based on pseudo-likelihoods and those based on the bounded influence ideas of Hampel (1968, 1974) and Stahel (1981).

2.1 The MLE. Suppose that we observe (y_i,x_i), $i=1,\ldots,n$ that are independently distributed according to the model given by equations (1.1) and (1.2). Let

$$C(\lambda) = \prod_{i=1}^{n} (y_i^{1/n})^{\lambda-1}$$

be the nth root of the Jacobian of $(y_1,\ldots,y_n) \to (y_1^{(\lambda)},\ldots, y_n^{(\lambda)})$. Then, apart from a constant, the log-likelihood is

$$(2.1) \qquad -(\tfrac{1}{2}) \sum_{i=1}^{n} \{\log(\sigma^*)^2+[(y_i^{(\lambda)} - x_i^t\beta)/(\sigma^*C(\lambda))]^2\} \quad ,$$

where $\sigma^* = \sigma/C(\lambda)$. The MLE maximizes (2.1).

2.2 Pseudo-likelihood estimators. In the evolution of robust estimators for an ordinary linear regression, much insight was gained and many advances were made through the idea of pseudo-likelihoods. Basically, this idea involves modifying the usual log-likelihood so as to obtain estimators with certain desirable properties. The estimators of Huber (1973), Mallows, and Schweppe all can be looked at in this way. (The last two estimators are discussed in Peters, Samarov, and Welsch (1982) and Krasker and Welsch (1982).)

For the estimation of power transformations, a first step towards a development of pseudo-likelihood estimators has already taken place, and in this section we will carry the idea further. We will introduce the general

class and ultimately nominate one member of this class for detailed study in the examples.

Let $\rho(\cdot)$ be any function with derivative $\psi(\cdot)$, and let $\{(s_i,t_i)\}$ be sets of weights between zero and one. We define the logarithm of the pseudo-likelihood to be

$$(2.2) \qquad -\sum_{i=1}^{N} s_i t_i \{\log \sigma + \rho((Y_i^{(\lambda)} - x_i^T\beta)/(\sigma t_i C(\lambda)))\}.$$

The pseudo-likelihood estimators are in principle defined to maximize (2.2).

If $\rho(x) = x^2/2$ and $s_i = t_i = 1$, then (2.2) reduces to the usual log-likelihood (2.1), so that the pseudo-likelihood estimators include the MLE. If

$$(2.3) \qquad \begin{aligned} \rho'(x) = \psi(x) &= x & |x| \le K \\ &= K \text{ sign } (x) & |x| > K \end{aligned}$$

and $s_i = t_i = 1$, we get the Huber-type estimators proposed by Carroll (1980) and Bickel and Doksum (1981). As a further protection against massive outliers in the response, Carroll (1982) proposed using the function

$$(2.4) \qquad \begin{aligned} \rho'(x) = \psi(x) & \\ = -\psi(-x) &= x & 0 \le x < a \\ &= a & a \le x < b \\ &= a(c-x)/(c-b) & b \le x < c \\ &= 0 & x \ge c , \end{aligned}$$

with $s_i = t_i = 1$. The form (2.4) is a type of "Hampel" function, and Carroll (1982) shows that it can be particularly useful in designed experiments which have large outliers; see this last paper for computational conventions.

The estimators which appear in the literature are useful and provide protection against outliers in the response for certain designed experiments, but they are all susceptible to the joint influence of outliers in the response and the design, see Cook and Wang (1982). The purpose of the terms $\{(s_i,t_i)\}$ in (2.2) is to provide some protection against the effects of response outliers occurring at high-leverage design points. The particular choice of $\{(s_i,t_i)\}$ turns out to be crucial. We have had some success with the following method, which is derived from a formal solution of the optimality problem given in equation (3.3) of Krasker and Welsch (1982). Let $a_x > 1$ be a constant and by iteration find the p×p matrix A solving

$$A = N^{-1} \sum_{i=1}^{N} w_{x,i} x_i x_i^T ,$$

where the design weights $w_{x,i}$ satisfy (recall that p = dimension of x_i)

$$w_{x,i} = \min\{1, pa_x (x_i^T A^{-1} x_i)^{-1}\} \tag{2.5}$$

The quantity $(x_i^T A^{-1} x_i)^{\frac{1}{2}}$ is a robust version of the Mahalanobis distance of x_i from the centroid of the design and $w_{x,i}$ measures the potential for downweighting of the ith observation caused by x_i being outlying.

If we choose $s_i = w_{x,i}$ and $t_i = 1$, we have a Mallows-type transformation, essentially using a weighted log-pseudo-likelihood which automatically downweights any outlying design points. We have had more success with the Schweppe-type idea wherein $s_i = t_i = w_{x,i}$, which more strongly downweights outlying responses when they occur at high leverage design points. The cost of using this extra weighting is a loss of efficiency for clean data.

The member of the class of pseudo-likelihood estimators which is used in the examples is described here; throughout the paper it will be denoted the pseudo-MLE. In equation (2.5), choose $a_x = 1.4$ and define

$$s_i = t_i = w_{x,i} .$$

For fixed λ, the estimates of (β,σ) will formally satisfy

$$\sum_{i=1}^{N} x_i w_{x,i} w_{r,i}\, r_i(\lambda) = 0, \tag{2.6}$$

$$\sum_{i=1}^{N} w_{x,i}^2 \{w_{r,i} r_i^2(\lambda) - 1\} = 0 \ , \quad \text{where} \tag{2.7}$$

$$r_i(\lambda) = (Y_i^{(\lambda)} - x_i^T\beta)/(\sigma w_{x,i} C(\lambda)) \tag{2.8}$$

$$w_{r,i} = \psi(r_i(\lambda))/r_i(\lambda). \tag{2.9}$$

For each fixed λ, we obtain a preliminary estimate of (β,σ) by solving (2.6)-(2.7) using the Huber function (2.3) with K = 1.5; this is accomplished through a standard iteratively reweighted least squares algorithm. We then do one step of the algorithm towards the solution to (2.6)-(2.7) using the Hampel function (2.4) with a = 1.5, b = 3.0, c = 7.0. This defines for each fixed λ a value of the logarithm of the pseudo-likelihood (2.2) using the Hampel $\rho(\cdot)$ function. The result is maximized for λ between -1.0 and 1.0 by a grid search. Although the Hampel $\rho(\cdot)$ function is not convex, in all examples we have found a unique maximum on [-1,1].

The procedure, besides giving an estimator of λ, also yields a design diagnostic and a response diagnostic. The *design diagnostic* is simply a listing as an index plot of the squared weights $\{w_{x,i}^2\}$; we choose squares because they are more dramatic to the eye. The response diagnostic is a standard one. We can rewrite (2.6) into the more standard iteratively reweighted least squares format

$$\sum_{i=1}^{N} w_{r,i}\, x_i (Y_i^{(\lambda)} - x_i^T\beta) = 0 \ ,$$

where the *response diagnostic* is

$$w_{r,i} = \psi(r_i(\hat{\lambda}))/r_i(\hat{\lambda}) \ .$$

As we shall show in the examples, the estimate of λ, the design diagnostics, the response diagnostics, and the residuals from the robust fit provide a wealth of useful information.

2.4 Bounded influence estimators. None of the estimators introduced so far has a bounded influence function, though in the examples we have studied, Schweppe-type estimators handle outliers reasonably well. A bounded influence transformation [BIT(a)] estimator, depending on the constant a, can be constructed by a method applicable for general parametric estimation problems. The method is given by Stahel (1981) and has its origins in work by Hampel (1968, 1974). Let $\theta = (\beta,\sigma,\lambda)$ and let $f(y_i,x_i;\theta)$ be the density of y_i given x_i when θ obtains. Denote the gradient of the log-likelihood by $\ell(y_i,x_i;\theta) = \nabla_\theta \log f(y_i,x_i;\theta)$. Now let A be a matrix solving

$$A = n^{-1} \sum_{i=1}^{n} E_y\{w^2(y,x_i;\theta)[\ell(y,x_i;\theta)][\ell(y,x_i;\theta)]^t\}, \tag{2.7}$$

where

$$w(y,x;\theta) = \min\{1, a\sqrt{p+2}[(\ell(y,x;\theta)^t A^{-1}(\ell(y,x;\theta)]^{-\frac{1}{2}}\} \ , \tag{2.8}$$

$a > 1$, and of course $(p+2)$ is the number of parameters being estimated. Then $\hat{\theta}$ solves

$$\sum_{i=1}^{n} w(y_i,x_i;\hat{\theta})[\ell(y_i,x_i;\hat{\theta})] = 0. \tag{2.9}$$

The matrix A is a robust version of the second moment matrix of $\ell(y,x;\theta)$ and (2.8) shows that the weights are based upon a Mahalanobis-type distance

of $\ell(y_i,x_i;\hat{\theta})$ from the centroid of $\{\ell(y_i,x_i;\hat{\theta})\colon i=1,\ldots,n\}$. As $a \to \infty$, $w(y_i,x_i;\hat{\theta}) \to 1$ for all i and $\hat{\theta}$ converges to the MLE.

The influence function evaluated at (y_i,x_i) is

$$IF(y_i,x_i;\hat{\theta}) = B^{-1}w(y_i,x_i;\hat{\theta})\ell(y_i,x_i;\hat{\theta})$$

where

$$B = n^{-1}\sum_{i=1}^{n} E_y\{w^2(y,x_i;\hat{\theta})\ell(y,x_i;\hat{\theta})\ell^t(y,x_i;\hat{\theta})\} .$$

The asymptotic covariance matrix of $\hat{\theta}$ is $V = B^{-1}AB^{-1}$. A measure of influence motivated by Krasker and Welsch is

$$\gamma_2 = \max_i[IF(y_i,x_i;\hat{\theta})^t V^{-1} IF(y_i,x_i;\hat{\theta})]^{\frac{1}{2}} = \max_i[\ell(y_i,x_i;\hat{\theta})^t A^{-1}\ell(y_i,x_i;\hat{\theta})]^{\frac{1}{2}}$$

Stahel (1981) calls γ_2 the self-standardized gross-error sensitivity. "Self-standardized" refers to the fact that the influence function is being normed by the estimator's own asymptotic covariance matrix. By the construction of $\hat{\theta}$,

$$\gamma_2 \le a.$$

Note that $(\partial/\partial\beta)\ell(y_i,x_i;\theta)$ is linear in ε_i but $(\partial/\partial\sigma)\ell(y_i,x_i;\theta)$ is quadratic in ε_i. Thus as $|\varepsilon_i| \to \infty$, the weight $w(y_i,x_i;\theta)$ goes to zero sufficiently fast so that

$$w(y_i,x_i;\hat{\theta}) \,\|\, (\partial/\partial\beta)\ell(y_i,x_i;\theta) \,\| \to 0.$$

Therefore, as far as estimation of β is concerned, $\hat{\theta}$ is similar to regression M-estimates with a "psi function" which redescends to 0, and we can expect the BIT to be somewhat similar to the pseudo-MLE. For the regression problem, this fact was noted by Hampel (1978).

Since $(\partial/\partial\lambda)\ell(y_i,x_i;\theta)$ is a function of $\log(y_i)$, if y_i is either close to 0 or large, then the ith observation may have a high influence even if $|\varepsilon_i|$ is small and x_i has only small leverage. Among the estimators we consider, only the BIT bounds this source of influence. We define

$$w_{T,i} = w(y_i, x_i; \hat{\theta}) ,$$

which measures the downweighting of the ith observation due to this total influence being large.

Both θ and A were estimated by an iterative algorithm that was modeled on the Krasker-Welsch algorithm of Peters, Samarov, and Welsch (1982).

The weighting of $\ell(y_i, x_i; \theta)$ introduces a bias, which we suspect is appreciable only for σ. The bias could be removed (asymptotically) by subtracting a suitably defined constant vector α from $\ell(y_i, x_i; \theta)$. However, α depends on θ and we have not been able to find a convergent algorithm for estimating α, θ, and A.

3. Confidence Intervals for λ

Large-sample confidence intervals for λ can be constructed using either (i) the asymptotic normal distribution of $(\hat{\lambda}-\lambda)$ or (ii) modified likelihood-ratio tests.

Confidence intervals for λ were derived from likelihood-ratio tests in the original paper of Box and Cox (1964). A modification of their method is possible for the pseudo-MLE (or, more generally, when the estimator of λ is defined as the minimizer of a real-valued function). As discussed by Carroll (1980, 1982), let $L_{max}(\lambda)$ be the maximum over β and σ of (2.2) with λ held fixed. To text H_0: $\lambda=\lambda_0$, we use the test statistic

$$T(\lambda_0) = 2D[L_{max}(\hat{\lambda}) - L_{max}(\lambda_0)] ,$$

where the correction factor

$$D = \left(\sum_{i=1}^{N} \dot{\psi}(r_i/\sigma t_i)\xi_i^2 w_{x,i}/t_i\right)\left(\sum_{i=1}^{N} \psi_k^2(r_i/\sigma t_i)\xi_i^2 w_{x,i}\right)^{-1},$$

$$\xi_i = (\partial/\partial\lambda)(x_i^t\beta)^{(\lambda)}\,|\hat{\lambda},\hat{\beta}(\hat{\lambda}),$$

is introduced so that $T(\lambda_0)$ is asymptotically (as $n \to \infty$ and $\sigma \to 0$ simultaneously) chi-square with 1 d.f. when H_0 holds. See Bickel and Doksum (1981) for an introduction to "small-σ" asymptotics. Also see Schrader and Hettmansperger (1980) for a similar correction factor. The correction factor is based on small-σ theory, and we have found that this test will tend to be conservative when the noise is large relative to the signal, as occurs especially in the first example.

We saw no way to develop a likelihood-ratio-type test for BIT. But this estimator is approximately normally distributed and its covariance matrix can be estimated by $n^{-1}(\hat{B}^{-1}\hat{A}\hat{B}^{-1})$. $\hat{A}$ and $\hat{B}$ are estimated by using expectation with respect to the empirical distribution of (y_i,x_i) in the definitions of A and B. For example.

$$\hat{B} = n^{-1}\sum_{i=1}^{N} w(y_i,x_i;\hat{\theta})\ell(y_i,x_i;\hat{\theta})\ell^t(y_i,x_i;\hat{\theta}).$$

4. Examples

4.1 Introductory remarks. In this section we will apply the following estimators to three previously published data sets:

1) MLE = the maximum likelihood estimator

2) The pseudo-MLE of section 2.

3) BIT(1.5) = the bounded influence estimator with a = 1.5.

4) BIT(1.3) = the bounded influence estimator with a = 1.3.

When calculating the first two estimators and their confidence intervals, the parameter λ was restricted to lie between -1.0 and +1.0. When calculating BIT(1.5) and BIT(1.3), no restrictions were placed on λ.

4.2 Salinity data. These data were obtained during collaboration with members of the Curriculum in Marine Sciences at the University of North Carolina and were used in Ruppert and Carroll (1980) to illustrate several robust regression estimators. Atkinson (1983) used them to illustrate regression diagnostics and power transformations. The response is salinity or salt concentration (SAL) in North Carolina's Pamlico Sound measured biweekly during the spring seasons of years 1972 through 1977. The explanatory variables are salinity lagged by 2 weeks (SALLAG), TREND which is the number of biweekly periods since the beginning of the spring season, and the volume of river discharge into the sound (DSCHG). The data were a minor part of a larger study aimed at predicting brown shrimp harvest.

The salinity data is a seemingly manageable , yet extremely complex, data set. In the interest of clarity and for future researchers, it may be worthwhile to describe the physical background of the data. Water salinity SAL was measured at certain marshy areas located near the estuaries of the major rivers which empty into Pamlico Sound; these areas were chosen because this environment was thought to be of importance during the growth cycle of brown shrimp. The physical structure suggests that the salinity measurements in the nursery areas will exhibit some autocorrelation within years, and we found this correlation convenient to express through the lagged salinity SALLAG. River discharge DSCHG measures the volume of river water flowing into the sound. Originally, it was thought that the higher the river flow,

the more fresh water would diffuse into the brown shrimp breeding grounds, thus lowering salinity. Inspection of the data revealed periods of very heavy discharge, see for example data points #5, #16. The marine biologists working with us suggested that these large values of DSCHG represented a real change in the system, in that at some point the rivers would be flowing so fast that increases in DSCHG would not lead to much increased fresh water diffusion into the breeding areas, and hence past some point the effect of increased DSCHG would be negligible or possibly even *positive*. We handled this phenomenon in an ad hoc fashion by simply setting DSCHG equal to 26.0 whenever it actually exceeded 26.0. In Ruppert and Carroll (1980), large DSCHG values were left unchanged to obtain a data set with anomalous observations for testing robust procedures. As we shall see, much of the problem with analyzing these data can be traced to those points with recorded DSCHG $\geq$ 26.0. A linear time trend TREND was also added into the model because North Carolina usually warms quickly in spring, which suggests possible evaporation and increased salinity.

Incidentally, the salinity data analyzed here did not form a part of the final prediction model used by the North Carolina Department of Fisheries, although our experience gained in working with these data was helpful. The final model, based on only six years of data, was used blindly but successfully by the Fisheries Department for four years, 1978-1981.

There are an almost infinite variety of models which can be fit to the salinity data. We shall study only two such models. When transforming the response SAL it seems sensible to simultaneously transform the lagged response SALLAG. While we believe it is certainly permissible for illustrative purposes to consider models that only trans-

form the response SAL and not the lagged response, it is important to keep in mind some reasonable alternatives.

The bivariate relationship between SAL and SALLAG is fairly linear with no apparent outliers; see Figure 4.1. A scatter diagram (not shown here) of SAL and TREND suggests a possible weak linear or quadratic relationship. Figure 4.2 is a scatter diagram of the residual of SAL regressed on SALLAG plotted against DSCHG. DSCHG is distributed with a noticeable positive skewness. The rather linear relationship between these residuals and DSCHG when DSCHG is below about 26 seems to break down for larger values of DSCHG. Observations #5 and #16 have much higher values of SAL than would be expected if the linear relationship held across all values of DSCHG. These observations also are high leverage points.

Observation #3 is also somewhat unusual. It has the third highest value of DSCHG, very low values of SAL and SALLAG, and the lowest possible value of TREND. However, in regard to the relationship between SAL and SALLAG, TREND, and DSCHG, #3 does conform reasonably well with the bulk of the data. In Figure 4.2 cases #9, #15 and #17 appear as somewhat outlying.

Atkinson (1983) uses these data to illustrate diagnostic methods. He considers what we shall call the BASIC model, wherein a transformed value of water salinity SAL is regressed linearly on lagged salinity, river discharge and TREND. He first modifies #16 from 33.443 to 23.443, leading to $\hat{\lambda} = 0.46$. Using a "constructed variable", he finds that observation #3 is highly influential for estimating λ, since after deleting it we obtain $\hat{\lambda} = -0.15$. With #16 corrected and #3 deleted, Atkinson suggests the log transformation. The entire process is effectively the same as having deleted both #3 and #16.

Despite the fact that the BASIC model does not transform the predictor lagged salinity, we will illustrate our methods on this model so as to compare them with Atkinson's. In Table 4.1 we exhibit details of estimates and diagnostics. The pseudo-MLE suggests that it is #5 and #16 which do not fit the data well, while #3 is an unusual design point; in light of Atkinson's analysis, the emergence of #5 as possibly influential strongly suggests masking. Somewhat more expectedly the BIT methods identify #16 as an outlier with #3 and then #5 as possibilities.

The results of our analysis indicate the possibility that the effect of observation #5 is being masked by #3 and #16. That this is the case is illustrated by Table 4.2, where we array the various estimates of λ having deleted the combinations (3,16) (5,16) and (3,5,16). Observation #5 is highly influential for the MLE of λ after #3 and #16 are deleted. This is classic case of masking, one of the better examples we have seen with real data.

The Cook and Wang LD transformation diagnostic is not immune to the masking in #5. For the full data, their diagnostic clearly identifies #3 and #16 as potential problems, but no hint is given that #5 might be unusual. After deleting #3 and #16, we recalculated the diagnostic, at which point #5 becomes an obvious candidate for special treatment.

The analysis of the BASIC model shows that in using diagnostics such as Atkinson's or Cook and Wang's to see if a point has a potentially large impact on the MLE for the transformation parameter λ, the careful analyst must be concerned with masking and should recompute the diagnostics after case deletion. While we do not claim that our methods, and especially our estimates, should be used blindly and will be immune to masking, we do hope that this example illustrates that robust methods

can provide valuable and non-redundant information to the data analyst.

We now turn to a model in which lagged salinity is also transformed; this is the TRANSFORMED model

$$(4.1) \qquad SAL^{(\lambda)} \doteq \beta_0 + \beta_1 SALLAG^{(\lambda)} + \beta_2 TREND + \beta_3 DSCHG.$$

As discussed previously, graphical methods and Figure 4.2 suggest that it is #5 and #16 that are very unusual, #9, #15, #17 somewhat less so, and that in fact #3 conforms reasonably well with the bulk of the data. To some extent this is confirmed by the robust method analyses, which are given in Table 4.3. Certainly #5, #16 are real outliers, but #3 appears to fit the data reasonably well, although it is an unusual design point. Observation #3 only becomes an outlier if one attempts to transform SAL but not SALLAG.

At this point in time analogues to the LD diagnostic of Cook and Wang have not been developed for the TRANSFORMED model (4.1), although it was very easy to modify our programs to construct the robust estimates and diagnostics for this new model. It is interesting to study Table 4.3, where observation #3 has an enormous effect on the estimate of λ only when #5 is included in the model. This fact may serve as a test for future diagnostics applied to model (4.1).

As stated at the outset, the salinity data are complex. For example, we have performed residual analyses on each of the models presented here, and the reader may wish to verify why we are not entirely satisfied with the results. Other analyses may also be contemplated; we have used our techniques on a model segmented at river discharge = 25, a model in which lagged river discharge is used rather than simple river discharge, and others. We have chosen the two models in this section not because they are the best possible, but rather because they are simple and

illustrate that our robust estimates and diagnostics can be used as an informative tool and not merely as a black box.

4.3 Stack loss data. This is an often analyzed data set; see among others Brownlee (1965), Daniel and Wood (1980), Andrews (1974), Dempster and Gasko-Green (1981). Observations 1, 3, 4 and 21 have been identified as outliers by these authors, with the second and fourth authors suggesting that observation 2 may also be outlying. Atkinson (1982) found that the MLE of λ is 0.30 for the first order model in the three explanatory variables with all observations included, while if #21 is excluded then the MLE is $\hat{\lambda} = 0.48$. Atkinson also considered a second order model.

We have only considered the first order model. The robust estimators and their associated diagnostics are given in Table 4.4; the estimators agree among themselves and with Atkinson that λ about $\frac{1}{2}$ is reasonable. The pseudo-MLE diagnostics indicate that observations #2, #4 and #21 are not well fit by the model, with #1, #17 also being unusual design points. The BIT methods suggest that #2, #4 and #21 are very poorly fit by the model, and suggest that #3, #16 are worth further investigation. If for each observation we evaluate the derivative of the log-likelihood with respect to λ evaluated at the BIT (1.3) estimate, then #2, #3, #4 and #21 stand out; their absolute values all exceed 64.0, while no other exceeds 38.0.

The Cook and Wang LD diagnostic originally identifies #21 as possibly interesting; although LD(#21) is only about 0.70 it still stands out. After removing #21, observation #16 now stands out, but LD(#16) is extremely small being only about 0.25. We think it would be unusual to continue the diagnostic-deletion process. At this point the diagnostic process would probably delete #21, take $\lambda = \frac{1}{2}$ and begin ordinary regression

diagnostic-deletion.

All the methods identified #21 as having the biggest potential for trouble. The methods we have introduced also identify potential difficulties with #2, #3 and #4. Faced with the conclusion that at least 4 of 21 data points do not fit a model, the analyst might reasonably question either the model or the data. One could attempt to accommodate the first 4 data points by changing the model, as many previous authors have done by suggesting "second order" models. Alternatively, following Daniel and Wood, one might question whether the first 4 observations are really valid data points.

We believe that for the first order model in the stack loss data, the robust estimates and diagnostics we have introduced provide valuable information. Combining previous diagnostics with our methods appears to have been quite successful.

4.4 Artificial data of Cook and Wang. This is an eleven case data set created by Cook and Wang (1983). These artificial data have the property that the model $\log y_i = x_i + \varepsilon_i$ fits the first ten cases, but the model $Ey = \beta_0 + \beta_1 x$ fits the complete data. Cook and Wang's LD diagnostic identifies #11 as particularly influential, because it is unusual in both the design and the response.

The various estimates of the transformation parameter λ are reproduced in Table (4.5), both with and without Case #11. Clearly, the MLE is dramatically affected by #11, as are the BIT estimates. The pseudo-MLE method downweights #11 sufficiently as to effectively exclude it from the analysis. The fact that the pseudo-MLE automatically downweights any unusual design point is the property which makes it an effective estimation technique in this example.

The pseudo-MLE and BIT methods also contain useful diagnostic information. An index plot or simply the numbers listed for the complete data tell us through the BIT methods that both #10 and #11 are potentially interesting, since the other nine case weights equal one. The pseudo-MLE suggests that both #10 and #11 are unusual design points, but that #11 is a response outlier as well. The reduced data analyses support this conclusion.

Cook and Wang conclude their analysis after deleting #11, having noted that when applied to the complete data their LD diagnostic identifies only #11 as potentially interesting. Of course, they have the somewhat unusual advantage of knowing the model which is intended to fit the data. In practice we would be faced with the problem of whether to repeat the diagnostic process. Because of the potential for masking found in the salinity data, for illustrative purposes we repeated the Cook and Wang diagnostic analysis on the reduced data set, without #11.

For the reduced data without #11, the Cook and Wang LD diagnostic strongly suggests further study of #10, since LD(#10) $\doteq$ 3.3 and all other LD statistics are below 1.2. There is a potential masking effect here in the LD statistics: #10 was not seen to be potentially influential for the complete data set. If we were to delete #10 as well as #11, the MLE becomes $\hat{\lambda} = 0.15$ with an associated 95% confidence interval which includes [-1.0, 1.0]. In combination with the BIT and pseudo-MLE analyses, the last confidence interval suggests that #11 is a design and response outlier, while #10 is a somewhat unusual design point that contains information about the transformation parameter λ.

It is probably not useful to draw too strong a conclusion from a small, specially constructed artificial data set. It seems clear from this example that no single method will suffice. It seems most sensible to

combine the information given from robust estimates, their associated diagnostics, specially constructed diagnostics such as Cook and Wang's LD statistic, and standard procedures such as the confidence intervals.

5. Conclusions

The classic work is robust estimation dealt with location parameters and was concerned with heavy-tailed deviations from the normal distribution. Earlier work on robust regression, e.g. Huber (1973), also concentrated on distributional robustness. However, regression models are at best only approximately correct, and they easily can break down at the extreme points of the factor space. Therefore, it was natural that techniques, such as the Krasker-Welsch (1982) estimator, that bound the influence of such points would be developed for linear models.

The use of transformations is a powerful technique for model building, since transformed data will often fit a simpler model than the raw data. As when using standard (untransformed) regression models, here too, one needs to guard against the possibility that one or a few extreme points in the factor space will overwhelmingly affect the data analysis, especially the choice of a transformation.

We have introduced several estimation techniques for the Box-Cox (1964) transformation model that, unlike the MLE, are relatively insensitive to outliers in both the design and in the residual. In several examples, these estimators proved valuable not only at the final estimation stage of the analysis, but also at the preliminary model identification stage. For model selection purposes, the robust estimators should be viewed as useful supplements to the diagnostic techniques of Atkinson (1982, 1983) and Cook and Wang (1983).

Model selection, the identification of outliers, and the choosing of a transformation must be done simultaneously. Also, the masking of some outliers by other outliers is a potential problem. For these reasons, at least, diagnostic information from as many sources as feasible should be used.

REFERENCES

ANDREWS, D.F. (1974). A robust method for multiple linear regression. Technometrics, 16, 523-531.

ATKINSON, A.C.. (1981). Two graphical displays for outlying and influential observations in regression. Biometrika, 68. 13-20.

ATKINSON, A.C. (1982). Regression diagnostics, transformations and constructed variables (with discussion). JRSS-B, 44, 1-36.

ATKINSON, A.C. (1983). Diagnostic regression analysis and shifted power transformations. Technometrics, 25, 23-33.

BELSLEY, D.A., KUH, E., and WELSCH, R.E. (1980). Regression Diagnostics. New York: Wiley.

BICKEL, P.J., and DOKSUM, K.A. (1981). An analysis of transformation revisited. J. Am. Statist. Assoc., 76, 296-311

BOX, G.E.P. and COX, D.R. (1964). An analysis of transformations (with discussion). J. Roy. Statist. Soc., Ser. B, 26, 211-246.

BROWNLEE, K.A. (1965). Statistical Theory and Methodology in Science and Engineering. New York: Wiley.

CARROLL, R.J. (1980). A robust method for testing transformations to achieve approximate normality. J. Roy. Statist. Soc., Series B, 42, 71-78.

CARROLL, R.J. (1982). Two examples of transformations when there are possible outliers. Applied Statistics, 31, 149-152.

COOK, R.D. and WANG, P.C. (1983). Transformation and influential cases in regression. Technometrics, 25, 337-343.

COOK, R.D. and WEISBERG, S. (1982). Residuals and Influence in Regression. New York and London. Chapman and Hall.

DANIEL, C. and WOOD, F.S. (1980). Fitting Equations to Data, 2nd Edition. New York: Wiley.

DEMPSTER, A.P., and GASKO-GREEN, M. (1981). New tools for residual analysis. Ann. Statist., 9, 945-959.

HAMPEL, F.R. (1968). Contributions to the Theory of Robust Estimation. Ph.D. Thesis. University of California, Berkeley.

HAMPEL, F.R. (1974). The influence curve and its role in robust estimation. J. Amer. Statist. Assoc., 62. 1179-1186.

HAMPEL, F.R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. 1978 Proceedings of the ASA Statistical Computing Section. ASA, Washington, D.C., 59-64.

HOCKING, R.R. (1983). Developments in linear regression methodology: 1959-1982 (with discussion). Technometrics, 25, 219-249.

HUBER, P.J. (1964). Robust estimation of a location parameter. Ann. Math. Statist., 35, 73-101.

HUBER, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Statist., 5, 799-821.

HUBER, P.J. (1983). Minimax Aspects of bounded-influence regression. J. Am. Statist. Assoc., 78, 66-80.

KRASKER, W.S. (1980). Estimation in linear regression models with disparate data points. Econometrica, 48, 1333-1346.

KRASKER, W.S. and WELSCH, R.E. (1982). Efficient bounded-influence regression estimation. J. Am. Statist. Assoc., 77, 595-604.

PETERS, S.C., SAMAROV, A.M., and WELSCH, R.E. (1982). Computational procedures for bounded-influence and robust regression. Center for Computational Research in Economics and Management Science. Alfred P. Sloan School of Management Science. MIT.

RUPPERT, D. and CARROLL, R.J. (1980). Trimmed least squares estimation in the linear model. J. Am. Statist. Assoc., 75, 828-838.

SCHRADER, R.M. and HETTMANSPERGER, T.P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. Biometrika, 67, 93-101.

SNEE, R.D. (1983). Discussion of "Developments in linear regression methodology: 1959-1982" by R.R. Hocking. Technometrics, 25, 230-237.

STAHEL, W.A. (1981). Robuste Schaetzungen: Infinitesimale Optimalitaet und Schaetzungen von Kovarianzmatrizen. Ph.D. Dissertation. Swiss Federal Institute of Technology, Zurich.

Table 4.1 Analysis of transformations for the BASIC salinity model using all the data. The 95% confidence intervals and standard errors are as described in the text, as are the case weights.

Method	Estimate of λ	95% C.I.	Standard Error		Weights for Selected Cases #3	#5	#9	#15	#16	#17
MLE	0.97	(0.17,1.00)	0.75							
Pseudo-MLE	0.69	(-0.64,1.00)	--	Residual	1.00	0.00	0.82	0.55	0.00	0.71
				Design	0.40	0.09	0.72	1.00	0.02	1.00
BIT(1.5)	0.51	--	0.36		0.43	0.51	0.46	0.55	0.04	0.89
BIT(1.3)	0.51	--	0.27		0.17	0.24	0.28	0.26	0.03	0.46

Table 4.2 Analysis of transformations for the TRANSFORMED salinity model (4.1) using all the data. Note here that both salinity and lagged salinity are transformed.

Method	Estimate of λ	95% C.I.	Standard Error		Weights for Selected Cases #3	#5	#9	#15	#16	#17
MLE	0.66	(-0.16,1.00)	0.46							
Pseudo-MLE	1.00	(-0.72,1.00)	--	Residual	1.00	0.00	0.84	0.52	0.00	0.69
				Design	0.40	0.09	0.72	1.00	0.02	1.00
BIT(1.5)	0.94		0.41		1.00	0.33	0.42	0.42	0.06	0.53
BIT(1.3)	1.06		0.41		0.95	0.14	0.30	0.23	0.03	0.33

Table 4.3 Estimates of λ for selected observations excluded. The BASIC and TRANSFORMED models are described in the text; the latter transforms the predictor lagged salinity, while the former does not.

	MODEL									
Estimator	BASIC					TRANSFORMED				
Observations deleted	None	16	3,16	5,16	3,5,16	None	16	3,16	5,16	3,5,16
MLE	.97	.46	-.15	.70	.39	.66	.50	.22	1.0	1.0
Pseudo-MLE	.69	.71	.22	.79	.40	1.0	1.0	1.0	1.0	1.0
BIT(1.3)	.51	.50	.25	.80	.45	1.06	1.09	1.06	1.16	1.19

Table 4.4 Estimates of λ for the complete stackloss data. A first order model was used.

Method	Estimate of λ	95% C.I.	Standard Error		Weights for Selected Cases #1	#2	#3	#4	#21
MLE	0.30	(-0.18,0.74)	0.38						
Pseudo-MLE	0.49	(0.13,0.74)	--	Residual	1.00	0.17	1.00	0.06	0.00
				Design	0.50	0.45	1.00	1.00	0.72
BIT(1.5)	0.39	--	0.29		1.00	0.72	1.00	0.47	0.27
BIT(1.3)	0.41	--	0.23		1.00	0.37	0.88	0.25	0.13

Table 4.5 Estimates of λ for the Data of Cook and Wang.

Estimator	$\hat{\lambda}$	95% CI	SE	Weights for Selected Cases		
				#10		#11
		Complete Data				
MLE	0.71	(0.04,1.00)	--			
Pseudo-MLE	-0.14	(-0.66,0.88)	--	1.00 0.39	Residual Design	0.00 0.01
BIT(1.3)	0.49	----	0.30	0.31		0.21
		Reduced Data (Case #11 Excluded)				
MLE	-0.18	(-1.00,0.57)	--	--		--
Pseudo-MLE	-0.25	(-0.72,0.86)	--	1.00 0.23	Residual Design	--
BIT(1.3)	-0.34	----	0.90	1.00		--

List of Figures

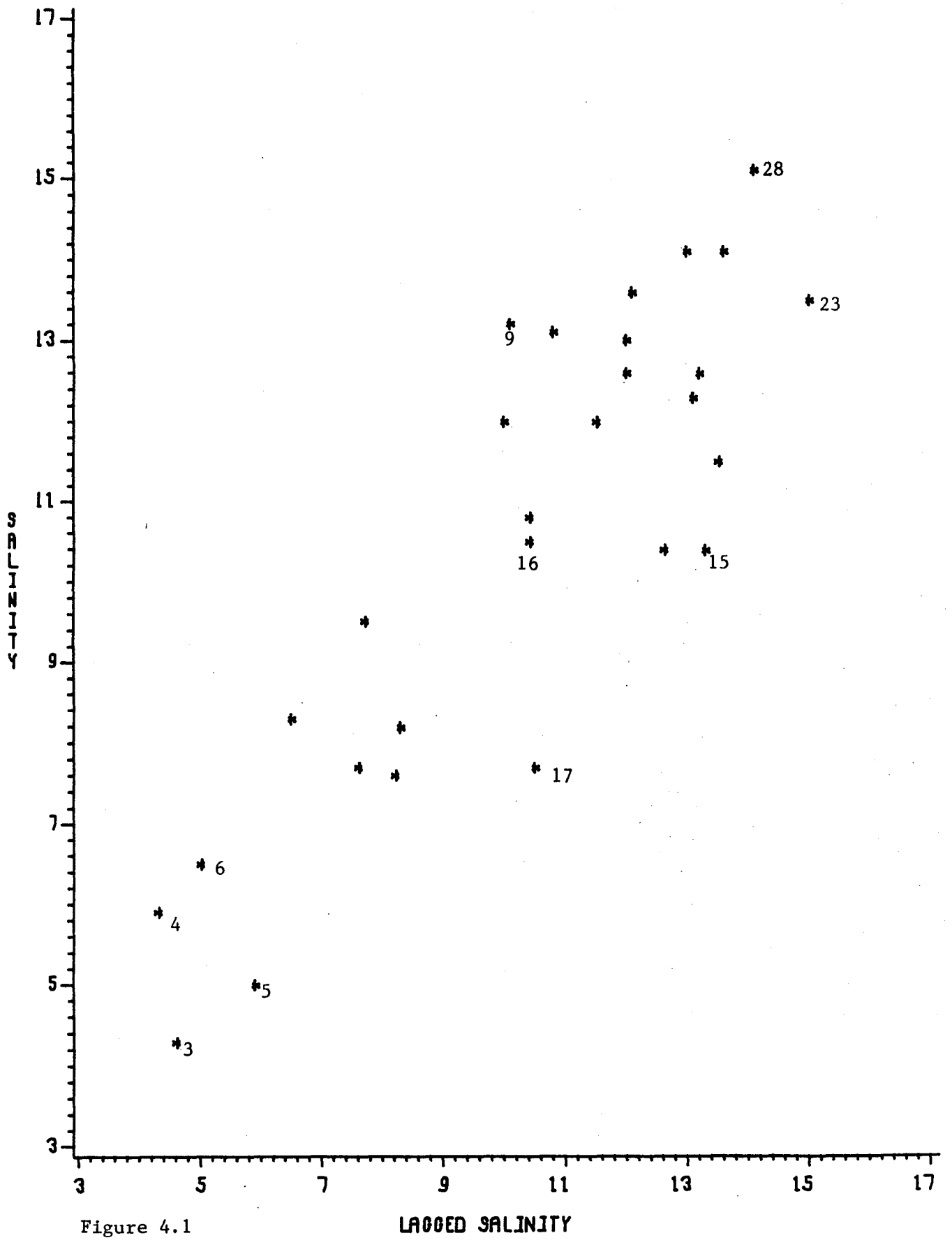
17
15
13
11
9
7
5
3
SALINITY
3
5
7
9
11
13
15
17
LAGGED SALINITY
28
23
9
16
15
17
6
4
5
3

Figure 4.1

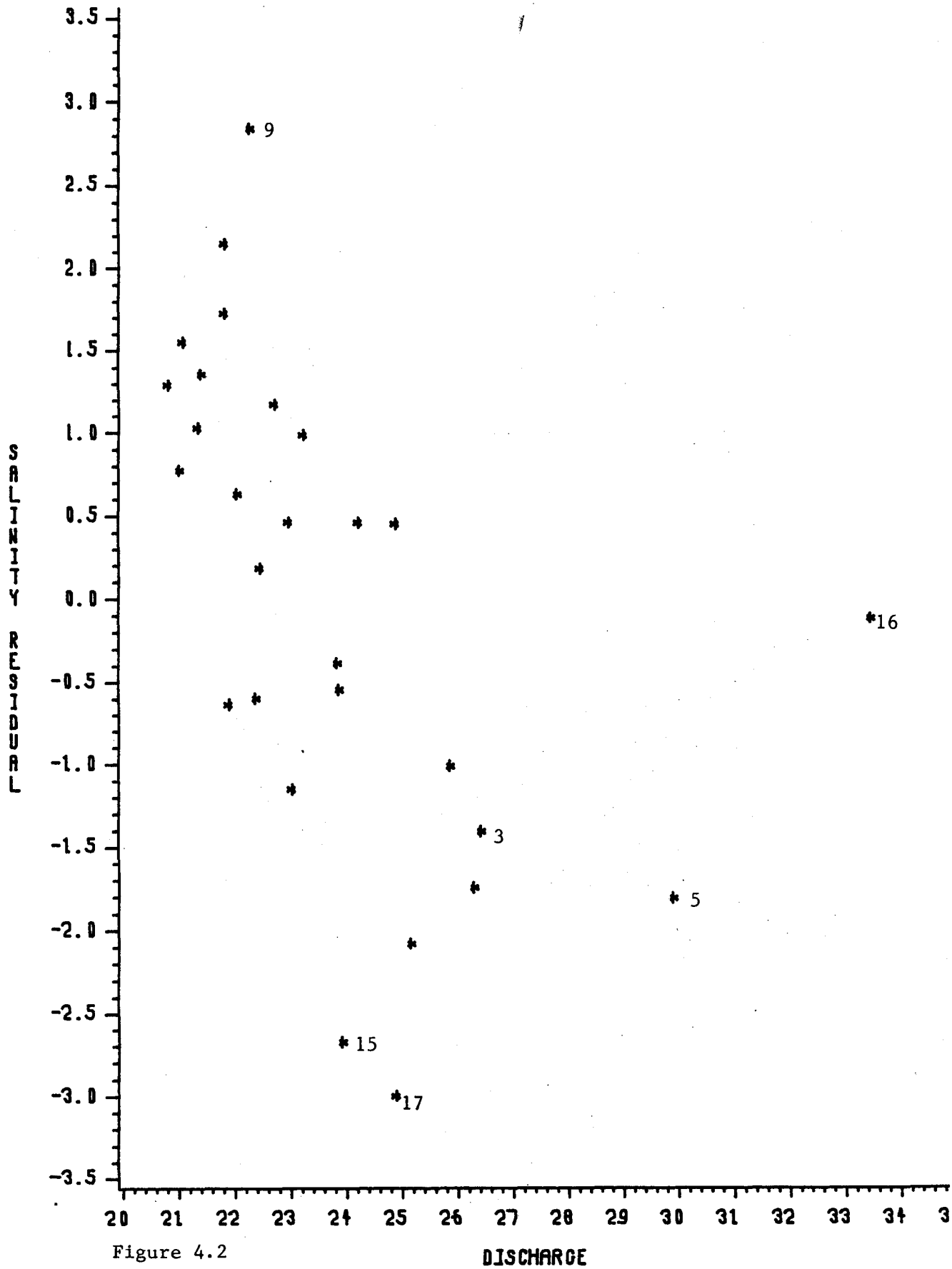
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
SALINITY RESIDUAL
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
DISCHARGE
9
16
3
5
15
17

Figure 4.2