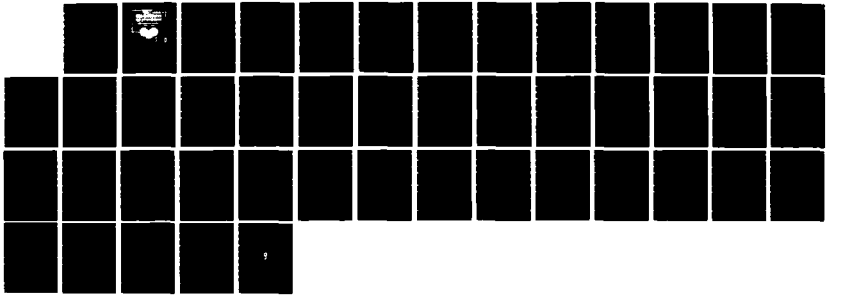
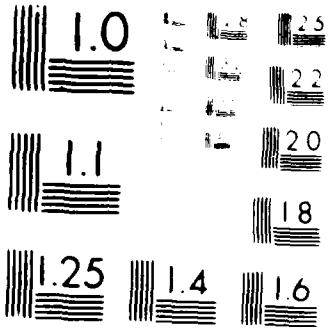
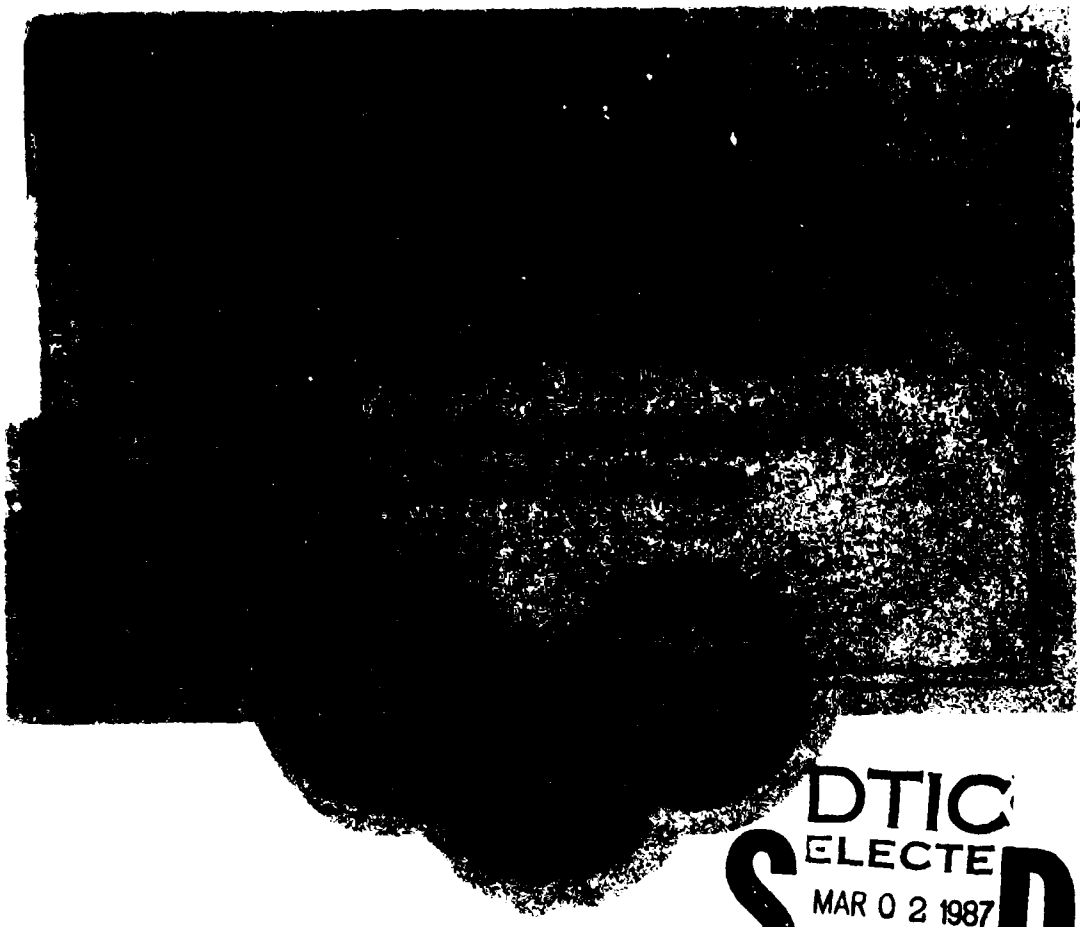
AD-A177 531 DIAGNOSTICS AND ROBUST ESTIMATION WHEN TRANSFORMING THE REGRESSION MODEL A.. (U) NORTH CAROLINA UNIV AT CHAPEL HILL INST OF STATISTICS R J CARROLL ET AL. OCT 86 1/1

UNCLASSIFIED MIMEO-SER-1706 AFOSR-TR-87-0266 F/G 12/1 NL




1.0
1.1
1.25
1.4
1.6
2.2
2.0
1.8

(2)



DIAGNOSTICS AND ROBUST ESTIMATION WHEN TRANSFORMING THE REGRESSION MODEL AND THE RESPONSE

R.J. CARROLL and D. RUPPERT

October 1986

Mimeo Series #1706

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

AD-A177 531

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b. RESTRICTIVE MARKINGS:
2a. SECURITY CLASSIFICATION AUTHORITY:	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE:	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): Mimeo Series #1706	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 87-0266
6a. NAME OF PERFORMING ORGANIZATION: Univ. of NC-Chapel Hill; 6b. OFFICE SYMBOL (If applicable):	7a. NAME OF MONITORING ORGANIZATION: Air Force Office of Scientific Research
6c. ADDRESS (City, State and ZIP Code): University of NC, Statistics Department, Phillips Hall, Chapel Hill, NC 27514	7b. ADDRESS (City, State and ZIP Code):
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR; 8b. OFFICE SYMBOL (If applicable):	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: ~~AFOSR~~-F-49620-85-C-0144
8c. ADDRESS (City, State and ZIP Code): Bolling Air Force Base, Washington, DC 20332	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2304	A5	

11. TITLE (Include Security Classification): "Diagnostics and Robust Estimation when Transforming the Regression Model and the Response"

12. PERSONAL AUTHOR(S): Carroll, R.J. and Ruppert, David

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
technical	FROM 8/86 TO 8/87	October 1986	40

16. SUPPLEMENTARY NOTATION:

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	regression analysis; heteroscedasticity; transformations; maximum likelihood.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

In regression analysis, the response is often transformed to remove heteroscedasticity and/or skewness. When a model already exists for the untransformed response, then it can be preserved by applying the same transform to both the model and the response. This methodology, which we call "transform both sides" has been applied in several recent papers, and appears highly useful in practice. When a parametric transformation family such as the power transformations is used, then the transformation can be estimated by maximum likelihood. The MLE, however, is very sensitive to outliers. In this article, we propose diagnostics to indicate cases influential for the transformation or regression parameters. We also propose a robust bounded-influence estimator similar to the Krasker-Welsch regression estimator.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS ☐	

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code)	22c. OFFICE SYMBOL
Lisa Brooks	(919) 962-2307	NM

DD FORM 1473, 83 APR — EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

Diagnostics and Robust Estimation When Transforming The Regression Model and The Response

Revised: September 1986

R. J. Carroll and David Ruppert

Department of Statistics

University of North Carolina

Chapel Hill, N. C. 27514

Abstract: In regression analysis, the response is often transformed to remove heteroscedasticity and/or skewness. When a model already exists for the untransformed response, then it can be preserved by applying the same transform to both the model and the response. This methodology, which we call "transform both sides" has been applied in several recent papers, and appears highly useful in practice. When a parametric transformation family such as the power transformations is used, then the transformation can be estimated by maximum likelihood. The MLE, however, is very sensitive to outliers. In this article, we propose diagnostics to indicate cases influential for the transformation or

regression parameters. We also propose a robust bounded-influence estimator similar to the Krasker-Welsch regression estimator.

Acknowledgement

The research of Professor Carroll was supported in full and the research of Professor Ruppert was supported in part by Air Force Office of Scientific research Contract AFOSR-S-49620-85-C-0144. Professor Ruppert's research was also partially supported by the National Science Foundation Grant DMS-8400602. We thank the referees for their helpful criticism and the encouragement to revise this article. We also thank Brian Aldershof for producing the figures.

1. Introduction

In regression analysis, the response y is often transformed for two distinct purposes, to induce normally distributed, homoscedastic errors and to improve the fit to some simple model involving an explanatory variable x. In many situations, however, y is already believed to fit a known model $f(x,\beta)$, β being a p-dimensional parameter. If a transformation of y is still needed to remove skewness and/or heteroscedasticity, then this model can be preserved by transforming y and $f(x,\beta)$ in the same manner. Specifically, let $y^{(\lambda)}$ be a transformation indexed by the parameter λ and assume that for some value of λ

$$y_i^{(\lambda)} = f^{(\lambda)}(x,\beta) + \sigma\epsilon_i \tag{1}$$

where $\epsilon_1,\ldots,\epsilon_N$ are independent and at least approximately normally distributed with variance 1. Notice the difference between (1) and the usual approach of transforming only the response, not $f(x,\beta)$, i. e.,

$$y^{(\lambda)} = f(x,\beta) + \sigma\epsilon_i . \tag{2}$$

It should be emphasized that model (1) is not intended as a substitute for (2). Both models are appropriate, but under quite different circumstances. Model (2) has been amply discussed by Box and Cox (1964) and others, e. g., Draper and Smith (1980), Cook and Weisberg (1982) and Carroll and Ruppert (1981).

Typically, in model (2), $f(x,\beta)$ is linear but in principle nonlinear models can be used. Model (1), which we call "transform both sides", has been investigated by Carroll and Ruppert (1984), Snee (1986), and Ruppert and Carroll (1986) and we will only summarize those discussions. According to (1), $f(x,\beta)$ has two closely related interpretations; $f(x,\beta)$ is the value of y when the error is zero and it is the median of the conditional distribution of y given x. In Carroll and Ruppert (1984), we were concerned with situations where a physical or biological model provides $f(x,\beta)$, but where the error structure is a priori unknown. Examples by Snee (1986), Carroll and Ruppert (1984), Ruppert and Carroll (1985), and Bates, Wolf, and Watts (1985) show that transforming both sides can be highly effective with real data, both when a theoretical model is available and, as Snee shows, when $f(x,\beta)$ is obtained empirically.

By estimating λ, σ, and β simultaneously, rather than simply fitting the original response y to $f(x,\beta)$, we achieve two purposes. First, β is estimated efficiently and therefore we obtain an efficient estimate of the conditional median of y. Second, we model the entire condtional distribution of y given x, and, in particular, we have a model which can account for the skewness and heteroscedasticity in the data. Carroll and Ruppert (1984) discuss the importance of modeling the conditional distribution of y in a special case, a spawner-recruit analysis of the Atlantic menhaden population. In section 6 we discuss the estimation of the conditional mean and conditional quantiles of y given x.

Many data sets we have examined have had severe outliers in the untransformed response y, but not in the residuals $e_i(\hat{\beta},\hat{\lambda}) = [y^{(\hat{\lambda})} - f^{(\hat{\lambda})}(x,\hat{\beta})]$; the transformation has accommodated, or explained, the outlying y's. There is still the danger, however, that a few outliers in y can greatly affect $\hat{\beta}$ and $\hat{\lambda}$. Outliers should not be automatically deleted or downweighted, especially when they appear to be part of the normal variation in the response, but it should be standard practice to detect and scrutinize influential cases.

When influential cases are present and they have an unacceptable effect on the MLE, then the best remedy is not simply to delete these cases but rather to apply a robust estimator. There are several reasons why this is so. First, as Hampel (1985) illustrates, robust estimates are generally somewhat more efficient than outlier rejection along with a classical estimator such as the MLE. Moreover, outlier rejection affects the sampling distribution of classical methods in ways that have not been fully studied. In contrast, the large-sample distribution of most robust estimators, in particular M-estimators which will be used here, can be easily calculated.

In this paper we propose a case-deletion diagnostic and a "bounded-influence" estimator.

Case deletion diagnostics for linear regression are discussed in Belsley, Kuh, and Welsch (1980) and Cook and Weisberg (1982), and have been extended to the response transformation model (2) by Cook and Wang (1983) and Atkinson (1986). The last two papers approximate the change in $\hat{\lambda}$ as single cases or subsets of cases are deleted.

Another approach to influence diagnostics is measuring changes in the statistical analysis under infinitesimal perturbations in the model. Cook (1986) gives an introduction to this theory, which he calls "local influence". We will not consider local influence for transformation models, but this seems a promising area for research.

When the response transformation model is used, $\hat{\beta}$ depends heavily on $\hat{\lambda}$, and it seems better to examine influence for $\hat{\beta}$ only after $\hat{\lambda}$ has been determined. In constrast, under the "transform both sides" model $\hat{\beta}$ and $\hat{\lambda}$ are only weakly related, with $\hat{\beta}$ determined by the median of the untransformed response, and $\hat{\lambda}$ determined by the skewness and heteroscedasticity. Therefore, influence for $\hat{\beta}$ and $\hat{\lambda}$ can be treating simultaneously.

For the "transform both sides" model we propose two approximations to the changes in $\theta = (\beta^T, \lambda)^T$ as cases are deleted. As shown in the next section, the MLE can be found as the least-squares estimate of a certain "pseudo-model". Both approximations start by linearizing this pseudo-model around the full-data estimate, and for each case take one step of an iterative procedure for finding the estimate without that case. The first approximation takes one step of the Newton-Raphson procedure, in effect using an accurate approximation to the Hessian matrix of the sum of squares for the pseudo-model. The second approximation is based on Atkinson's "quick estimate", and is equivalent to using one step of the Gauss-Newton rather than the Newton-Raphson algorithm. It is considerably less accurate than the first, but is more easily implemented on standard software and is useful for diagnostic purposes.

Subset deletion is simple in theory but can be unwieldly in practice because of the large number of possible subsets. If influential subsets are to be detected, some strategy is needed to search for them. An alternative to subset deletion and a good supplement to single case deletion diagnostics is to compare the fit of a highly robust estimator with that of the MLE. In the example of section 6, a robust estimator reveals an observation whose influence was masked by another observation and could not be detected by single case-deletion diagnostics.

Bounded-influence estimators place a bound on the influence function of each observation. Bounded-influence regression estimators have been proposed by Krasker (1980), Hampel (1978), and Krasker and Welsch (1982). The last paper and Hampel et al. (1986) provide a good overview.

Huber (1983) has questioned the need for bounded-influence estimators. They appear to be based on the pessimistic philosophy that nature will place response outliers precisely where they can do the most harm, on the high leverage points.

We disagree with Huber and feel that such pessimism is justified. Apparent response outliers are often due to gross errors in the measurement or recording of the explanatory variables x. In addition, response outliers can be caused by model breakdown. Both an incorrect model for the median response or an incorrect specification of the variance function, say a constant variance model where the variance actually depends on the median or mean, will lead to response outliers.

In the latter case, the y may be outlying only relative to the assumed variance, not the true variance, but this is enough to cause problems. For all these reasons, outlying y values are more likely at unusual x values. Huber's objection to bounded-influence estimation follows logically from his model that the errors are identically distributed with a heavy-tailed density, but this model is unrealistic in many modeling situations.

Moreover, even if one accepts Huber's conclusions about regression modeling, they apply only when there is no need to estimate a transformation parameter. The analog of leverage for λ is the derivative of the residual $e_i(\theta)$ with respect to λ. A response outlier will usually make this derivative large. Therefore, response outliers can induce high leverage points in transformation models even at x's that are in no way unusual.

Carroll and Ruppert (1985) proposed a bounded-influence transformation (BIT) estimator extending the Krasker-Welsch estimator to the response transformation model (2). In this paper we adapt this estimator to the "transform both sides" model. We also discuss computational aspects and propose a simple one-step estimator that can be implemented on standard software packages such as SAS.

2. Weighted Maximum Likelihood Estimation

All estimators used in this paper are found by maximizing a weighted log-likelihood. When the weights are identically 1, then the

estimator is the MLE. The robust estimators introduced in section 4 are weighted MLE's with the weights less than 1 for influential cases and equal to 1 otherwise. Let $w_1,\ldots,w_N$ be fixed weights. For now, they will depend on the y's but not the parameter $\theta = (\beta^T, \lambda)^T$. In section 4 the weights will depend on θ, but θ will be fixed at a preliminary estimate $\hat{\theta}_p$.

Throughout this paper $y^{(\lambda)}$ is the modified power transformation family used by Box and Cox (1964);

$$y^{(\lambda)} = (y^{\lambda} - 1)/\lambda \text{ if } \lambda \neq 0,$$
$$= \log(y) \text{ if } \lambda = 0.$$

Our analysis will be conditional on the observed x's. This is appropriate both for fixed and random x's. Let $\theta = (\beta^T,\lambda)^T$ be the vector of the transformation and regression parameters, and let $g_i(y_i,\theta,\sigma)$ be the conditional density of y_i given x_i. The log-likelihood for y_i is

$$\ell_i(y_i,\theta,\sigma) = \log g_i(y_i,\theta,\sigma) =$$

$$-(1/2)\log(2\pi\sigma^2) + (\lambda - 1)\log(y_i) - (2\sigma^2)^{-1}[e_i(\theta)]^2,$$

where

$$e_i(\theta) = y_i^{(\lambda)} - f^{(\lambda)}(x_i,\beta).$$

The weighted log-likelihood for $y_1,\ldots,y_N$ is

$$L(\theta,\sigma) = \Sigma_{i=1}^{N} w_i \ell_i(y_i,\theta,\sigma). \tag{3}$$

For fixed θ

$$\hat{\sigma}^2(\theta) = \{\Sigma_{i=1}^{N} w_i (e_i(\theta))^2\} / \{\Sigma_{i=1}^{N} w_i\}$$

maximizes $L(\theta,\sigma)$ over θ. Let $\dot{y}$ be the weighted geometric mean of $y_1,\ldots,y_N$ defined by

$$\log(\dot{y}) = \{\Sigma_{i=1}^{N} w_i \log(y_i)\} / \{\Sigma_{i=1}^{N} w_i\}.$$

The weighted MLE of θ maximizes

$$L_{max}(\theta) = L(\theta,\sigma(\theta)) =$$

$$\Sigma_{i=1}^{N} w_i \{-(1/2)\log(2\pi\hat{\sigma}^2) + (\lambda - 1)\log(y_i) - 1/2\} = \tag{4}$$

$$-(1/2)\,\Sigma_{i=1}^{N} w_i \{\log[\Sigma_{i=1}^{N} w_i (e_i(\theta) / \dot{y}^{\lambda-1})^2] + 1\}.$$

Since the weights w_i are fixed, $\hat{\theta}$ minimizes

$$SS(\theta) = \Sigma_{i=1}^{N} w_i [e_i(\theta)/\dot{y}^{\lambda}]^2. \tag{5}$$

Following Box and Cox (1964), $\hat{\theta}$ can be computed as follows. For fixed λ, minimize $SS(\theta)$ in β by ordinary (typically nonlinear) least-squares and call the minimizer $\hat{\beta}(\lambda)$. Plot $L_{max}(\hat{\beta}(\lambda),\lambda)$ on a grid and maximize graphically or numerically. This technique is particularly attractive when f is *not* transformed and $f(x,\beta) = x^T\beta$ for then (5) can be minimized in β by linear least-squares. When transforming both sides, this technique is less attractive computationally but for the unweighted MLE it does give the confidence interval

$$\{\lambda: L_{max}(\hat{\beta}(\lambda),\lambda) \geq L_{max}(\hat{\beta}(\hat{\lambda}),\hat{\lambda}) - (1/2)\,\chi^2_1(1-\alpha)\},$$

where $\chi^2_1(1-\alpha)$ is the $(1-\alpha)$ quantile of the chi-square distribution with one degree of freedom. Minimizing (5) simultaneously in λ and θ is straightforward with standard nonlinear regression software. One first creates a dummy variable D_i which is 0 for all cases. Define

$$z_i(\theta) = e_i(\theta) \cdot \dot{y}^{\lambda} = [y^{(\lambda)} - f^{(\lambda)}(x_i,\beta)] \cdot \dot{y}^{\lambda}.$$

One fits the "pseudo-model"

$$D_i = z_i(\theta), \tag{6}$$

with "response" D_i, "regression function" $z_i(\theta)$, "regression parameter" θ, and "independent variable" (y_i,x_i). The dummy variable D_i is created

because nonlinear least squares software typically does not allow the response to depend on the parameters. In (6) the real response is incorporated into the model so that is can be transformed by λ.

The standard error of $\hat{\lambda}$ that is output when fitting (6) with a least-squares package should not be used. It is not that same as the standard error from the inverse Fisher information and it does not consistently estimate the large-sample standard deviation of $\hat{\lambda}$; see section 5.

3. Diagnostics

A simple and easily interpreted way of measuring the influence of the ith case is to recompute $\hat{\theta}$ with this case deleted. Since nonlinear estimation can be computer intensive, we will describe two simple approximations to these case-deletion diagnostics.

Let $\hat{\theta}$ and $\hat{\theta}_{(i)}$ be the MLE of θ with and without the ith case, and let $\Delta_i^E = (\Delta^E\beta_i, \Delta^E\lambda_i) = \hat{\theta} - \hat{\theta}_{(i)}$ be the exact change in $\hat{\theta}$ upon deletion of this case. Let $\nabla z_i(\theta)$ and $\nabla^2 z_i(\theta)$ be the gradient and Hessian of $z_i(\theta)$. The first step in the approximation to Δ_i^E will be to ignore the change in $\dot{y}$ when y_i is deleted. From the previous section, $\hat{\theta}_{(i)}$ is the solution to

$$H_{(i)} = \Sigma_{j \neq i} z_j(\theta) \nabla z_j(\theta) = 0. \tag{7}$$

An approximate solution to (7) is obtained by taking one step of the Newton-Raphson algorithm beginning at $\hat{\theta}$. This requires the differential of the left-hand side of (7) which is

$$(8) \qquad \nabla H_{(i)}(\theta) = \Sigma_{j \neq i} \{\nabla z_j(\theta)\ \nabla^T z_j(\theta) + z_j(\theta)\ \nabla^2 z_j(\theta)\}.$$

Then the one-step approximation to $\hat{\theta}_{(i)}$ is

$$\tilde{\theta}_{(i)} = \hat{\theta} - [\nabla H_{(i)}(\hat{\theta})]^{-1} H_{(i)}(\hat{\theta})$$

$$= \hat{\theta} - [\nabla H_{(i)}(\hat{\theta})]^{-1} (z_j(\hat{\theta})\ \nabla z_j(\hat{\theta})), \text{ since}$$

$$\Sigma_{j=1}^{N} z_j(\hat{\theta})\ \nabla z_j(\hat{\theta}) = 0.$$

To simplify $\nabla H_{(i)}(\hat{\theta})$ we first note that by the law of large numbers

$$\dot{y} \cong \exp\{E[\Sigma_{i=1}^{N} \log(y_i)/N \mid x_1, \ldots, x_N]\} = \dot{\mu}, \text{ say.}$$

If $\dot{y}$ is replaced by $\dot{\mu}$ in $z_i(\theta)$, then $\partial / \partial\beta\ z_i(\theta)$ does not depend on $y_1, \ldots, y_N$. Thus, only the lower right-hand element of $\nabla^2 z_j(\theta)$ depends on y_j, so that $z_j(\theta)$ is independent of all other entries of $\nabla^2 z_j(\theta)$. Therefore, letting θ_0 be the true parameter we have

$$E[z_j(\theta_0)\ \nabla^2 z_j(\theta_0)] \cong \text{diag}\{0, \ldots, 0, E[z_j(\theta_0)\ (\partial^2 / \partial\lambda^2\ z_j(\theta_0))]\}$$

since $Ez_j(\theta_0) = 0$.

Therefore, letting

$$B_j = \Sigma_{j \neq i} \{ \nabla z_j(\hat{\theta}) \nabla^T z_j(\hat{\theta}) + D_j(\hat{\theta}) \}, \text{ where} \tag{9}$$

$$D_j = D_j(\theta) = \text{diag}\{0, \ldots, 0, z_j(\hat{\theta}) (\partial^2 / \partial \lambda^2 z_j(\hat{\theta}))\}$$

we have $\nabla H_{(j)}(\hat{\theta}) \cong B_j$.

The approximate influence diagnostic for the ith case is

$$\Delta_i^A = (\Delta^A \beta_i, \Delta^A \lambda_i) = -B_j^{-1} \nabla z_j(\hat{\theta}) z_j(\hat{\theta}) \cong \hat{\theta} - \hat{\theta}_{(i)}.$$

The inverse of B_j can be easily computed using the well-known identity [Rao (1973, page 33)]

$$(A - uv^T)^{-1} = A^{-1} + A^{-1}uv^TA^{-1}/(1 - v^TA^{-1}u), \tag{10}$$

which holds for nonsingular $p \times p$ matrices A and p-dimensional vectors u and v such that $v^TA^{-1}u \neq 1$ (so that $A - uv^T$ is nonsingular).

Let

$$C = \Sigma_{j=1}^{N} \{\nabla z_j(\hat{\theta}) \nabla^T z_j(\hat{\theta}) + z_j(\hat{\theta}) \nabla^2 z_j(\hat{\theta})\},$$

$$C_j = C - \nabla z_j(\hat{\theta}) \nabla^T z_j(\hat{\theta}),$$

$$u_j = (0, \ldots, 0, | z_j(\hat{\theta}) \nabla^2 z_j(\hat{\theta})|^{1/2}), \text{ and}$$

$$v_j = \text{sign}[z_j(\hat{\theta})\, \nabla^2{}_j(\hat{\theta})]\, u_j.$$

Then $D_j = u_j v_j^T$ so that $B_j = C_j - u_j v_j^T$. Using (10) twice we have

$$(11) \qquad C_j^{-1} = C^{-1} + C^{-1}\nabla z_j(\hat{\theta})\nabla^T z_j(\hat{\theta})C^{-1}/(1 - \nabla^T z_j(\hat{\theta})C^{-1}\nabla z_j(\hat{\theta}))$$

and then

$$(12) \qquad B_j^{-1} = C_j^{-1} + C_j^{-1}u_j v_j^T C_j^{-1}/(1-u_j^T C_j^{-1} v_j).$$

Using (11) and (12) allows us to compute $B_1^{-1},\ldots,B_N^{-1}$ using only the single matrix inversion needed to calculate C^{-1}. If we ignore the $D_j(\hat{\theta})$ in (9), then we obtain an even simpler, but less accurate, approximation

$$\Delta_i^Q = (\Delta^Q\beta_i,\ \Delta^Q\lambda_i) = -\,[\Sigma_{j\neq i}\nabla z_j(\hat{\theta})\nabla^T z_j(\hat{\theta})]^{-1} z_j(\hat{\theta})\nabla z_j(\hat{\theta}).$$

$\Delta\lambda_i^Q$ is analogous to the "quick-estimate" diagnostic used by Atkinson (1986) for the response transformation model. The advantage of Δ_i^Q is that it can be computed using standard software packages that calculate linear regression diagnostics. To do this, one creates a linearized model with

$$-\,(z_1(\hat{\theta}),\ldots,z_N(\hat{\theta}))^T$$

as the vector of dependent variables and

$$(\nabla z_1(\hat{\theta}), \ldots, \nabla z_N(\hat{\theta}))^T$$

as the design matrix. Then Δ_i^Q, i = 1,...,N, are the diagnostics DFBETA of Belsley, Kuh, and Welsch (1980, page 13). Some software packages compute a scaled version DFBETAS, which is equally useful for diagnostics. Cook's D or DFFITS from the linearized model can be used as measures of total influence for β and λ.

If we used one step of the Gauss-Newton, rather than the Newton-Raphson, algorithm when solving (7) then we would obtain Δ_j^Q, not Δ_j^A. The difference between the Gauss-Newton and Newton-Raphson algorithms is that the former uses an approximate Hessian which in the present notation consists of ignoring the term $\Sigma_{j=1}^N z_j(\hat{\theta})\nabla^2 z_j(\hat{\theta})$ in the Hessian of the sum of squares. For regression models without a transformation parameter, the residuals are uncorrelated with their Hessians and the Gauss-Newton approximation is acceptable.

In the example of section 6 and in all other examples that we have examined, $|\lambda_i^Q|$ substantially overestimates large values of $|\lambda_i^E|$ while λ_i^A is a good approximation to λ_i^E. The overestimation results from positive correlation between $z_i(\hat{\theta})$ and $\partial^2/\partial\lambda^2\ z_i(\hat{\theta})$ causing

$$\Sigma_{j=1}^N z_j(\hat{\theta})\ [\partial^2/\partial\lambda^2\ z_j(\hat{\theta})]$$

to be positive and of the same magnitude as $\Sigma_{j=1}^N [\partial/\partial\lambda\ z_j(\hat{\theta})]^2$.

If we use λ_j^Q only as a diagnostic, then the overestimation of $|\Delta\lambda_i^E|$ is not a serious problem and it does not prevent us from detecting influential cases.

The diagnostic Δ_i^E or its approximations are vectors showing influence separately for each parameter. An overall measure of influence of the ith case is the exact likelihood distance defined by Cook and Weisberg (1982, section 5.2) as

$$(13) \qquad LD_i^E = 2[L_{max}(\hat{\theta}) - L_{max}(\hat{\theta}_{(i)})] \cong (\Delta_i^E)^T(\nabla^2 L_{max}(\hat{\theta}))\Delta_i^E.$$

The approximation in (13) is also found in Cook and Weisberg (1982) and follows from a Taylor expansion using $\nabla L_{max}(\hat{\theta}) = 0$. We can define an accurate approximation, LD_i^A by replacing Δ_i^E in (13) with Δ_i^A. As a quick approximation one can use

$$LD_i = (\Delta_i^Q)^T[\Sigma_{i=1}^N \nabla z_i(\hat{\theta})\nabla^T z_i(\hat{\theta}))]\Delta_i^Q,$$

which is a constant multiple of Cook's D from the linearized model. From the above discussion we can expect that LD_i^D will not be an accurate approximation to LD_i^E.

4. Robust Estimation

Once influential cases have been identified, we must decide how they should be treated. In some situations, the statistician will feel

that they are valid data, that they do not indicate model deficiencies, and that they should be allowed full influence on the analysis. Then the MLE can be used.

In other situations, the influential cases will be suspected as gross errors. Alternatively, the influential cases may indicate a model deficiency, but either the sparsity of data near their x values will not allow a better model to be developed or the data analyst will hesitate to add complexity to the model merely to accommodate one or at most a few observations. Then a robust estimator should be used.

This section is concerned with the robust estimation of θ. The scale parameter σ can be estimated separately with a robust scale functional, e. g. the median absolute deviation (MAD) applied to the residuals from a robust fit. Let $s_i(y_i,\theta) = \nabla \ell_i(y_i,\theta,\hat{\sigma}(\theta))$ be the score function, i. e. the gradient of the log-likelihood for y_i. The weighted MLE satisfies

$$\Sigma_{i=1}^{N} w_i \, s_i(y_i,\hat{\theta}) = 0.$$

The ordinary (unweighted) MLE is sensitive to cases with large values of s_i, in particular, to cases with large values of residual $e_i(\hat{\beta},\hat{\lambda})$, $\partial/\partial\beta$ $(f^{(\hat{\lambda})}(x_i,\beta))$, or $\partial/\partial\lambda$ $s_i(y_i,\theta)$, corresponding to response, high leverage points, and points having high influence for λ, respectively.

A robust bounded-influence estimator can be found by letting w_i decrease as some norm of $s_i(y_i,\theta)$ increases. To do this the weights

must be allowed to depend on θ. Thus we define the estimator $\tilde{\theta}$ as the solution to

$$(14) \qquad \Sigma_{i=1}^{N} w_i(y_i,\tilde{\theta})\, s_i(y_i,\tilde{\theta},\hat{\sigma}(\tilde{\theta})) = 0,$$

where w_i is a suitable scalar weighting function and, as in section 2, $\hat{\sigma}^2(\tilde{\theta})$ is the weighted variance of the residuals.

Let $\phi_i(y,\theta) = w_i(y,\theta)s_i(y,\theta)$. The minimum requirement for robustness is to have $\phi_i(y,\theta)$ bounded as a function of i, y , and θ. Otherwise, a single outlier can cause an arbitrarily large change in $\tilde{\theta}$.

It should be mentioned that a bounded-influence estimators may not have a high breakdown point. The breakdown point is the largest percentage of contamination that an estimator can tolerate before it can be overwhelmed by the contaminants. This means that if the percentage of contaminants exceeds the breakdown point, then the estimate can be forced to take an arbitrary value by choosing the contaminants in a sufficiently nasty way. The estimators that we define here are related to the Krasker-Welsch regression estimator, and like that estimator they will have a breakdown point at most $(p+1)^{-1}$, where (p+1) is the total number of parameters.

In the case of the linear model, Rousseeuw (1984) and Rousseeuw and Yohai (1984) have proposed estimators with near 50% breakdown points, the best possible. However, these estimators have poor efficiency. Yohai (1985) shows how to achieve both high asymptotic efficiency and a near 50% breakdown point, but again only in the case of linear

regression. In the future, we hope to develop similar estimators for transformation models. A major difficulty will be computational complexity.

When choosing w, asymptotic efficiency measured by the covariance matrix of $\tilde{\theta}$ must be balanced against robustness measured by the supremum of some norm of $\phi_i(y,\theta)$, the so-called gross-error sensitivity. For univariate parameters there is a unique optimal w_i which minimizes the asymptotic variance subject to a given bound on the gross-error sensitivity (Hampel 1968, 1974). For multivariate parameters such as θ the balancing of robustness and variance raises philosophical questions, since there are many ways of comparing covariance matrices or of norming vector functions. Different norms on $\phi_i(y_i,\theta)$ give rise to different definitions of gross-error sensitivity.

The approach we take generalizes the Krasker-Welsch (1982) bounded-influence regression estimates. Whether the Krasker-Welsch estimator optimizes the asymptotic covariance matrix in any meaningful sense is an open question, but its efficiency at the normal model is usually close to that of the MLE (Ruppert 1985). Krasker and Welsch (1982) bound the so-called self-standardized gross-error sensitivity, which we denote by γ_2. We will describe γ_2 only briefly and the interested reader is referred to the original paper of Krasker and Welsch or to Hampel et al. (1986, chapter 4) for further details.

First we note that the influence function of $\tilde{\theta}$ satisfying (14) is

$$IF_i(y_i,\theta) = B^{-1}\phi_i(y_i,\theta), \text{ where}$$

$$B = - N^{-1}\Sigma_{i=1}^{N} \nabla^T E[\phi_i(y_i,\theta)].$$

This definition of the influence function is conditional on $x_1,\ldots,x_N$ but coincides with the usual definiton when the x's are independent and identically distributed and summation over i is replaced by expectation with respect to the x's. The definition of B is analogous to that on page 5 of Carroll and Ruppert (1985) where w is incorrectly squared.

The asymptotic covariance matrix of $\tilde{\theta}$ is $B^{-1}A(B^{-1})^T$ where

$$A = N^{-1}\Sigma_{i=1}^{N} E[\phi_i(y_i,\theta)\phi_i^T(y_i,\theta)].$$

An intuitively reasonable way to norm influence function is to use the asymptotic covariance matrix of the estimator. See Krasker and Welsch (1982) for further motivation and discussion. The resultant measure of influence is the so-called self-standardized gross-error sensitivity defined as

$$\gamma_s = \max_i \| IF(y_i,\tilde{\theta}) \|_V = \max_i \| w_i(y_i,\tilde{\theta})\, s_i(y_i,\tilde{\theta}) \|_A,$$

where $\| v \|_M = (v^T M^{-1} v)^{1/2}$ for any vector v and positive definite matrix M. This definition of γ_s is analogous to equation (15) of Carroll and Ruppert (1985) where w^2 has been incorrectly omitted from the last term.

To robustify the MLE, we will choose the weights so that γ_s does not exceed a predetermined bound. γ_s must be at least $(p+1)^{1/2}$ (Krasker

and Welsch 1982). From experience with this and other problems we suggest bounding γ_s by $a(p+1)^{1/2}$ where "a" is between 1.1 and 1.6. The choice a = 1.5 worked well on the data set in section 6.

If an observation has low influence then it should not be downweighted. Otherwise, it should be downweighted just enough to keep γ_s below the given bound. Therefore the weights should be

$$(15) \qquad w_i(y_i,\theta) = \min\{1,\ a(p+1)^{1/2}/\|s_i(y_i,\theta)\|_A\}.$$

In (15) A must be replaced by an estimate $\hat{A}$.

To calculate $\tilde{\theta}$ we used a simple iterative scheme:

(1) Fix a > 1. Let C be the total number of iterations that will be used. Set c=1. Let $\hat{\theta}_p$ be a preliminary estimate, possibly the MLE. Set $w_i = 1$ for all i.

(2) Define

$$\hat{A} = N^{-1}\Sigma_{i=1}^{N} w_i^2\, s_i(y_i,\hat{\theta}_p)s_i^T(y_i,\hat{\theta}_p).$$

(3) Using (15) update the weights:

$$w_i = \min\{1,\ a(p+1)^{1/2}/\|s_i(y_i,\hat{\theta}_p\|_{\hat{A}}\}.$$

(4) Using the methods of section 2, find the weighted MLE with these weights, and call it $\tilde{\theta}$.

(5) If $c < C$, set $\hat{\theta}_p = \tilde{\theta}$, $c = c + 1$, and return to step (2). Otherwise, stop.

It is possible to implement this algorithm on standard software packages, in particular SAS. Steps (2) and (3) can be computed with a

matrix language such as PROC MATRIX on the 1982 version of SAS. Step (4) can be performed using a weighted nonlinear least-squares routine such as PROC NLIN on SAS. By using macros on SAS or a similar package, it is possible to put the matrix computations and the least-squares routines into an iterative loop. We initially used SAS, but now prefer the matrix programming language GAUSS.

One or two iterations seem adequate for diagnostic purposes, but this algorithm sometimes converges slowly, particularly when there are extremely influential points. This was the case with the example in section 6 where the algorithm did not stabilize until ten iterations. Unfortunately, the slow convergence gave the appearance that the algorithm had converged after only two or three iterations.

We found that a fully iterative version of the algorithm could be easily implemented with the GAUSS on an IBM PC-AT, and computation time for ten or more iterations was acceptable.

Although bounded-influence estimation limits the effect of any case on the estimate, all cases regardless of how deviant from the bulk of the data will have some influence. Hampel el al. (1986) discuss the need for robust estimators that completely reject extreme outliers. For estimation of a location parameter, this can be done with a redescending "psi function".

We can define an analog to a redescending psi-function here. Let $\Psi(x)$ be an odd function with $\Psi(x) \geq 0$ for $x \geq 0$, and define the weights

$$(16) \qquad w(y_i,\theta) = \Psi(\|s_i(y_i,\theta)\|_A^{\wedge}) / \|s_i(y_i,\theta)\|_A^{\wedge}.$$

Equation (16) reduces to (15) if Ψ is the Huber psi-function

$$\Psi(x) = x \qquad |x| \le b$$
$$= b\ \mathrm{sign}(x) \qquad \text{otherwise.}$$

with $b = a(p+1)^{1/2}$.

If for some $R > 0$, $\Psi(x) = 0$ for all $|x| > R$, then extreme outliers are completely rejected. In the example we use Hampel's three-part redescending psi-function

$$\Psi(x) = x \qquad 0 \le x \le b_1$$
$$= b_1 \qquad b_1 \le x \le b_2$$
$$= b_1(b_3 - x)/(b_3 - b_2) \qquad b_2 \le x \le b_3 \qquad (17)$$
$$= 0 \qquad b_3 \le x,$$

with $(b_1, b_2, b_3) = (p+1)^{1/2}(1.5, 3.5, 8.0)$.

In section 6, the redescending estimate was computed by the same algorithm used to calculate the bounded-influence estimate. In step 1, w_i and $\hat{\theta}_p$ were from the last iteration of the bounded-influence estimate. In step (3) the weights were calculated using (16) and (17) instead of (15).

5. Estimating The Covariance Matrix of $\hat{\theta}$

The asymptotic covariance matrix of the MLE can be found by (i) inverting the Fisher information matrix or (ii) using the covariance matrix of the influence function, the covariance being with respect to the empirical distribution of (y_i, x_i), $i=1,\ldots,N$. Method (ii) is based on the asymptotic theory of M-estimation; see for example Hampel et al. (1986, section 4.2c). One advantage of (ii) is that the asymptotic covariance is consistently estimated even if the ϵ_i are not normally distributed (Huber 1967) or do not have a constant variance. In fact, method (ii) is similar to the jackknife which Wu (1986) advocates as a consistent estimate of the least-squares covariance matrix under heteroscedasticity. Moreover, method (ii) can be used for the bounded-influence and redescending estimates as well.

Method (i): The observed Fisher information matrix for $(\hat{\theta},\hat{\sigma})$ is

$$-\nabla^2 L(\hat{\theta},\hat{\sigma}), \tag{18}$$

where the weights in (3) are all unity and ∇^2 means the Hessian with respect to (θ,σ). We could invert (18) and then take the upper left $(p+1)^2$ corner corresponding to $\hat{\theta}$, but by Patefield (1977, 1985) this is equivalent to the easier computation of inverting the $(p+1)^2$ matrix

$$-\nabla^2 L_{max}(\hat{\theta}),$$

where ∇^2 now is the Hessian with respect to θ only

Method (ii): The MLE, bounded-influence estimator, and the redescending estimator are all defined by the equation

$$(19) \qquad \Sigma_{i=1}^{N} w_i(y_i,\hat{\theta})\, s_i(y_i,\hat{\theta}) = \Sigma_{i=1}^{N} \phi_i(y_i,\hat{\theta}) = 0$$

and only differ in the choice of the weights. The asymptotic covariance matrix of any estimator solving an equation of the form (19) can be estimated by $[\hat{B}^{-1}\ \hat{A}\ (\hat{B}^{-1})^T]$ where

$$\hat{B} = \nabla^T \Sigma_{i=1}^{N} \phi_i(y_i,\hat{\theta}) \text{ and}$$

$$\hat{A} = \Sigma_{i=1}^{N} \phi_i(y_i,\hat{\theta})\phi_i^T(y_i,\hat{\theta}).$$

The asymptotic theory of M-estimation also shows that the standard error $\hat{A}$ when fitting the "pseudo-model" (see section 2) are incorrect. The pseudo-model finds the MLE by minimizing

$$\Sigma_{i=1}^{N} z_i^2(y_i,\hat{\theta})$$

or solving

$$(20) \qquad \Sigma_{i=1}^{N} z_i(y_i,\hat{\theta})\, \nabla z_i(y_i,\hat{\theta}) = 0.$$

When the pseudo model is fit by a nonlinear least squares package the estimated covariance matrix is

$$(21) \qquad \hat{\sigma}^2(\hat{\theta})\ \{\Sigma_{i=1}^{N} \nabla z_i(y_i,\hat{\theta})\ \nabla^T_i(y_i,\hat{\theta})\}^{-1}.$$

The asymptotic covariance of the solution to (20) is consistently estimated by

$$(22) \qquad \hat{B}_{LS}^{-1}\hat{A}_{LS}(\hat{B}_{LS}^{-1})^T$$

where $\hat{B}_{LS} = \nabla^T[\Sigma_{i=1}^{N} z_i(y_i,\hat{\theta})\ \nabla z_i(y_i,\hat{\theta})]$ and $\hat{A}_{LS} = \Sigma_{i=1}^{N} \nabla z_i(y_i,\hat{\theta})\ \nabla^T z_i(y_i,\hat{\theta})$. It is not hard to see that (21) and (22) have different limits so that (21) is inconsistent. In practice (21) and (22) can be considerably different, especially in the estimated variance of $\hat{\lambda}$.

6. An Example From Fisheries Analysis

In this section we look at an example. Our goals are (a) to see the type of data analytic information that can come from the diagnostics and the robust estimators, (b) to see how well the robust estimators handle outlying data, and (c) to compare the accuracy of the "accurate" approximation Δ_i^A to the "quick" approximation Δ_i^Q.

When managing a fish stock, one must model the relationship between the annual spawning stock size and the eventual production of new catchable sized fish (returns or recruits) from the spawning. Ricker and Smith (1975) give numbers of spawners (S) and returns (R) from 1940

until 1967 for the Skeena River sockeye salmon stock.

Using some simple assumptions about factors influencing the survival of juvenile fish, Ricker (1954) derived the theoretical model

$$R = \beta_1 S \exp(\beta_2 S) = f(S,\beta)$$

relating R and S. Other models have been proposed, e. g. by Beverton and Holt (1957). However, the Ricker model appears to fit adequately and, in particular, gives almost the same fit to this stock as the Beverton-Holt model.

From Figure 1, a plot of R against S, it is clear that recruitment is highly variable and heteroscedastic, with the variance of R increasing with its mean. Several cases appear somewhat outlying, in particular #5, #18, #19, and #25.

An index plot of D_i^A was constructed; see Figure 2. Clearly case #12 stands out as the most influential by this measure, and #5, #19, and #25 are only moderately influential by comparison. We will examine these cases more closely.

The exact case-deletion statistic Δ_i^E and the approximations Δ_i^A and Δ_i^Q are given in table 1 for these four cases.

Case #12 has a high influence on all three parameters. In particular, $\Delta\lambda_i^E = .51$, showing that the MLE of λ decreases from .31 to -.2 when # 12 is deleted. Why is #12 so influential?

To see why, look again at Figure 1.. Observation #12 is not far removed from the median relative to the variation in all the data, but

it is quite far removed relative to the variation in the other data with similar values of the independent variable S.

The 27 data points besides #12 suggest that the data are extremely heteroscedastic, with the variation in recruitment increasing very rapidly with the median recruitment. The effect of #12 is to increase the apparent recruit variation for low median recruitment and to suggest that the heteroscedasticity is not so nearly pronounced. Seen in this light, the much more severe transformation ($\hat{\lambda} = -.2$ instead of $\hat{\lambda} = .31$) when #12 is deleted is not surprising.

Besides suggesting less heteroscedasticity than seen in the remainder of the data, #12 has a large negative residual which suggests less right skewness as well.

To further analyze the influence of #12 on $\hat{\lambda}$ we will introduce two alternative estimators of λ. These will be discussed fully in a forthcoming paper by Aldershof and Ruppert. For fixed λ, let $\hat{\beta}(\lambda)$ be the MLE of β and define $\hat{\theta}(\lambda) = (\hat{\beta}(\lambda)^T, \lambda)^T$. The skewness estimator $\hat{\lambda}_{sk}$ is the value of λ such that the skewness coefficient of the residuals $\{e_i(\hat{\theta}(\lambda))\}$ is 0. The heteroscedasticity estimator $\hat{\lambda}_{het}$ is the value of λ such that the correlation between $\{e_i(\hat{\theta}(\lambda))\}$ and $\{\log(f(x_i, \hat{\beta}(\lambda))\}$ is 0.

When case #12 is omitted the value of $\hat{\lambda}_{sk}$ only changes from .58 to .46, but $\hat{\lambda}_{het}$ changes much further, from .16 to -.86. The major effect of deleting #12 is to increase the heteroscedasticity.

Case #12 was the year 1951 when a rock slide drastically reduced recruitment (Ricker and Smith 1975). For this reason we are quite

comfortable downweighting it severely. In fact, it seems best to reject #12 entirely. This, in effect, is what the redescending estimator does; see below.

Compared to case #12, cases #5, #19, and #25 all have the opposite effect on $\hat{\lambda}$. Deleting any of them decreases the apparent recruitment variance when the number of expected recruits is large, in effect suggesting less heteroscedasticity. For this reason $\Delta\lambda_i^E$ is positive for $i = 5, 19,$ and 25.

The effect of #12 on $\hat{\beta}$ is not as easily analyzed as its effect on $\hat{\lambda}$. Deleting #12 increases $\hat{\beta}_1$ from 3.29 to 3.77 and decreases $\hat{\beta}_2$ from -7.00 to -9.54. These effects tend to cancel but not completely. As shown in Figure 3, the net effect of including #12 is a decrease in $f(S,\hat{\beta})$ for small S and an increase for large S. The former effect is plausible since #12 has a low value of S and a negative residual. The increase in $f(S,\hat{\beta})$ for large S is not so plausible, but it is a consequence of using the Ricker model for the median recruitment.

Having analyzed the exact changes Δ_i^E, we turn to the accuracy of the approximations Δ_i^A and Δ_i^Q. The accuracy of Δ_i^Q is poor for cases #5 and #12, especially #12, and these are precisely the cases of interest. The same is true of the approximations to LD_i^E; the quick approximation is rather inaccurate.

The quick estimate does better for cases #19 and #25, but this is merely fortuitious. The overestimation by the quick estimate is small here and cancels the error from not recalculating $\dot{y}$ after case-deletion. To see this we can examine the changes in the MLE induced by

case deletion with $\dot{y}$ is kept equal to the geometric mean of all the y's. These changes are -.17, .77, -.013, and .011 for i = 5, 12, 19, and 25, respectively. In all four cases, the change is closer to $\Delta^{A}\lambda_i$ than to $\Delta^{Q}\lambda_i$.

We calculated 20 iterations of the bounded-influence estimate with the tuning constant a = 1.5, and using this as a starting value we calculated 10 iterations of the redescending estimate with $(b_1, b_2, b_3) = (p+1)^{1/2}(1.5, 3.5, 8.0)$.

The bounded-influence estimate changed rapidly for the first five iterations, more slowly for the next five, and then was stable for the last ten iterations. To show the behavior of the algorithm, the value of $\tilde{\theta}$ and non-unity values of w_i are given in table 2 for iterations 1, 2, 3, 4, 5, 10, and 20. It is interesting to examine w_4. This is 1 for the first three iterations but eventually decreases to .65, less than w_{19} and w_{25}.

Case #4 is influential, but its influence is masked by #12. Robust estimation or subset deletion diagnostics seem necessary to detect the influence of #4.

The values of $\tilde{\theta}$ and nonunity values of w_i are also given in table 2 for the redescending estimate. Case #12 is completely rejected, e. g., $w_{12} = 0$, for all iterations. The value of w_4 decreases slowly for three iterations and then jumps downward to almost 0 on the fourth iteration. With #4 nearly deleted the weights w_{16}, w_{19}, and w_{25} readjust on the fifth and sixth iterations. The redescending estimate is stable after the sixth iteration.

The first iteration of the redescending estimator is nearly identical to the MLE without #12. After #4 is strongly downweighted, $\hat{\lambda}$ decreases to about -.3 which suggests even slightly stronger heteroscedasticity. As both #12 and #4 are completely or nearly completely rejected, #16 becomes influential and is slightly downweighted.

We do not necessarily advocate using the fully-iterated redescending estimator. The bounded-influence estimator or the first iterate of the redescender give a good fit to the bulk of the data and reject the outlier #12. Without further information about these data, we are not comfortable downweighting #4 and #16 as much as the fully iterated redescending estimator downweights them.

If forced to choose one estimate, we would choose the one-step redescender. The residuals $\{e_i(\tilde{\theta})\}$ from this estimate are plotted in Figure 4. Ignoring $e_{12}(\tilde{\theta})$, the remaining residuals show only slight heteroscedasticity and almost no skewness.

The redescending estimator completely rejects #12 so there is no need to refit with this anomalous case removed. However, it is instructive to see what happens if this is done. Both the bounded-influence and the redescending estimators converge rapidly, with the first iterate equal for practical purposes to the fully-iterated estimate. This suggests that the algorithm converges slowly only in the presence of an extremely influential point such as #12.

In table 3 we give the standard errors of the MLE by method (i), inverting the observed Fisher information matrix. We also give the

method (ii) standard errors for the MLE, the one-step and iterated (c = 10) bounded-influence estimator, and the one-step and iterated (c = 5) redescending estimator.

The method (i) and method (ii) standard errors of the MLE of λ are moderately different, but both are considerably smaller than the standard error of $\hat{\lambda}$ for the robust estimators. Downweighting #12 seems to increase the variability of $\hat{\lambda}$, but since #12 is known to be an unusual year it seems wise to use a robust estimator despite the higher variability. The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ are smaller for the robust estimators than for the MLE.

At least in this example, robust estimation appears to cause a loss in efficiency of $\hat{\lambda}$ but a gain in efficiency of $\hat{\beta}$. However, this loss in efficiency for λ is only apparent, not real. There are two ways of looking at this, either conditioning or not conditioning on the event that a rock slide occurred in 1951. Conditional on there being exactly one rock slide and it occuring in a year when the number of spawners was low, the MLE has a small variance but a large bias and consequently a large mean square error relative to the robust estimators. If we do not condition on the occurence of exactly one slide, but rather admit that some other number of slides could have occurred and that these could have occurred at any years, then it is clear that the MLE is really much more variable than its standard error shows.

The change in $\hat{\lambda}$ when #12 is deleted is large relative to the standard errors of $\hat{\lambda}$. Notice that $\Delta^{E}\lambda_{12}$ is over twice the standard error of the MLE and about 1.65 times the standard error of the one-step redescending estimator.

Table 1. Deletion diagnostics for the most influential cases. The superscripts denote the method of calculation: E = exact, A = accurate approximation, Q = quick approximation.

DIAGNOSTIC	CASE NUMBER			
	5	12	19	25
$\Delta^{E}\lambda_{i}$	-.10	.51	-.034	.033
$\Delta^{A}\lambda_{i}$	-.14	.61	-.015	-.015
$\Delta^{Q}\lambda_{i}$	-.32	1.57	-.036	-.034
$\Delta^{E}\beta_{1,i}$	-.22	-.48	.08	.12
$\Delta^{A}\beta_{1,i}$	-.22	-.45	.11	.15
$\Delta^{Q}\beta_{1,i}$	-.34	-.43	.10	.14
$\Delta^{E}\beta_{2,i}$	1.48	2.54	-1.02	-1.25
$\Delta^{A}\beta_{2,i}$	1.47	2.57	-1.14	-1.38
$\Delta^{Q}\beta_{2,i}$	1.80	3.72	-1.12	-1.36
LD_{i}^{E}	.58	9.64	.32	.38
LD_{i}^{A}	.72	8.7	.27	.34
LD_{i}^{Q}	1.29	24.3	.27	.34

Table 2. Estimates of β and λ and the case weights for the robust estimators. BIE = bounded-influence estimator. RE = redescending estimator.

	ESTIMATES			CASE WEIGHTS					
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\lambda}$	w_4	w_5	w_{12}	w_{16}	w_{19}	w_{25}
MLE	3.29	-7.00	.31	1	1	1	1	1	1
BIE: c=1	3.49	-8.00	.27	1	.71	.58	1	1	1
c=2	3.59	-8.54	.20	1	.63	.34	1	1	1
c=3	3.66	-8.87	.12	1	.62	.20	1	1	.99
c=4	3.70	-9.11	.06	.98	.62	.13	1	1	.98
c=5	3.74	-9.31	.02	.91	.62	.087	1	1	.96
c=10	3.84	-9.71	-.07	.69	.63	.041	.97	1	.91
c=20	3.85	-9.74	-.08	.65	.64	.038	.92	1	.91
RE: c=1	3.91	-10.1	-.21	.65	.64	0	.92	1	.91
c=2	3.93	-10.1	-.23	.52	.71	0	.86	.97	.86
c=3	3.92	-10.0	-.23	.46	.75	0	.81	.94	.82
c=4	3.91	-9.97	-.23	.43	.76	0	.78	.92	.81
c=5	4.11	-10.7	-.36	.006	.77	0	.76	.90	.79
c=10	4.00	-9.94	-.34	.005	.91	0	.64	.71	.61

Table 3: Standard errors of $\hat{\beta}$ and $\hat{\lambda}$.

	ESTIMATOR		
	$\hat{\beta}_1$	β_2	$\hat{\lambda}$
MLE (inverse Fisher information - method (i))	.75	3.40	.21
(as an M-estimator - method (ii))	.66	3.46	.16
BIE: One-step (c=1)	.54	3.01	.42
c=10	.61	3.51	.39
RE: One-step (c=1)	.51	2.79	.35
Fully iterated (c=10)	.43	2.37	.35

LIST OF FIGURES

Figure 1

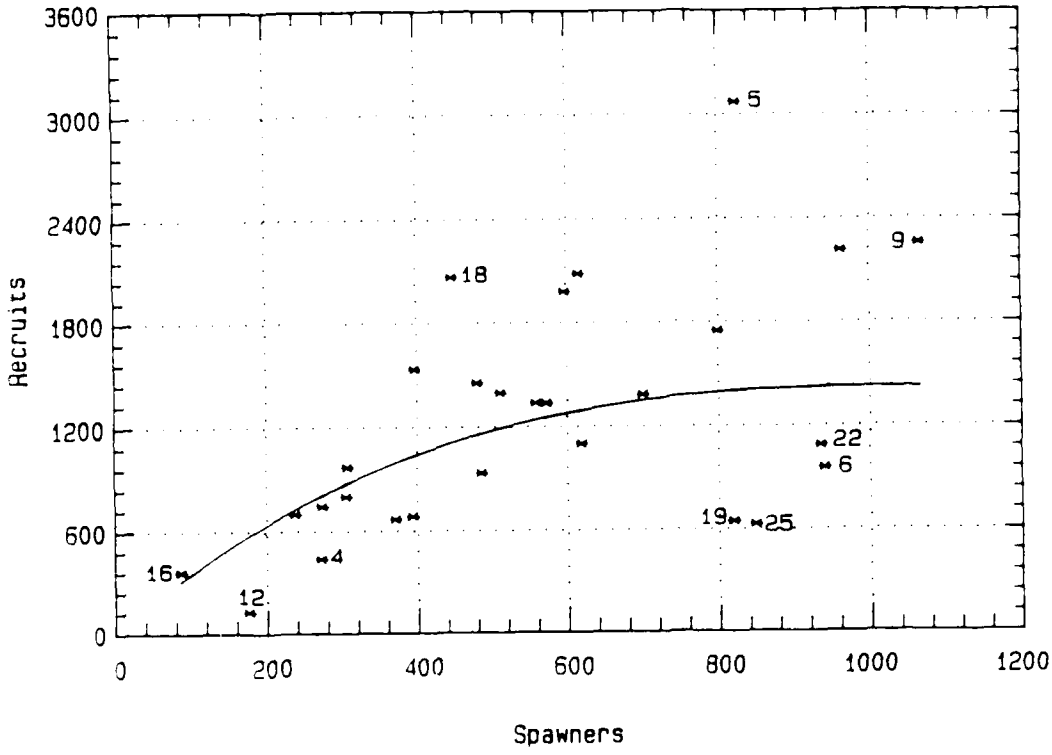


Figure 2

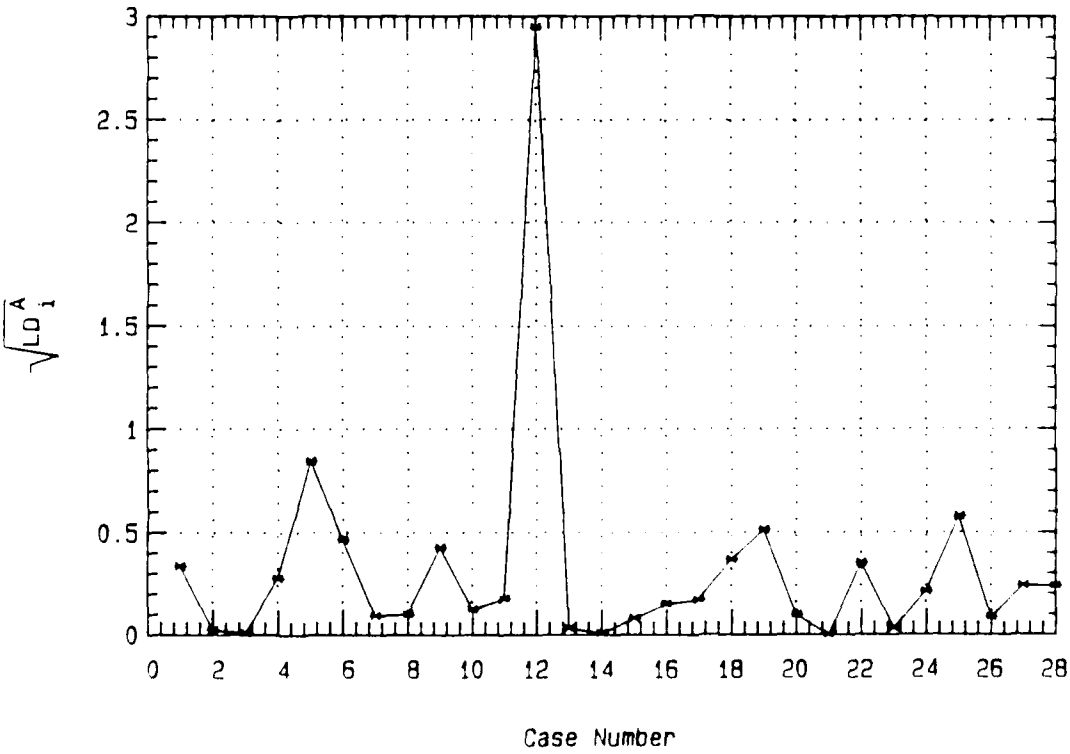


Figure 3

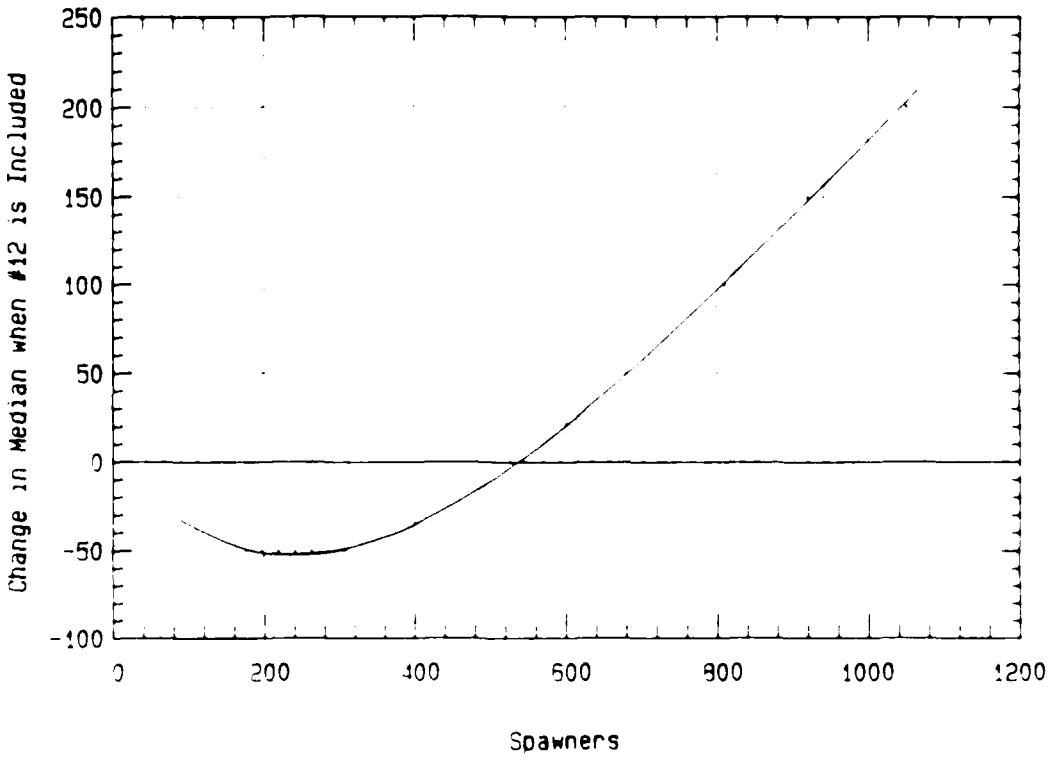
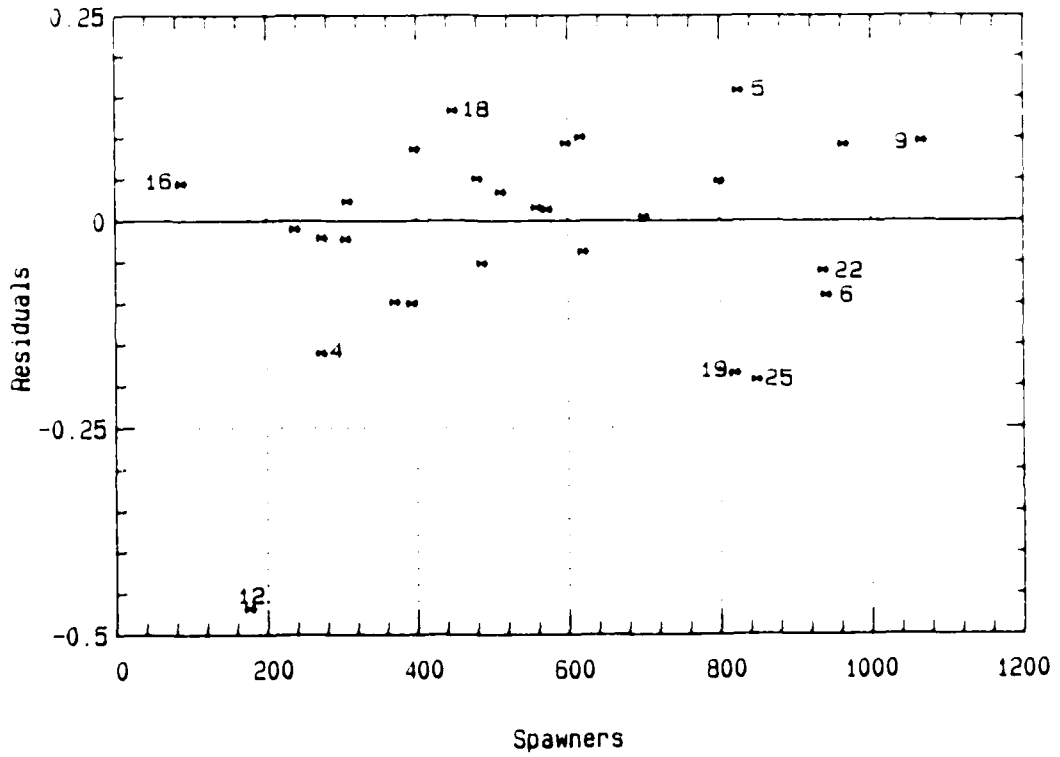


Figure 4



REFERENCES

Atkinson, A. C. (1986). "Diagnostic tests for transformations." **Technometrics**, 28, 29-38

Bates, D. M. and Watts, D. G. (1980). "Relative curvature measures of nonlinearity." **Journal of the Royal Statistical Society, Series B**, 42, 1 25.

Bates, D. M., Wolf, D. A., and Watts, D. G. (1986). "Nonlinear least squares and first-order kinetics." in **Proceedings of Computer Science and Statistics: Seventeenth Symposium on the Interface**. David Allen, ed. New York: North-Holland.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). **Regression Diagnostics**. New York: Wiley.

Box, G. E. P. and Cox, D. R. (1964). "An analysis of transformations (with discussion)." **Journal of the Royal Statistical Society, Series B**, 26, 211-246.

Box, G. E. P. and Hill, W. J. (1974). "Correcting inhomogeneity of variance with power transformation weighting." **Technometrics**, 16, 385-389.

Carr, N. L. (1960). "Kinetics of catalytic isomerization of n-pentane." **Industrial and Engineering Chemistry**, 52, 391-396.

Carroll, R. J. and Ruppert, D. (1981). "On prediction and the power transformation family." **Biometrika**, 68, 609-616.

Carroll, R. J. and Ruppert, D. (1984). "Power transformations when fitting theoretical models to data." **Journal of the American Statistical Association**, 79, 321-328

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

Cook, R. D. (1986). "Assessment of local influence." **Journal of the Royal Statistical Society - Series B**. To appear.

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.

Draper, N. and Smith, H. (1981). **Applied Regression Analysis, 2nd Edition**. Wiley: New York.

Hampel, F. R. (1974). "The influence curve and its role in robust estimation." **Journal of the American Statistical Association**, 62, 1179-1186.

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

Hampel, F. R. (1985). "The breakdown points of the mean combined with some rejection rules." **Technometrics**, 27, 95 107.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). **Robust Statistics.** John Wiley and Sons: New York.

Krasker, W. S. and Welsch, R. E. (1982). "Efficient bounded-influence regression estimation." **Journal of the American Statistical Association**, 77, 595-604.

Patefield, W. M. (1977). "On the maximized likelihood function." **Sankhya**, B, 39, 92-96.

Patefield, W. M. (1985). "Information from the maximized likelihood function." **Biometrika**, 72, 664-668.

Rao, C. R. (1973). **Linear statistical inference and its applications, 2nd. edition.** John Wiley and Sons: New York.

Ricker, W. E. (1954). "Stock and recruitment." **Journal of Fisheries Research Board of Canada**, 11, 559,623.

Ricker, W. E. and Smith, H. D. (1975). "A revised interpretation of the history of the Skeena River sockeye salmon (Oncorhynchus nerka)." **Journal of the Fisheries Research Board of Canada**, 32, 1369-1381.

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

Rousseeuw, P. J. and Yohai, V. (1984). "Robust regression by means of S-estimators." in **Robust and Nonlinear Time Series Analysis**, Franke, J., Hardle, W., and Martin, R. D., ed. Springer-Verlag: Berlin.

Ruppert, D. (1985). "On the bounded influence regression estimator of Krasker and Welsch." **Journal of the American Statistical Association**, 80, 205 208.

Ruppert, D. and Carroll, R. J. (1985). "Data transformations in regression analysis with applications to stock-recruitment relationships." in **Resource Management: Proceedings of the Second Ralf Yorque Workshp held in Ashland, Oregon, July 23-25, 1984**. S. Levin, ed. Springer Verlag: Berlin.

Snee, R. D. (1986). "An alternative approach to fitting models when re-expression of the response is useful." **Journal of Quality Technology**. To appear.

Wu, C. F. J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis. **Annals of Statistics**, 14. To appear.

Yohai, V. (1985). **High breakdown point and high efficiency robust estimates for regression.** Technical Report No. 66, Department of Statistics, University of Washington, Seattle, Washington.

END

4-88-87

DTIC