SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

TECHNICAL REPORT NO. 762

November 1987
Revised September 1988

Transformations to Symmetry and Homoscedasticity

David Ruppert and Brian Aldershof^{*}

^{*} David Ruppert is Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853. Brian Aldershof is Graduate Student, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. This research was supported by NSF Grants DMS-8400602, DMS-8701201, and DMS-8800294. The authors thank Ray Carroll and two referees for helpful comments.

ABSTRACT

This paper considers transformations in regression to eliminate skewness and heteroscedasticity of the response. We work with the "transform—both—sides" model where the relationship between the median response and the independent variables has been identified, at least tentatively. To preserve this relationship, the response and the regression model are transformed in the same way. Extending the work of others for the location parameter case, we propose an estimator $\hat{\lambda}_s$ that eliminates skewness. We also develop an estimator $\hat{\lambda}_h$ to eliminate heteroscedasticity and an estimator $\hat{\lambda}_{hs}$ that attempts to induce both symmetry and homoscedasticity. Both $\hat{\lambda}_h$ and $\hat{\lambda}_{hs}$ appear new. By comparing $\hat{\lambda}_s$ and $\hat{\lambda}_h$ we develop a test of the null hypothesis that there exists a transformation to both symmetry and homoscedasticity.

We study the question, when does the estimator of λ behaves (in terms of asymptotic variance) as if the regression parameter β were known (and vice versa)? The results are of use for telling when the optimal estimator of λ does not depend upon the regression model. In addition, we present two examples and discuss computation of the estimators.

The transformation to symmetry is defined by setting the usual third—moment skewness coefficient of the residuals equal to 0, and the variance—stabilizing transformation is defined by setting the correlation between the squared residuals and the logarithms of fitted values equal to 0. Under normality, both these estimators achieve minimum asymptotic variance within certain natural classes of estimators. The estimator of a transformation to both symmetry and constant variance uses a weighted average of the equations separately defining transformations to symmetry and homoscedasticity, and the optimal weights can be estimated. Robust modifications of the estimators are considered.

One example studies efficiency in a class of robust estimators of transformations to symmetry. A second example, using real data, compares standard errors from asymptotic theory with bootstrap standard errors.

The transformation parameter is nearly orthogonal to the regression parameters. For this reason we concentrate on estimation of the transformation parameter. A good estimate of the transformation parameter is especially important when constructing prediction intervals for future responses. For this reason we concentrate on estimation of the transformation parameter. A good estimate of the transformation parameter is especially important when constructing prediction intervals for future responses.

Key Words Transform—both—sides; Skewness; Test for Heteroscedasticity; M—estimation; Power transformations; Robustness

1. INTRODUCTION

Data transformations have long been used to induce symmetrically distributed and homoscedastic data. In regression problems, the proper transformation could be calculated if one had sufficient knowledge of the conditional distribution of the response y , given covariates, $\mathbf{x}$. Without such knowledge the data themselves are used to choose the transformation. This paper is concerned with estimation of a transformation of y to symmetry and/or homoscedasticity. It is assumed that a theoretical model relating y to $\mathbf{x}$ is available, and the transformation must not destroy this relationship.

In a fundamental paper, Box and Cox (1964) proposed choosing a transformation from within a parametric family by the method of maximum likelihood. Let y be a positive response and let $h(y, \lambda)$ be a transformation of y depending upon the parameter λ . For example, h could be the modified power transformation family given by

$$(1.1) \quad \begin{aligned} h(y, \lambda) &= y^{(\lambda)} = (y^\lambda - 1)/\lambda \quad \text{if } \lambda \neq 0, \\ &= \log(y) \quad \text{if } \lambda = 0. \end{aligned}$$

The Box and Cox assume that for some λ

$$(1.2) \quad h(y, \lambda) = \mathbf{x}^T \boldsymbol{\beta} + \sigma \epsilon,$$

where ϵ has a standard normal distribution and $\mathbf{x}^T \boldsymbol{\beta}$ is a simple linear model. They estimate λ , $\boldsymbol{\beta}$, and σ by maximum likelihood. Model (1.2) assumes that the transformation of y to $h(y, \lambda)$ can achieve three simultaneous objectives: (a) normally distributed errors, (b) a constant variance, and (c) a simple model (given by $\mathbf{x}^T \boldsymbol{\beta}$) for the relationship between the covariates $\mathbf{x}$ and the response y .

In this paper, we will be concerned with a problem somewhat different than that addressed by Box and Cox (1964). Suppose that from a theoretical (e. g. biological or physical) model or empirical knowledge we have already postulated a relationship between y and $\tilde{x}$, say that the median of y given x is

$$f(\tilde{x}, \beta),$$

where f is a possibly nonlinear regression model. If y is skewed and/or heteroscedastic, then we may wish to transform y , but transforming y alone destroys the postulated model, $y = f(\tilde{x}, \beta) + \text{error}$. In this situation, Carroll and Ruppert (1984) propose "transforming both-sides" of the regression equation, that is, using the model

$$(1.3) \quad y_i^{(\lambda)} = f^{(\lambda)}(\tilde{x}_i, \beta) + \epsilon_i,$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. F. Model (1.3) has been discussed in detail by Carroll and Ruppert (1984, 1987, and 1988), and Snee (1986).

If ϵ is distributed symmetrically about 0 or merely has median 0, then (1.3) implies that the median of y given $\tilde{x}$ is $f(\tilde{x}, \beta)$, regardless of the value of λ . Thus, the value of λ is not determined by the conditional median of y given $\tilde{x}$. Rather, it is the asymmetry and the mean–variance relationship of y that determines λ . If (1.3) holds with $\lambda = \lambda_0$ and ϵ symmetrically distributed, then $y^{(\lambda)}$ is right-skewed when $\lambda > \lambda_0$ and left-skewed when $\lambda < \lambda_0$. Also, by a Taylor approximation,

$$(1.4) \quad \text{var}(y^{(\lambda)}) \cong \sigma_\epsilon^2 \{f(\tilde{x}, \beta)\}^{2(\lambda - \lambda_0)}.$$

It is often assumed that the distribution F of ϵ in the transform–both–sides model is normal. However, a difficulty with model (1.1), (1.3) (or with model (1.1), (1.2))

is that since y is positive, ϵ cannot have a distribution whose support is $(-\infty, \infty)$ except in special cases, e.g., when λ is zero.

It appears preferable to assume that ϵ is not normally distributed but rather has a symmetric distribution with bounded support (and perhaps close to normal in some sense). One could use a parametric family of such distributions, e.g. a family of symmetrically truncated normal distributions or a family of symmetric beta distributions. However, there seems to be no reason for preferring one such family over another, and it may be easier to work with the nonparametric, or rather semiparametric, model that assumes merely that ϵ has a symmetric distribution.

Hinkley (1975) and Taylor (1985) have investigated power transformations to symmetry. They were primarily concerned with the one-sample problem, which can be expressed as the Box–Cox model

$$(1.5) \quad y^{(\lambda)} = \mu + \sigma\epsilon$$

or as the transform–both–sides model

$$(1.6) \quad y^{(\lambda)} = \mu^{(\lambda)} + \sigma\epsilon.$$

In this case, the mean model fits and the variance is constant for all λ , so symmetry is the only objective. Given data $(y_i, \mathbf{x}_i)$, let $\hat{\beta}(\lambda)$ be the least-squares estimate of β when fitting (1.1)–(1.2) with λ fixed, and let

$$\hat{\epsilon}_i(\lambda) = y_i^{(\lambda)} - \mathbf{x}_i^T \hat{\beta}(\lambda)$$

be the i th residual. The estimators studied by Hinkley and Taylor use a measure of skewness, for example the usual third-moment skewness coefficient. Then $\hat{\lambda}$ is defined as the value of λ such that this skewness measure of $\{\hat{\epsilon}_i(\lambda)\}$ is zero.

The advantage of the Hinkley and Taylor estimators is that when a transformation to symmetry does exist, that is, (1.5) (or equivalently (1.6)) holds for some λ with ϵ symmetrically distributed, then these estimators will consistently estimate the correct λ . Except when $\lambda = 0$, the normal-theory MLE will not be consistent (Hinkley 1975). In the one-sample location problem, the estimator based on the third-moment skewness coefficient is nearly as efficient as the MLE, but using a quantile-based skewness measure leads to an inefficient, albeit robust, estimator of λ (Taylor 1985). However, when the model is more complex, then these symmetry-estimators can be quite inefficient since they do not use the information about λ coming from the mean-variance relationship of $y^{(\lambda)}$ or (in the Box-Cox model) from the relationship between $y^{(\lambda)}$ and $\bar{x}$.

In this paper we apply the transform-to-symmetry estimators of Hinkley and Taylor to the transform-both-sides model, (1.3). We also consider estimators that transform to homoscedasticity, that is, to a null value of some test for heteroscedasticity. Finally, we explore the possibility of combining the two approaches to achieve a transformation to symmetry and homoscedasticity.

It is informative to consider two limiting cases: (1) $\sigma^2(f(\underline{x}_i, \underline{\beta})) \rightarrow 0$, where $\sigma^2(f(\underline{x}_i, \underline{\beta}))$ is the sample variance of $\{f(\underline{x}_i, \underline{\beta}): i = 1, \dots, n\}$, and (2) $\sigma \rightarrow 0$, where $F(\cdot) = F_0(\cdot/\sigma)$ for some fixed F_0 . In case (1) the responses, y_i , become homoscedastic, in fact i. i. d., and all sample information about λ comes from the shape of their distribution. In case (2) the estimator that transforms to a zero third moment of the residuals is nearly efficient (fully efficient if $\lambda = 0$); see theorem 1(b) and the remarks following that theorem.

In case (2) for fixed $\underline{x}_i$, y_i varies over an interval of order $O(\sigma)$. As $\sigma \rightarrow 0$, the transformation $y_i \rightarrow y_i^{(\lambda)}$ will be nearly equal to its linear Taylor approximation over the range of y_i ; therefore $y_i^{(\lambda)}$ and y_i have nearly the same distribution except for a location and scale change (the scale change depends on $\underline{x}_i$ — this is crucial for otherwise the transformation could not induce homoscedasticity), so all sample information about λ comes from the heteroscedasticity of the y_i 's. In case (2) the estimator that transforms to a zero value of the score test for heteroscedasticity is fully efficient; see theorem 4.

There are two reasons for using the transform—both—sides (TBS) model, rather than fitting y to $f(\underline{x}, \beta)$ by nonlinear least—squares. First, transforming to symmetry and homoscedasticity results in a more efficient estimate of β . Second, by "backtransformation", TBS models the skewness and heteroscedasticity of the original response. This is quite important when estimating conditional (given $\underline{x}$) quantiles of y , or when setting prediction intervals for y . In some examples, e.g. the one in section 5, the choice of $\hat{\lambda}$ has relatively little effect on $\hat{\beta}$. However, as discussed further in section 5, the length and shape of prediction intervals is very sensitive to the value of $\hat{\lambda}$.

2. TRANSFORMATIONS TO SYMMETRY

In this section we assume that F is symmetric about 0. When estimating λ we will restrict ourselves to measures of skewness that are M—estimators. This is a large and flexible class, yet sufficiently narrow to allow a compact theoretical development. Suppose we have a sample $y_1, \dots, y_n$ with mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$. Define

$$z_i = (y_i - \hat{\mu})/\hat{\sigma}$$

and let ψ be an odd function. Then we can define the "skewness" of $y_1, \dots, y_n$ as

$$\gamma_\psi = \frac{1}{n} \sum_{i=1}^n \psi(z_i).$$

If $\psi(y) = y^3$, then γ_ψ is the usual third-moment skewness coefficient. This choice of ψ does not give a robust estimator of λ . Taylor (1985) considers other choices of ψ .

Taylor's (1985) estimator can be generalized to the TBS model as follows. For fixed λ , let $\hat{\epsilon}_i(\lambda) = y_i^{(\lambda)} - f^{(\lambda)}(\tilde{x}_i, \hat{\beta}(\lambda))$ where $\hat{\beta}(\lambda)$ is the least-squares estimate of β . Let $\hat{\mu}(\lambda)$ and $\hat{\sigma}(\lambda)$ be the mean and standard deviation of $\{\hat{\epsilon}_i(\lambda)\}$, and define

$$(2.1) \quad z_i(\lambda) = [\hat{\epsilon}_i(\lambda) - \hat{\mu}(\lambda)] / \hat{\sigma}(\lambda).$$

Then $\hat{\lambda}$ is defined by

$$\frac{1}{n} \sum_{i=1}^n \psi\{z_i(\lambda)\} = 0.$$

If λ equals λ_0 , the true value of this parameter, then $\hat{\mu}(\lambda)$ will converge to zero as $n \rightarrow \infty$, so to obtain a consistent estimate of λ_0 it is not necessary to center the $\hat{\epsilon}_i$ when defining $z_i(\lambda)$. However, the asymptotic variance of $\hat{\lambda}$ depends on whether the $z_i(\lambda)$ have mean zero or not. We have followed Taylor (1985) and centered $z_i(\lambda)$ as in (2.1) so that $\hat{\lambda}$ has a simply expressed asymptotic variance, in fact, the same asymptotic variance as Taylor obtained for the location problem; see Theorem 1(b). It would be interesting to know how greatly the introduction of $\hat{\mu}$ affects the asymptotic variance, in particular, to establish a bound of the difference between the asymptotic variance of $\hat{\lambda}$ with and without $\hat{\mu}$. Now let μ be the parameter that $\hat{\mu}$ estimates. Although $\mu = 0$ now, we are going to reparametrize so that μ is not zero.

Now let $\theta = (\lambda, \mu, \beta^T, \sigma)^T$ be the vector of all parameters. The estimator $\hat{\theta}$ defined above can also be defined as the solution to the following M-estimating equation. Define

$$\bar{f}^{(\lambda)}(\beta) = \frac{1}{n} \sum_{i=1}^n f^{(\lambda)}(x_i, \beta).$$

Let

$$(2.2) \quad r_i(\theta) = [y_i^{(\lambda)} - f^{(\lambda)}(\underline{x}_i, \beta)] / \sigma$$

and

$$r_i^c(\theta) = \{y_i^{(\lambda)} - [f^{(\lambda)}(\underline{x}_i, \beta) - \bar{f}^{(\lambda)}(\beta)] - \mu\} / \sigma.$$

The superscript "c" is to indicate that the residual has been centered. Having μ in r_i^c allows us to center $f^{(\lambda)}(\underline{x}_i, \beta)$ at its mean, which will be convenient later. This reparametrization affects the value of μ but does not change the meaning of β , λ , and σ . Define

$$(2.3) \quad \Psi_s(r, r^c, \underline{x}, \theta) = \begin{bmatrix} \psi(r^c) \\ r^c \\ r f_{\beta}^{(\lambda)}(\underline{x}, \beta) \\ [(r^c)^2 - 1] \end{bmatrix},$$

where $f_{\beta}^{(\lambda)}(\underline{x}, \beta) = \partial / \partial \beta f^{(\lambda)}(\underline{x}, \beta)$. The subscript "s" denotes "skewness" estimator and distinguishes this estimator from later ones. Then $\hat{\theta}$ solves

$$(2.4) \quad \sum_{i=1}^n \Psi_s(r_i(\hat{\theta}), r_i^c(\hat{\theta}), \underline{x}_i, \hat{\theta}) = 0.$$

Note that the uncentered residual appears in the third component of Ψ_s that is used to estimate β , but the centered residuals are used to define the other parameters.

Definition (2.3) can be generalized to include robust estimators. The technique we will introduce is essentially Mallows's (1975) estimator; see Li (1985). Let the functions η and χ be odd and even respectively; for robustness both should be bounded. Let $\omega(\underline{\theta}, \underline{x})$ be a weight function taking values in $[0, 1]$; for robustness

$$\omega(\underline{\theta}, \underline{x}) f_{\underline{\beta}}^{(\lambda)}(\underline{x}, \underline{\beta})$$

should be bounded as a function of $\underline{\theta}$ and $\underline{x}$. Then replace (2.3) by

$$\Psi_{\underline{s}}(\mathbf{r}, \mathbf{r}^c, \underline{x}, \underline{\theta}) = \begin{bmatrix} \psi(\mathbf{r}^c) \\ \eta(\mathbf{r}^c) \\ \eta(\mathbf{r}) f_{\underline{\beta}}^{(\lambda)}(\underline{x}, \underline{\beta}) \omega(\underline{\theta}, \underline{x}) \\ \chi(\mathbf{r}^c) \end{bmatrix},$$

and let $\hat{\underline{\theta}}$ be the solution to (2.4) with this new definition of Ψ . The parameter σ is no longer the standard deviation of F but rather is defined as the solution to

$$\int \chi(\epsilon/\sigma) dF(\epsilon) = 0;$$

see Huber (1981, section 5.2).

The consistency and asymptotic normality of $\hat{\underline{\theta}}$ can be investigated using Huber's (1967) results on M-estimation or related techniques. A complete technical discussion of consistency, though potentially interesting, would be a paper in itself. Instead, we discuss constraints on ψ , η , ω , and χ imposed by Huber's conditions. First of all, it is essential that if $\hat{\underline{\theta}}$ in (2.4) is fixed at $\underline{\theta}$, say, then the expectation of the LHS of (2.4) is 0 if and only if θ is the "true" parameter $\underline{\theta}_0$. Therefore, ψ , η , and χ should be nondecreasing on

$[0, \infty)$, ψ and η should be odd, and χ should be even. η and χ might be Huber's (1964) "Proposal 2," e.g.

$$(2.5) \quad \eta(x) = \min\{|x|, b\} \text{sign}(x)$$

and

$$(2.6) \quad \chi(x) = \eta^2(x) - K,$$

where $K = \int_{-\infty}^{\infty} \eta^2(x) d\Phi(x)$ and Φ is the standard normal distribution. If

$$\eta(x) = \text{sign}(x)$$

then $\hat{\beta}$ and $\hat{\mu}$ will be consistent if $\hat{\lambda}$ is fixed at an arbitrary constant, e.g. 1; see Carroll and Ruppert (1984). If $\hat{\beta}$ and $\hat{\mu}$ are consistent, then $\hat{\lambda}$ will be consistent if

$$(2.7) \quad E\{\psi(r_1^c(\lambda, \mu_0, \beta_0^T, \sigma_0)^T)\} = 0$$

if and only if $\lambda = \lambda_0$. If F has a positive density on a nondegenerate interval $(-a, a)$, then it can be shown that (2.7) holds if ψ is odd, nondecreasing, and strictly increasing in a neighborhood of 0.

If $\eta(x) = \text{sign}(x)$ and $\psi(x) = x$, then $\hat{\lambda}$ is determined so that the mean and median of the residuals are both equal to 0 — here the centering of the residuals is crucial.

However, Taylor (1985) found this estimate rather efficient.

Later, we show that $\psi(x) = x^3$ is optimal if F is normal. For robustness, it seems sensible to modify x^3 slightly and use

$$(2.8) \quad \psi(x) = \min\{b^3, |x|^3\} \text{sign}(x),$$

where b could be the same value as used for Huber's Proposal 2. The efficiency of (2.8) as a function of b is explored in example 5.1. Using (2.5), (2.6), and (2.8) can be considered a generalization of Proposal 2.

Now we will simply assume consistency and asymptotic normality and study the form of the asymptotic variance matrix.

Let θ_0 be the true value of the parameter θ . The asymptotic variance matrix of $\hat{\theta}$ is

$$V_s = B_s^{-1} A_s B_s^{-T},$$

where

$$(2.9) \quad A_s = E \sum_{i=1}^n \Psi_s(r_i(\theta_0), r_i^c(\theta_0), x_i, \theta_0) \Psi_s^T(r_i(\theta_0), r_i^c(\theta_0), x_i, \theta_0)$$

$$(2.10) \quad B_s = -(\partial/\partial\theta) E \sum_{i=1}^n \Psi_s(r_i(\theta_0), r_i^c(\theta_0), x_i, \theta_0),$$

and where for any invertible matrix M we define $M^{-T} = (M^{-1})^T$. V_s can be consistently estimated by replacing θ_0 with $\hat{\theta}$ and replacing the expected sums in (2.9) and (2.10) with their observed values; call this estimator $\hat{V}_s$.

If we only wished to estimate V_s , then there would be no need to develop a formula for V_s . However, we will see that some insight can be gained by examining the form of V_s .

Now assume that model (1.3) holds with ϵ symmetrically distributed and that $\hat{\theta}$ is asymptotically normal with variance matrix V_s , or more precisely,

$$V_s^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, I), \text{ as } n \rightarrow \infty.$$

Note that the asymptotic variance matrix is actually a sequence of matrices; the n -th matrix in the sequence is the large-sample theory approximation to the variance of $\hat{\theta}$ calculated from the first n observations. Of course, if $\{x_i\}$ is suitably behaved, e. g. an i.i.d. sequence, then V_S will be simply n^{-1} times a constant matrix.

We have the following results.

Theorem 1: (a) $(\hat{\lambda}, \hat{\beta}, \hat{\mu})$ and $\hat{\sigma}$ are, in general, asymptotically correlated, but the asymptotic variance of $(\hat{\lambda}, \hat{\beta}, \hat{\mu})$ is the same as if σ were known and did not need to be estimated. Therefore we can ignore the nuisance parameter σ and focus attention on the variance of $(\hat{\lambda}, \hat{\beta}, \hat{\mu})$.

(b) The asymptotic distribution of $(\hat{\lambda}, \hat{\mu})$ is the same as it would be if β were known. In particular, the asymptotic distribution of $(\hat{\lambda}, \hat{\mu})$ is the same as Taylor (1985) obtained for the location problem.

(c) In general, $\hat{\beta}$ is asymptotically correlated with $(\hat{\lambda}, \hat{\mu})$ and the asymptotic distribution of $\hat{\beta}$ is different than it would be if (λ, μ) were known.

Since Taylor (1985) has made a detailed study of the choice of ψ , it is very useful to obtain the same asymptotic variance so that his conclusions carry over. The proofs of all theorems are in section 6.

3. TRANSFORMATIONS TO HOMOSCEDASTICITY

As in section 2, we assume that model (1.3) holds with $y^{(\lambda)}$ given by (1.1). In this section, however, the distribution F need not be symmetric, but we require that $E\eta(\epsilon/\sigma) = 0$, e. g., if $\eta(\epsilon) = \text{sign}(\epsilon)$ ($\eta(\epsilon) = \epsilon$) then the median (mean) of ϵ must be 0. If we transform y to $y^{(\lambda)}$ and estimate β and σ by least-squares, then we expect the squared

residuals to be positively (negatively) correlated with the fitted values if $\lambda > \lambda_0$ ($\lambda < \lambda_0$). This suggests using the correlation between the squared residuals and the fitted values of $f(\mathbf{x}_i, \beta)$ to test if $\lambda = \lambda_0$, and estimating λ by the value giving zero correlation. Such tests were proposed by Anscombe (1961). Following Bickel (1978) we will introduce more general tests for heteroscedasticity.

We will not need the parameter μ , so define $\Delta = (\lambda, \sigma, \beta)$. Since $\mathbf{r}_i(\theta)$ defined by (2.2) does not depend on μ , we now write $\mathbf{r}_i(\Delta)$ instead of $\mathbf{r}_i(\theta)$. Now we also assume that ω does not depend on μ and write $\omega(\Delta, \mathbf{x})$.

Let $b(\cdot)$ be a monotonically increasing function. Define

$$\bar{b}(\beta) = n^{-1} \sum_{i=1}^n b[f(\mathbf{x}_i, \beta)]$$

and

$$S(\Delta) = \sum_{i=1}^n \{b[f(\mathbf{x}_i, \beta)] - \bar{b}(\beta)\} \chi[\mathbf{r}_i(\Delta)].$$

Bickel (1978) proposed using S to test for heteroscedasticity. We will define $\hat{\Delta}$ by simultaneously solving

$$S(\hat{\Delta}) = 0,$$

along with equations that estimate β and σ . More explicitly, let

$$(3.1) \quad \Psi_h(\mathbf{r}, \mathbf{x}, \Delta) = \begin{bmatrix} \{b[f(\mathbf{x}, \beta)] - \bar{b}(\beta)\} \chi(\mathbf{r}) \\ \chi(\mathbf{r}) \\ \eta(\mathbf{r}) f_{\beta}^{(\lambda)}(\mathbf{x}, \beta) \omega(\Delta, \mathbf{x}) \end{bmatrix}.$$

Then $\hat{\Delta}$ solves

$$(3.2) \quad \sum_{i=1}^n \Psi_h(r_i(\hat{\Delta}), \mathbf{x}_i, \hat{\Delta}) = 0.$$

If (1.4) holds exactly then

$$SD(y_i^{(\lambda)}) = \sigma \{1 + \log[f(\mathbf{x}_i, \beta)](\lambda - \lambda_0) + o(\lambda - \lambda_0)\}$$

as $\lambda \rightarrow \lambda_0$, and so the score test for heteroscedasticity uses $b(\mathbf{x}) = \log(\mathbf{x})$; see Cook and Weisberg (1983). This choice of $b(\cdot)$ is approximately optimal for small σ ; see theorem 3.

Now assume that $\hat{\Delta}$ is consistent and asymptotically normal with asymptotic variance $\mathbf{V}_h = \mathbf{B}_h^{-1} \mathbf{A}_h \mathbf{B}_h^{-T}$ where

$$(3.3) \quad \mathbf{B}_h = \partial / \partial \Delta \mathbf{E} \sum_{i=1}^n \Psi_h(r_i(\Delta), \mathbf{x}_i, \Delta),$$

and

$$(3.4) \quad \mathbf{A}_h = \mathbf{E} \sum_{i=1}^n \Psi_h(r_i(\Delta), \mathbf{x}_i, \Delta) \Psi_h^T(r_i(\Delta), \mathbf{x}_i, \Delta).$$

Even when F is asymmetric, $\hat{\Delta}$ is consistent and asymptotically normal with variance $\mathbf{V}_h$. However, the form of $\mathbf{V}_h$ can be simplified if F is assumed to be symmetric, and, therefore, we make this assumption for the remainder of this paper.

Theorem 2: (a) In general, $(\hat{\lambda}, \hat{\sigma})$ and $\hat{\beta}$ are asymptotically correlated and the asymptotic variance of $\hat{\beta}$ is different than it would be if $(\hat{\lambda}, \hat{\sigma})$ were known. However, the asymptotic variance of $(\hat{\lambda}, \hat{\sigma})$ is the same as if $\hat{\beta}$ were known. Therefore, we can focus attention on $(\hat{\lambda}, \hat{\sigma})$, ignoring $\hat{\beta}$.

(b) The asymptotic variance of $\hat{\lambda}$ is the same as it would be if $\hat{\sigma}$ were known.

The asymptotic variance of $\hat{\lambda}$ becomes simple under "small- σ " asymptotics where the scale parameter σ converges to 0. Several authors (for example, Bickel and Doksum 1982 and Carroll and Ruppert 1981, 1984) have applied "small- σ " asymptotics to similar transformation problems. We will study the asymptotic formula for the variance, V_h , with n fixed and σ converging to 0.

Theorem 3: Under small- σ asymptotics,

- (a) *the asymptotic variance of $\hat{\beta}$ is the same as it would be if (λ, σ) were known,*
- (b) *the asymptotic variance of $\hat{\lambda}$ converges to*

$$(3.5) \quad \frac{E\chi^2(\epsilon/\sigma) \sum_{i=1}^n \{b[f(x_i, \beta)] - \bar{b}(\beta)\}^2}{\left[E[(\epsilon/\sigma)\dot{\chi}(\epsilon/\sigma)] \sum_{i=1}^n \{b[f(x_i, \beta)] - \bar{b}(\beta)\} \log[f(x_i, \beta)] \right]^2},$$

and

- (c) *the optimal choice of $b(\cdot)$ becomes*

$$(3.6) \quad b(x) = \log(x).$$

With this choice of $b(\cdot)$, (3.5) becomes

$$(3.7) \quad \frac{E\chi^2(\epsilon/\sigma)}{\left[E[(\epsilon/\sigma)\dot{\chi}(\epsilon/\sigma)] \right]^2 \sum_{i=1}^n \{b[f(x_i, \beta)] - \bar{b}(\beta)\}^2}.$$

The variance in (3.7) depends on χ only through the ratio $E\chi^2(\epsilon/\sigma)/E[(\epsilon/\sigma)\chi(\epsilon/\sigma)]$. Therefore, the relative efficiency of two possible choices of χ is the same as in the simpler scale parameter problem; see Huber (1964,1981) for a discussion of the later.

The choice $b(x) = \log(x)$ is highly nonrobust, especially because observations with small predicted values can be very influential. Indeed, it has been our experience with modeling the variances of heteroscedastic data that such observations are often anomalous, being more variable than predicted by models fitting the remainder of the data. If $b(x) = \log(x)$ is used, then the data should be carefully scrutinized. Alternatively, b could be a suitably truncated version of $\log(x)$. Without truncation of the log function, observations with unusually small fitted values may have high, and deleterious, influence on $\hat{\Delta}_{\sim}$.

Let $\hat{\Delta}_{\text{ML}} = (\hat{\lambda}_{\text{ML}}, \hat{\sigma}_{\text{ML}}, \hat{\beta}_{\text{ML}})$ be the normal-theory maximum-likelihood estimator, that is, the estimator that maximizes the likelihood assuming that $F(\cdot) = \Phi(\cdot/\sigma)$. Although this assumption is, strictly-speaking false for $\lambda \neq 0$, it can hold in the limit as $\sigma \rightarrow 0$. For σ fixed, Hernandez and Johnson (1980) have given a detailed study of the normal-theory MLE for the Box-Cox model, and the extension to the transform-both-sides model is straightforward.

Carroll and Ruppert (1984) have shown that as $\sigma \rightarrow 0$, $\hat{\beta}_{\text{ML}}$ has the same asymptotic variance as when λ is known, and theorem 3(a) shows that $\hat{\Delta}_{\sim}$ has the same behavior. In fact, more is true:

Theorem 4: As $\sigma \rightarrow 0$, the likelihood equations satisfied by $\hat{\Delta}_{\text{ML}}$ converge to (3.5) with $b(x) = \log(x)$, $\eta(x) = x$, and $\chi(x) = x^2 - 1$, so $\hat{\Delta}_{\sim}$ and $\hat{\Delta}_{\text{ML}}$ are asymptotically equivalent as $\sigma \rightarrow 0$.

The approximation (1.4) becomes exact as the scale parameter $\sigma \rightarrow 0$. For this reason, we could conjecture that as $\sigma \rightarrow 0$, λ behaves as a parameter in a variance function. Therefore, Theorem 3(a) is not unexpected; in heteroscedastic regression models the estimate of β using estimated reciprocal variances as weights is asymptotically equivalent to the estimate using the true variances (Carroll and Ruppert 1982, 1988).

4. TRANSFORMATIONS TO SYMMETRY AND HOMOSCEDASTICITY

Model (1.3) postulates a single transformation to both symmetry and homoscedasticity. While there may be a transformation to symmetry and another transformation to homoscedasticity, there is no guarantee that a single transformation will induce both. Let $\hat{\theta}_s = (\hat{\lambda}_s, \hat{\mu}_s, \hat{\beta}_s^T, \hat{\sigma}_s)^T$ be the transformation to symmetry estimator proposed in section 2, and let $\hat{\Delta}_h = (\hat{\lambda}_h, \hat{\sigma}_h, \hat{\beta}_h^T)^T$ be the transformation to homoscedasticity of section 3. By comparing $\hat{\lambda}_s$ to $\hat{\lambda}_h$ we can test the hypothesis that there is a transformation to both symmetry and homoscedasticity. Such tests will be discussed in this section.

If we accept the hypothesis that a transformation to both symmetry and homoscedasticity exists, then we could estimate it by a weighted average of $\hat{\lambda}_s$ and $\hat{\lambda}_h$. However, in some sampling situations either $\hat{\lambda}_s$ or $\hat{\lambda}_h$ is unstable. For example, $\hat{\lambda}_s$ is highly variable if σ is small so that transformations have only minor effects on distributional shape. In practice, we truncate the estimates of λ at ± 1 to avoid instability. There appears to be no satisfactory way of combining the estimates if one of them has been truncated. A better approach, the one taken here, is to solve a weighted average of equations (2.3)–(2.4) and (3.1)–(3.2) that define $\hat{\theta}_s$ and $\hat{\Delta}_h$, respectively.

Let $\underline{\theta}_s$ be the limit of $\hat{\underline{\theta}}_s$ and let $\underline{\Delta}_h$ be the limit of $\hat{\underline{\Delta}}_h$ as $n \rightarrow \infty$ (we assume that these limits exist). If $\lambda_s = \lambda_h$, then $\underline{\beta}_s = \underline{\beta}_h$ and $\sigma_s = \sigma_h$, but in general $\underline{\beta}_s \neq \underline{\beta}_h$ and $\sigma_s \neq \sigma_h$ if $\lambda_s \neq \lambda_h$.

To test

$$(4.1) \quad H_0: \lambda_s = \lambda_h \text{ versus } H_1: \lambda_s \neq \lambda_h,$$

we need the joint limiting distribution of $\hat{\lambda}_s$ and $\hat{\lambda}_h$. Let $\hat{\underline{\theta}}_J = (\hat{\underline{\theta}}_s^T, \hat{\underline{\Delta}}_h^T)^T$ be the joint estimator of $\underline{\theta}_s$ and $\underline{\Delta}_h$. Then $\hat{\underline{\theta}}_J$ is an M-estimator solving (2.4) and (3.2) simultaneously. If we define $\underline{\Psi}_J(\hat{\underline{\theta}}_J, \underline{x}_i)$ by stacking $\underline{\Psi}_s(r_i(\hat{\underline{\theta}}_s), r_i^c(\hat{\underline{\theta}}_s), \underline{x}_i, \hat{\underline{\theta}}_s)$ and $\underline{\Psi}_h(r_i(\hat{\underline{\Delta}}_h), \underline{x}_i, \hat{\underline{\Delta}}_h)$, then the asymptotic variance of $\hat{\underline{\theta}}_J$ is estimated by

$$\hat{\underline{V}}_J = \hat{\underline{B}}_J^{-1} \hat{\underline{A}}_J \hat{\underline{B}}_J^{-T},$$

where

$$\hat{\underline{B}}_J = \partial / \partial \underline{\theta} \sum_{i=1}^n \underline{\Psi}_J(\hat{\underline{\theta}}_J, \underline{x}_i)$$

and

$$\hat{\underline{A}}_J = \sum_{i=1}^n \underline{\Psi}_J(\hat{\underline{\theta}}_J, \underline{x}_i) \underline{\Psi}_J^T(\hat{\underline{\theta}}_J, \underline{x}_i).$$

Let $\hat{\sigma}^2(\hat{\lambda}_s)$, $\hat{\sigma}^2(\hat{\lambda}_h)$, and $\hat{\sigma}(\hat{\lambda}_s, \hat{\lambda}_h)$ be the estimated variances and covariance of the indicated estimators. Then the test statistic

$$(4.2) \quad (\hat{\lambda}_s - \hat{\lambda}_h) / \{ \hat{\sigma}^2(\hat{\lambda}_s) - 2 \hat{\sigma}(\hat{\lambda}_s, \hat{\lambda}_h) + \hat{\sigma}^2(\hat{\lambda}_h) \}^{1/2}$$

is asymptotically standard normal and can be used to test (4.1).

If the null hypothesis is rejected, then an alternative model should be found. One possibility is a heteroscedastic regression model without a transformation; this is appropriate if the untransformed data are symmetric, but with a nonconstant variance. Another possibility is to combine the transform–both–sides model with a nonconstant variance function. Both types of models are discussed in detail in Carroll and Ruppert (1988).

Suppose that we accept H_0 at a sufficiently large significance level that we are willing to proceed as though H_0 were true. If H_0 is true, then $\lambda_s = \lambda_h$ so let λ_0 be the common value of λ_s and λ_h . It is convenient to redefine Ψ_h so that its components align with those of Ψ_s ; the original definition of Ψ_h is useful when proving the theorems of section 3 (see section 6) but is cumbersome now. Let

$$\Psi_h(r, r^c, \underline{x}, \underline{\theta}) = \begin{bmatrix} b^c[f(\underline{x}, \underline{\beta})] \chi(r) \\ \eta(r^c) \\ \eta(r) f_{\underline{\beta}}^{(\lambda)}(\underline{x}_i, \underline{\beta}) \omega(\theta, x) \\ \chi(r) \end{bmatrix}$$

In this redefinition the component $\eta(r^c)$ has been added to estimate μ and the other components have been rearranged. The functions Ψ_s and Ψ_h differ only in their first component, the one used to estimate λ . Ψ_h now depends on r^c as well as r and on θ instead of just $\underline{\Delta}$.

Another possibility, one that we use in example 2 of section 5, is to omit the component $\eta(r^c)$ from Ψ_s and Ψ_h and to replace r^c by r and $\underline{\theta}$ by $\underline{\Delta}$ throughout; this means that μ is not estimated. When defining $\hat{\lambda}_{hs}$, the rationale for introducing μ is less clear. Recall that μ was introduced so that $\hat{\lambda}_s$ would have the same asymptotic variance $\hat{\lambda}$ as Taylor (1985) studied. $\hat{\lambda}_{hs}$ will not have this variance whether μ is used or not. Let

$\hat{\Delta}_u = (\hat{\lambda}_u, \hat{\sigma}_u, \hat{\beta}_u)$ be the combined heteroscedasticity–skewness estimator without centering by μ ("u" means uncentered).

To estimate λ define

$$\Psi_{hs}(r, r^c, x, w, \theta) = w \Psi_s(r, r^c, x, \theta) + (1 - w) \Psi_h(r, r^c, x, \theta),$$

$0 \leq w \leq 1$. Then define $\hat{\theta}_{hs}$ as the solution to

$$(4.3) \quad \sum_{i=1}^n \Psi_{hs}(r_i(\hat{\theta}_{hs}), r_i^c(\hat{\theta}_{hs}), x_i, \hat{\theta}_{hs}) = 0.$$

If $w = 1$, then $\hat{\theta}_{hs} = \hat{\theta}_s$. If $w = 0$, then $\hat{\theta}_{hs}$ equals $\hat{\Delta}_h$ except that $\hat{\theta}_{hs}$ has the extra component that estimates μ . Presumably a value of w strictly between 0 and 1 would be better than either of these extremes. We propose letting w minimize the asymptotic variance of $\hat{\lambda}_{hs}$.

Let $\hat{V}_{hs}(w)$ be the estimated variance matrix of $\hat{\theta}_{hs}$ with w fixed. $\hat{V}_{hs}$ is obtained in the same way as $\hat{V}_s$ in section 2 (or $\hat{V}_h$ of section 3) but with Ψ_s replaced by Ψ_{hs} . More explicitly, let $\hat{A}_h$ and $\hat{B}_h$ be defined by (3.3) and (3.4) but with the new definition of Ψ_h .

Define

$$(4.4) \quad \hat{A}_{s \cdot h} = \sum_{i=1}^n \Psi_s(r_i(\hat{\theta}_{hs}), r_i^c(\hat{\theta}_{hs}), x_i, \hat{\theta}_{hs}) \Psi_h^T(r_i(\hat{\theta}_{hs}), r_i^c(\hat{\theta}_{hs}), x_i, \hat{\theta}_{hs})$$

Then

$$\hat{V}_{hs} = \hat{B}_{hs}^{-1}(w) \hat{A}_{hs}(w) \hat{B}_{hs}^{-T}(w),$$

where

$$\hat{\mathbf{B}}_{\mathcal{H}\mathcal{S}}(\mathbf{w}) = \mathbf{w}\hat{\mathbf{B}}_{\mathcal{S}} + (1-\mathbf{w})\hat{\mathbf{B}}_{\mathcal{H}}$$

and

$$\hat{\mathbf{A}}_{\mathcal{H}\mathcal{S}}(\mathbf{w}) = \mathbf{w}^2\hat{\mathbf{A}}_{\mathcal{S}} + (1-\mathbf{w})^2\hat{\mathbf{A}}_{\mathcal{H}} + \mathbf{w}(1-\mathbf{w})(\hat{\mathbf{A}}_{\mathcal{S}\cdot\mathcal{H}} + \hat{\mathbf{A}}_{\mathcal{S}\cdot\mathcal{H}}^{\mathbf{T}}).$$

Define $\hat{\mathcal{C}} = (\hat{\mathbf{A}}_{\mathcal{S}}, \hat{\mathbf{A}}_{\mathcal{H}}, \hat{\mathbf{A}}_{\mathcal{S}\cdot\mathcal{H}}, \hat{\mathbf{B}}_{\mathcal{S}}, \hat{\mathbf{B}}_{\mathcal{H}})$ and let $\hat{\mathbf{w}}(\hat{\mathcal{C}})$ be the value of $\mathbf{w}$ that minimizes $\hat{\sigma}^2(\hat{\lambda})$ of $\hat{\mathbf{Y}}_{\mathcal{H}\mathcal{S}}(\mathbf{w})$.

$\hat{\mathcal{C}}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{H}\mathcal{S}}$ are both M-estimates, defined jointly by (2.9), (2.10), (3.3), (3.4), (4.4), and (4.3), where $\mathbf{w} = \hat{\mathbf{w}}(\hat{\mathcal{C}})$ in (4.3). When estimating the asymptotic variance of $\hat{\boldsymbol{\theta}}_{\mathcal{H}\mathcal{S}}$ it would be burdensome if we needed to examine the joint limiting distribution of $\hat{\boldsymbol{\theta}}_{\mathcal{H}\mathcal{S}}$ and $\hat{\mathcal{C}}$. Fortunately, this is not the case. The following theorem shows that for large enough sample sizes we can act as if $\hat{\mathbf{w}}$ were fixed and use $\hat{\mathbf{Y}}_{\mathcal{H}\mathcal{S}}(\hat{\mathbf{w}})$ to estimate the variance of $\hat{\boldsymbol{\theta}}_{\mathcal{H}\mathcal{S}}$.

Theorem 5: The asymptotic variance of $\hat{\boldsymbol{\theta}}_{\mathcal{H}\mathcal{S}}$ is the same as it would be if $\mathcal{C}$ were known.

A referee has suggested testing for symmetry after using the variance-stabilizing transformation (or vice versa). While we will not pursue this idea here, it would be interesting to compare this proposal with test (4.2).

5. EXAMPLES

Example 5.1. This example is a small numerical study of particular robust class of $\hat{\lambda}_{\mathcal{S}}$.

Let $\eta_{\mathbf{b}}$, $\chi_{\mathbf{b}}$, and $\psi_{\mathbf{b}}$ be given by (2.5), (2.6) and (2.8).

$$\log(\mathbf{y}) = \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

and $\epsilon \sim F$ where F is a contaminated normal distribution, $F(x) = (1-\gamma)\Phi(x) + \gamma\Phi(x/3)$. Since $f(x, \beta) \equiv 0$, this is the location problem studied by Taylor (1985), but by Theorem 1(b) the results here study apply to general f . Table 1 gives the asymptotic variance of $\hat{\lambda}_s$ for $b = 1, 2, 3, 4, 5, 8$, and 12.

For the normal distribution, any b greater than or equal to 4 is near fully efficient, while $b = 3$ is 94% $(=.667/.711)100\%$ efficient. When Huber's proposal 2 is used to estimate a location parameter, typically b is between 1 and 2. Here $b = 1$ and $b = 2$ have efficiency 11.7% and 60.1%, respectively. It might be useful to use one value of b in ψ_b and another in η_b and χ_b , but we will not pursue this idea.

When the amount of contamination is small ($\gamma = .001$ or $\gamma = .01$), then $b = 3$ is optimal among the six values of b in the table. Moreover, the variance increases dramatically as b increases beyond 3.

When $\gamma \geq .1$, then large values of b are optimal. Also, when $\gamma \geq .2$ then contamination is beneficial, since the variance of $\hat{\lambda}$ for the optimal b is smaller than when there is no contamination. (The contamination does increase the variance of $\hat{\mu}$.) Folklore says that the information about asymmetry is in the tails. Here is evidence that the folklore is correct. However, we have seen that for very small amounts of contamination ($\gamma \leq .01$) it is best to downweight the extreme tails by using b around 3 or 4.

Overall, $b = 4$ seems a good choice, having reasonable high efficiency for all values of γ . The last column of Table 1 gives the variance of $\hat{\lambda}$ maximized over the values of γ in the preceding six columns, and $b = 4$ minimizes this maximum variance.

5.2. Salmon Data

As a numerical example we use the Skeena River sockeye salmon data from Ricker and Smith (1975). These data consist of yearly values of recruits (R) and spawners (S) from this fishery. In a given year, the value of S is the total number of fish that spawn,

that is, the number of fish returning to the river to spawn minus the catch. The value of R for any year is the total number of fish produced by spawning this year that eventually (usually after four years) return to spawn themselves. The data are listed and plotted in Carroll and Ruppert (1988). The transform-both-sides model as well as related heteroscedastic regression models have been fit to these data by Ruppert and Carroll (1985), Carroll and Ruppert (1987, 1988), and Carroll, Cressie, and Ruppert (1987). After examining a robust estimator and influence diagnostics, Carroll and Ruppert (1987) suggest that one year, 1951, should be eliminated from the analysis. The number of recruits was very low this year because of a rock slide, and as a result, this observation has tremendous influence on the fit. The year 1955 has a low value of S since the spawning population that year came from 1951, but the recruitment in 1955 is nearly what would be expected given this low value of S . We have retained 1955.

For simplicity and because these data have already been carefully examined for influential observations and outliers, we use least-squares estimation, $\eta(x) = x$. For the skewness estimator we use $\psi(x) = x^3$. With these choices of η and ψ , $\hat{\beta}$ and $\hat{\lambda}$ can (and will) be estimated without simultaneous estimation of the scale parameter σ . Also for simplicity, we use $\hat{\Delta}_u$, the heteroscedasticity-skewness estimator without the centering constant. For $\hat{\lambda}_h$ we use $b(x) = \log(x)$.

As a regression model we use the Ricker (1954) model,

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

though other models exist that appear to fit equally well (Ruppert and Carroll 1985 and Carroll and Ruppert 1988).

The most prominent aspect of the data is their heteroscedasticity. In fact, $\hat{\lambda}_h = -.86$, suggesting a fairly radical transformation to correct for the nonconstant variance. The data exhibit moderate right skewness and $\hat{\lambda}_s$ is .45. Since $\hat{\lambda}_s$ and $\hat{\lambda}_h$ are

rather different, the next step should be to test whether $\lambda_s = \lambda_h$. The appropriate test statistic, t , is given by (4.2). The variance matrix of $(\hat{\beta}_s, \hat{\beta}_h, \hat{\lambda}_s, \hat{\lambda}_h)$ was calculated numerically and the result was $t = 1.39$, so we decided to proceed as though $\lambda_s = \lambda_h$ (at least for illustrative purposes).

Although the test that $\lambda_s = \lambda_h$ does not reject this hypothesis, it is worrisome that $\hat{\lambda}_s$ and $\hat{\lambda}_h$ are so different. However, different choices of λ do not lead to substantially different $\hat{\beta}$. For example, $\hat{\beta}_s$ is $(3.8, 8.7 \times 10^{-4})$ while $\hat{\beta}_h$ is $(4.1, 11.6 \times 10^{-4})$. Going from $\hat{\beta}_s$ to $\hat{\beta}_h$ increases each component of $\hat{\beta}$ by less than one standard deviation. Also, the increase in the first component tends to offset increase in the second component, so the fitted curves are similar, especially relative to the spread in the data. Not transforming ($\lambda = 1$) leads to $\hat{\beta} = (4.0, 8.5 \times 10^{-4})$ that is similar to $\hat{\beta}_h$ and $\hat{\beta}_s$. One might ask, "Is it worth the bother to transform?" The answer is "yes". First, there are examples where the untransformed data contain wild outliers due to the skewness or heteroscedasticity. Without proper transformation these outliers seriously degrade the estimates. Moreover, modeling skewness and heteroscedasticity is essential if one is predicting future responses. Carroll and Ruppert (1988, Table 4.4) construct 90% prediction intervals for recruitment. When $S = 1066$, the largest value of S in the data set, then the prediction intervals are (744, 8569), (752, 3331), (678, 2531), and (534, 2381), for $\lambda = -.86, -.3, .5$, and 1 , respectively. If $S = 511$ (a typical value of S), then the prediction intervals are (660, 4091), (635, 2564), (498, 2170), and (263, 2110) for the above values of λ . Clearly, these intervals vary drastically with λ , especially for $S = 1066$. We believe that $\lambda = -.3$, which is near $\hat{\lambda}_{hs} = -.26$, gives realistic prediction intervals, since it accounts for both asymmetry and heteroscedasticity. Not transforming leads to prediction intervals that fit the data poorly; see Carroll and Ruppert (1988, example 4.1) for further discussion. In fisheries management, having realistic prediction intervals is probably as important as estimating β .

The weight that minimizes the estimated asymptotic variance of $\hat{\lambda}$ is $\hat{w} = .85$, giving $\hat{\lambda}_{hs} = -.26$. This agrees well with the maximum likelihood estimate of $-.20$ found by Carroll and Ruppert (1987).

Recall that the asymptotic variance of $\hat{\lambda}_{hs}$ is the same as if the weight w were known. For this small data set ($n = 27$) can we treat w as fixed? To answer this question we performed a bootstrap experiment. The bootstrap data were generated from (1.5) using $(\lambda, \beta) = (\hat{\lambda}_{hs}, \hat{\beta}_{hs})$. The errors, ϵ , were generated by sampling with replacement from the "symmetrized residuals", $\{\hat{\epsilon}_1, \dots, \hat{\epsilon}_n, -\hat{\epsilon}_1, \dots, -\hat{\epsilon}_n\}$, where

$$\hat{\epsilon}_i = y_i^{(\hat{\lambda}_{hs})} - f^{(\hat{\lambda}_{hs})}(x_i, \hat{\beta}_{hs}).$$

Under the bootstrap distribution the null hypothesis that $\lambda_s = \lambda_h$ is true. If we did not "symmetrize" the residuals, then error distribution would be somewhat skewed and in the bootstrap sampling distribution the "true" λ_s and λ_h would be unequal. Thus the bootstrap would not mimick the assumed distribution of the data. We used 200 bootstrap repetitions. We also calculated $\hat{\lambda}_s$ and $\hat{\lambda}_h$ from the bootstrap samples, but these proved occasionally unstable so we truncated them at ± 1 . $\hat{\lambda}_h$ was truncated about 1 sample in 6 and $\hat{\lambda}_s$ about 1 sample in 12. For this reason we did not calculate bootstrap means and standard deviations for $\hat{\lambda}_h$ and $\hat{\lambda}_s$. Fortunately, $\hat{\lambda}_{hs}$ was stable and was in the interval $(-1, 1)$ on all 200 bootstrap samples.

The bootstrap mean and standard errors are given in Table 1 along with the original estimates and the standard errors from $\hat{V}_{hs}$. The bootstrap mean and standard deviation of $\hat{w}$ are also given. It is clear from the standard deviation of $\hat{w}$ (.22) that $\hat{w}$ is far from constant. Moreover, the bootstrap standard error of $\hat{\lambda}_{hs}$, which is .45, is considerably larger than the standard error from $\hat{V}_{hs}$, .279. This suggests that the approximation of treating $\hat{w}$ as fixed is adequate only for larger sample sizes. This phenomenon is similar to

one observed by Carroll (1979). Switzer's (1970) adaptive location estimator is defined as the 5%, 10%, or 25% trimmed mean, whichever has the smallest estimated variance. Asymptotically, Switzer's estimator behaves as if the minimum–variance trimming proportion were known. Carroll found that for small samples, Switzer's estimator is considerably more variable than asymptotics suggest.

Theorem 3(a) suggests estimating the variance matrix of $\hat{\beta}$ from the nonlinear least–squares fit of $y_i^{(\lambda)}$ on $f^{(\lambda)}(\mathbf{x}_i, \beta)$, with $\lambda = \hat{\lambda}_{\text{hs}}$ treated as fixed. The estimated variance matrix is the

$$\hat{\mathbf{V}}_{\text{nl}} = \hat{\sigma}^2 \left[\sum_{i=1}^n f_{\beta}^{(\lambda)}(\mathbf{x}_i, \beta) [f_{\beta}^{(\lambda)}(\mathbf{x}_i, \beta)]^T \right]^{-1},$$

where $\hat{\sigma}^2$ is the mean square for error. The nonlinear least–squares standard errors are also given in Table 1.

To study the asymptotic standard errors for larger sample sizes, we conducted a second bootstrap experiment, this time with $n = 81$. The results, which appear in Table 1, show that for $n = 81$ the standard error of $\hat{\lambda}_{\text{hs}}$ from $\hat{\mathbf{Y}}_{\text{hs}}(\hat{\mathbf{w}})$ agrees to two decimals with the bootstrap standard error. For samples of this size, treating $\hat{\mathbf{w}}$ as fixed seems to be an adequate approximation. For $\hat{\beta}_{\text{hs}}$, the nonlinear least–squares standard errors are closer to the bootstrap standard errors than the standard errors from $\hat{\mathbf{Y}}_{\text{hs}}(\hat{\mathbf{w}})$. A possible reason for this is that the nonlinear least–squares standard errors and the bootstrap standard errors both use the assumption that the errors, ϵ_i , are homoscedastic. The estimate $\hat{\mathbf{Y}}_{\text{hs}}(\hat{\mathbf{w}})$ is related to the jackknife estimate of variance that does not assume homoscedasticity; see Wu (1986).

5.3. Implementation

We programmed these estimators in the matrix language GAUSS. In principle, it should be possible to use a nonlinear equation solving method, such as the Newton–Raphson method, to solve for $\hat{\beta}$, $\hat{\lambda}_{\text{hs}}$, and $\hat{w}$ simultaneously. These methods are potentially fast and efficient but proved to be very unreliable, having trouble with both non–convergence and local extrema. Instead, we used a bisection approach for λ that is similar to one used by Box and Cox (1964). For fixed λ and w , $\hat{\beta}(\lambda, w)$ was determined using the Gauss–Newton method. This gave a value of the first component of Ψ_{hs} , a positive value indicating that a more severe transformation (i. e. λ smaller) was needed and a negative value indicating the opposite. Then λ was adjusted until equation (4.3) was solved for this fixed value of w . The standard deviation of $\hat{\lambda}_{\text{hs}}$ was then found from $\hat{V}_{\text{hs}}(w)$.

This process was repeated for all w on a grid of values between 0 and 1. The grid was repeatedly refined around the minimizing value of w until sufficient accuracy was obtained.

6. PROOFS

We will make repeated use of the following result.

Lemma 1: Let $\tilde{A}$ and $\tilde{B}$ be $q \times q$ matrices such that $\tilde{B}$ is nonsingular and

$$\tilde{B} = \begin{bmatrix} \tilde{B}_{11} & 0 \\ \tilde{B}_{21} & \tilde{B}_{22} \end{bmatrix}$$

where $\tilde{B}_{11}$ is $q_1 \times q_1$ and $\tilde{B}_{22}$ is $q_2 \times q_2$ ($q_1 + q_2 = q$). Partition $\tilde{A}$ in the same manner. Then the $q_1 \times q_1$ upper-left corner of $\tilde{B}^{-1} \tilde{A} \tilde{B}^{-T}$ is $\tilde{B}_{11}^{-1} \tilde{A}_{11} \tilde{B}_{11}^{-T}$.

Proof: This is a direct calculation using

$$\tilde{B}^{-1} = \begin{bmatrix} \tilde{B}_{11}^{-1} & 0 \\ -\tilde{B}_{22}^{-1} \tilde{B}_{21} \tilde{B}_{11}^{-1} & \tilde{B}_{22}^{-1} \end{bmatrix}. \quad \square$$

If the parameter θ is partitioned as $\theta = (\theta_1, \theta_2)$ and is estimated by an M-estimator, then by the standard asymptotic theory of M-estimation the asymptotic variance of $\hat{\theta}$ is $\tilde{B}^{-1} \tilde{A} \tilde{B}^{-T}$ for certain matrices $\tilde{B}$ and $\tilde{A}$ (Huber 1967). Lemma 1 tells us when the asymptotic variance of $\hat{\theta}_1$ is the same as it would be if the "nuisance parameter" θ_2 were known and did not need to be estimated; such knowledge, of course, can greatly simplify theoretical studies as well as the calculation of standard errors in practice.

If θ_1 is the nuisance parameter, then the condition that $\tilde{B}_{21} = 0$ is sufficient for $\hat{\theta}_2$ to have the same asymptotic variance as if θ_1 were known. If $\tilde{B}_{12} = \tilde{B}_{21} = 0$, then each of θ_1 and θ_2 can be estimated as well as if the other were known. It does *not* follow that $\hat{\theta}_1$ and $\hat{\theta}_2$ are asymptotically uncorrelated; this requires that $\tilde{A}_{12} = \tilde{A}_{21}^T = 0$. For maximum likelihood estimation $\tilde{B} = \tilde{B}^T = \tilde{A}$, and it is well-known that the following are equivalent: (a) The asymptotic variance of $\hat{\theta}_1$ is the same as when θ_2 is known, (b) The asymptotic variance of $\hat{\theta}_2$ is the same as when θ_1 is known, and (c) $\hat{\theta}_1$ and $\hat{\theta}_2$ are asymptotically uncorrelated.

Proof of Theorem 1: Let p be the dimension of β . Partition $\tilde{B}_s$ and $\tilde{A}_s$ as in lemma 1 with $q_1 = p + 2$ and $q_2 = 1$. Thus, $\tilde{B}_{s,12}$ is

$$E \frac{\partial}{\partial \sigma} [\psi(r^c) \eta(r^c) \eta(r^c) f_{\beta}^{(\lambda)}(\underline{x}, \beta) \omega(\theta, \underline{x})]^T,$$

and $B_{s,21}$ is

$$E(\partial/\partial \lambda \quad \partial/\partial \mu \quad \partial/\partial \lambda)^T [\chi(r^c)].$$

Then by lemma 1 we need only show that

$$B_{s,12} = 0.$$

This is easy to prove since $E(\dot{\psi}(\epsilon)\epsilon) = E(\dot{\eta}(\epsilon)\epsilon) = 0$ because $\dot{\psi}$ and $\dot{\eta}$ are even and ϵ is symmetrically distributed. In general, $B_{s,21} \neq 0$ and $A_{12} \neq 0$ so $\hat{\sigma}$ and $(\hat{\lambda}, \hat{\beta}, \hat{\mu})$ are asymptotically correlated and the asymptotic variance of $\hat{\sigma}$ is different than it would be if (λ, β, μ) were known. This proves (a).

To simplify notation let $\mathcal{C} = A_{s,11}$ and $\mathcal{D} = B_{s,11}$. By part (a), the asymptotic variance of $(\hat{\lambda}, \hat{\beta}, \hat{\mu})$ is $\mathcal{D}^{-1} \mathcal{C} \mathcal{D}^{-T}$. Partition $\mathcal{C}$ and $\mathcal{D}$ as in lemma 1 with $q_1 = 2$ and $q_2 = p$. To prove (b) by lemma 1 it suffices to prove that $\mathcal{D}_{12} = 0$. Since

$$\sum_{i=1}^n \left[f^{(\lambda)}(\underline{x}_i, \beta) - \bar{f}^{(\lambda)}(\beta) \right] \equiv 0,$$

it follows that

$$(2.7) \quad -\partial/\partial \beta \sum_{i=1}^n E \psi(r_i^c(\theta)) =$$

$$E \left[\dot{\psi}(r_1(\theta)) \sum_{i=1}^n \partial/\partial \beta [f^{(\lambda)}(\underline{x}_i, \beta) - \bar{f}^{(\lambda)}(\beta)]/\sigma \right] = 0,$$

and for the same reason

$$(2.8) \quad \partial/\partial\beta \sum_{i=1}^n E(\eta(r_i^c(\beta))) = 0.$$

(2.7) and (2.8) prove that $\mathbb{D}_{12} = 0$.

To prove (c), note that in general $\mathbb{D}_{21} \neq 0$ and $\mathbb{C}_{12} = \mathbb{C}_{21}^T \neq 0$ so that $(\hat{\lambda}, \hat{\mu})$ and $\hat{\beta}$ are asymptotically correlated and the asymptotic variance of $\hat{\beta}$ depends on whether (λ, μ) is estimated or not. $\square$

Proof of Theorem 2: (a) Let

$$b_i^c(\beta) = b[f(x_i, \beta)] - \bar{b}(\beta).$$

Applying lemma 1 to B_h and A_h with $q_1 = 2$ and $q_2 = p$, it suffices to prove that $\mathbb{B}_{h,12} = 0$, where

$$\mathbb{B}_{h,12} = \begin{bmatrix} E \partial/\partial\beta \sum_{i=1}^n b_i^c(\beta) \chi(r_i(\Delta)) \\ E \partial/\partial\beta \sum_{i=1}^n \chi(r_i(\Delta)) \end{bmatrix}.$$

Now, $E\chi[r_i(\Delta)] = E\chi(\epsilon/\sigma)$ and $E\dot{\chi}[r_i(\Delta)] = E\dot{\chi}(\epsilon/\sigma) = 0$ for all i . Moreover,

$$\sum_{i=1}^n b_i^c(\beta) \equiv 0.$$

Therefore,

$$E \partial/\partial \beta \sum_{i=1}^n b_i^c(\beta) \chi(r_i(\Delta)) =$$

$$E \chi[r_i(\Delta)] (\partial/\partial \beta) \sum_{i=1}^n b_i^c(\beta) + E \dot{\chi}(r_i(\Delta)) \sum_{i=1}^n b_i^c(\beta) [-f_{\beta}^{(\lambda)}(x_i, \beta)/\sigma] = 0.$$

Also,

$$E \partial/\partial \beta \sum_{i=1}^n \chi(r_i(\Delta)) = \sum_{i=1}^n \dot{\chi}(r_i(\Delta)) [-f_{\beta}^{(\lambda)}(x_i, \beta)/\sigma] = 0.$$

(b) Let $\mathcal{D} = \mathcal{B}_{h,11}$ and $\mathcal{C} = \mathcal{A}_{h,11}$ where $\mathcal{B}$ and $\mathcal{A}$ are partitioned as in the proof of (a).

Then partition $\mathcal{D}$ and $\mathcal{C}$ as in lemma 1 with $q_1 = q_2 = 1$. Then the proof follows from lemma 1 since

$$D_{12} = \partial/\partial \sigma \sum_{i=1}^n b_i^c(\beta) \chi(r_i(\Delta)) =$$

$$\dot{\chi}(r_i(\Delta)) [-r_i(\Delta)/\sigma] \sum_{i=1}^n b_i^c(\beta) = 0. \quad \square$$

Proof of Theorem 3: (a) Partition $\mathcal{B}_h$ and $\mathcal{A}_h$ as in lemma 1 with $q_1 = 2$ and $q_2 = p$. We now regard (λ, σ) as the nuisance parameter so we need to show that $\mathcal{B}_{h,21} \rightarrow 0$ as $\sigma \rightarrow 0$; see the remarks after lemma 1. Now $\mathcal{B}_{h,21} = [\mathcal{B}_{h,21}^{(1)} \quad \mathcal{B}_{h,21}^{(2)}]$ where

$$\mathcal{B}_{h,21}^{(1)} = \partial/\partial \lambda \sum_{i=1}^n \eta(r_i(\Delta)) f_{\beta}^{(\lambda)}(x_i, \beta) w(\Delta, x_i)$$

and

$$B_{h,21}^{(2)} = \partial/\partial\sigma \sum_{i=1}^n \eta(r_i(\Delta)) f_{\beta}^{(\lambda)}(\mathbf{x}_i, \beta) w(\Delta, \mathbf{x}_i).$$

Since $E[r_i(\Delta) \dot{\eta}(r_i(\Delta))] = 0$, $B_{h,21}^{(2)} = 0$. As in the proof of (3.8), as $\sigma \rightarrow 0$

$$E \partial/\partial\lambda \eta(r_i(\Delta)) = E \dot{\eta}(\epsilon/\sigma) (\partial r_i(\Delta)/\partial\lambda) \cong$$

$$E [\dot{\eta}(\epsilon/\sigma) (\epsilon/\sigma)] \log(f(\mathbf{x}_i, \beta)) = 0,$$

whence $B_{h,21}^{(1)} \cong 0$.

(b,c) From theorem 2(b), we know that the asymptotic variance of $\hat{\lambda}$ is $A_{\lambda}/B_{\lambda}^2$ where

$$(3.7) \quad A_{\lambda} = E \chi^2(\epsilon/\sigma) \sum_{i=1}^n \{b_i^c(\beta)\}^2$$

and

$$B_{\lambda} = \sum_{i=1}^n b_i^c(\beta) E\{\dot{\chi}(\epsilon/\sigma) (\partial r_i(\Delta)/\partial\lambda)\}.$$

Now letting $y = y_i$ and $f = f(\mathbf{x}_i, \beta)$,

$$\partial r_i(\Delta)/\partial\lambda = \sigma^{-1} \{ \lambda^{-2} [\log(y^{\lambda}) y^{\lambda} - \log(f^{\lambda}) f^{\lambda}] - \lambda^{-1} [y^{(\lambda)} - f^{(\lambda)}] \}.$$

As $\sigma \rightarrow 0$, $y \rightarrow f$. Therefore, since $d(x \log(x))/dx = 1 + \log(x)$, as $\sigma \rightarrow 0$

$$\log(y^{\lambda}) y^{\lambda} - \log(f^{\lambda}) f^{\lambda} \cong 1 + \log(f^{\lambda}) [y^{\lambda} - f^{\lambda}]$$

so

$$(3.8) \quad \partial \mathbf{r}_1(\hat{\Delta}) / \partial \lambda \cong \sigma^{-1} \{ \lambda^{-2} [1 + \log(f^\lambda)] [y^\lambda - f^\lambda] - \lambda^{-2} [y^\lambda - f^\lambda] \} =$$

$$\sigma^{-1} \log(f) [y^\lambda - f^\lambda] / \lambda \cong \log(f) (\epsilon / \sigma).$$

Therefore, as $\sigma \rightarrow 0$

$$(3.9) \quad \mathbf{B}_{\lambda} \cong E [(\epsilon / \sigma) \dot{\chi}(\epsilon / \sigma)] \sum_{i=1}^n \log[f(\mathbf{x}_i, \beta)] \mathbf{b}_i^c(\beta).$$

(3.7) and (3.9) prove (3.5). (3.6) follows by applying the Cauchy–Schwarz inequality to the denominator of (3.5). $\square$

As $\sigma \rightarrow 0$ certain elements of $\mathbf{B}_h$ converge to ∞ at rate $O(\sigma^{-1})$. This does not affect our argument. One can stabilize $\mathbf{B}_h$ through the post–multiplication of $\mathbf{B}_h$ by $\text{Diag}(1, \sigma, \dots, \sigma)$. This gives us the asymptotic variance of $(\hat{\lambda}, \hat{\sigma}/\sigma, \hat{\beta}/\sigma)$ which has a finite, non–zero limit as $\sigma \rightarrow 0$. It is intuitively reasonable that the variance of $\hat{\lambda}$ does not converge to 0 as $\sigma \rightarrow 0$; if $\sigma = 0$ then λ is not identifiable.

Proof of Theorem 4: The log–likelihood is

$$L(\hat{\Delta}) = -n/2 \log(2\pi\sigma^2) - 1/2 \sum_{i=1}^n r_i^2(\hat{\Delta}) + (\lambda - 1) \sum_{i=1}^n \log(y_i).$$

Differentiating $L(\hat{\Delta})$ with respect to β , σ , and the λ , we see that $\hat{\Delta}_{\text{ML}}$ solves

$$(3.10) \quad \sum_{i=1}^n r_i(\Delta) f_{\beta}^{(\lambda)}(\mathbf{x}_i, \beta) = 0,$$

$$(3.11) \quad \sum_{i=1}^n [r_i^2(\Delta) - 1] = 0,$$

and

$$(3.12) \quad \sum_{i=1}^n r_i(\Delta) (\partial r_i(\Delta) / \partial \lambda) - \log(y_i) = 0.$$

We now look at (3.12) as $\sigma \rightarrow 0$. By (3.8), $\partial r_i(\Delta) / \partial \lambda \cong \log(y_i) (\epsilon / \sigma) \cong \log[f(\mathbf{x}_i, \beta)] (\epsilon / \sigma)$, so as $\sigma \rightarrow 0$ (3.12) converges to

$$(3.12) \quad \sum_{i=1}^n [r_i^2(\Delta) - 1] \log[f(\mathbf{x}_i, \beta)].$$

Comparing (3.10), (3.11), and (3.13) to (3.1)–(3.2), we see that the likelihood equations converge to (3.2) as $\sigma \rightarrow 0$. $\square$

Proof of theorem 5: This is another application of lemma 1. The key result is that

$$\begin{aligned} & E \left[\partial / \partial \zeta \{ w(\zeta) \Psi_s(r_i(\zeta), r_i^c(\zeta), \mathbf{x}_i, \zeta) + [1 - w(\zeta)] \Psi_h(r_i(\zeta), r_i^c(\zeta), \mathbf{x}_i, \zeta) \} \right] \\ &= E \left[\Psi_s(r_i(\zeta), r_i^c(\zeta), \mathbf{x}_i, \zeta) - \Psi_h(r_i(\zeta), r_i^c(\zeta), \mathbf{x}_i, \zeta) \right] \partial w(\zeta) / \partial \zeta = 0, \end{aligned}$$

since $E \Psi_s(r_i(\zeta), r_i^c(\zeta), \mathbf{x}_i, \zeta) = E \Psi_h(r_i(\zeta), r_i^c(\zeta), \mathbf{x}_i, \zeta) = 0$. $\square$

The proof of Theorem 5 could be extended to a more general result about combining two or more unbiased estimating equations estimating the same parameter, but we will not do this here. Notice however that the same proof applies to $\hat{\Delta}_U$, the heteroscedasticity–skewness estimator without the centering constant μ .

REFERENCES

- Anscombe, F. J. (1967), "Examination of Residuals," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, ed. J. Neyman, Berkeley: University of California Press, pp. 1–36.
- Bickel, P. J. (1978), "Using Residuals Robustly I: Tests for Heteroscedasticity, Nonlinearity," *The Annals of Statistics*, 6, 266–291.
- Bickel, P. J. and Doksum, K. A. (1981), "An Analysis of Transformations Revisited," *Journal of the American Statistical Association*, 76, 296–311.
- Box, G. E. P. and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, Series B*, 26, 211–246.
- Carroll, R. J. (1979), "On estimating variances of robust estimators when the errors are asymmetric," *Journal of the American Statistical Association*, 74, 673–679.
- Carroll, R. J. and Ruppert, D. (1981), "On Prediction and the Power Transformation Family," *Biometrika*, 68, 609–615.
- Carroll, R. J. and Ruppert, D. (1982), "Robust Estimation in Heteroscedastic Linear Models," *The Annals of Statistics*, 10, 429–441.
- Carroll, R. J. and Ruppert, D. (1984), "Power Transformation When Fitting Theoretical Models to Data," *Journal of the American Statistical Association*, 79, 321–328.
- Carroll, R. J. and Ruppert, D. (1987), "Diagnostics and Robust Estimation When Transforming the Regression Model and the Response," *Technometrics*, 3, 287–300.
- Carroll, R. J. and Ruppert, D. (1988), *Transformation and Weighting in Regression*. New York and London: Chapman and Hall.
- Carroll, R. J., Cressie, N. A. C., and Ruppert, D. (1987), "A Transformation/Weighting Model for Estimating Michaelis–Menten Parameters," (under revision for *Biometrics*).
- Cook, R. D. and Weisberg, S. (1983), "Diagnostic for heteroscedasticity in regression," *Biometrika*, 72, 23–29.
- Hernandez, F. and Johnson, R. A. (1980), "The Large Sample Behavior of Transformations to Normality," *Journal of the American Statistical Association*, 75, 855–861.

- Hinkley, D. V. (1975), "On Power Transformations to Symmetry," *Biometrika*, 62, 101–111.
- Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73–101.
- Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1*. eds. L.M. LeCam and J. Neyman, Berkeley and Los Angeles: University of California Press, pp. 221–233.
- Huber, P. J. (1981), *Robust Statistics*. New York: John Wiley.
- Li, G., (1985), "Robust Regression," in *Exploring Data Tables, Trends, and Shapes*, eds. D. C. Hoaglin, F. Mosteller, and J. W. Tukey. New York: John Wiley, pp. 281–344.
- Mallows, C. L. (1975), "Robust Methods," *Proceedings of Symposia in Applied Mathematics*, 28, ed. R. Gnanadesikan. Providence, R.I.: American Mathematical Society, pp. 49–74.
- Ricker, W. E. (1954), "Stock and Recruitment," *Journal of the Fisheries Research Board of Canada*, 11, 559–623.
- Rudemo, M., Ruppert, D, and Streibig, J. (1987), "Random Effect Models in Nonlinear Regression with Applications to Bioassay," Mimeo Series #1727, Department of Statistics, University of North Carolina at Chapel Hill.
- Ruppert, D. and Carroll, R. J. (1985), "Data Transformations in Regression Analysis with Applications to Stock Recruitment Relationships," In *Resource Management: Lecture Notes in Biomathematics*, 61, ed. M. Mangel. New York: Springer Verlag, pp. 29–47.
- Snee, R. D. (1986), "An Alternative Approach to Fitting Models When Reexpression of the Response is Useful," *Journal of Quality Technology*, 18, 211–225.
- Switzer, P. (1972), "Efficiency robustness of estimators," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, eds. L. M. LeCam, J. Neyman, and E. L. Scott. Berkeley and Los Angeles: University of California Press, pp. 283–291.
- Taylor, J. M. G. (1985), "Power transformations to Symmetry," *Biometrika*, 72, 145–152.
- Wu, C. F. J. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis (With Discussion)," *The Annals of Statistics*, 14, 1261–1343.

*Table 1. Asymptotic variance of $\hat{\lambda}_s$ when $\log y = \mu + \epsilon$
and $\epsilon \sim F = (1-\gamma)N(0,1) + \gamma N(0,9)$.*

b	$\gamma = 0$ (Normal)	$\gamma = .001$	$\gamma = .01$	$\gamma = .1$	$\gamma = .2$	$\gamma = .4$	Maximum over γ
1	5.713	5.717	5.371	4.688	4.135	3.228	5.717
2	1.110	1.067	1.101	1.313	1.551	0.938	1.551
3	0.711	0.718	0.850	1.709	1.044	0.306	1.709
4	0.669	0.734	1.251	1.346	0.567	0.193	1.346
5	0.667	0.844	1.819	0.984	0.407	0.175	1.819
8	0.667	1.253	2.457	0.692	0.355	0.172	2.457
12	0.667	1.393	2.493	0.683	0.355	0.172	2.493

Table 2. Bootstrap results for Salmon data.

	$\hat{\beta}_{hs,1}$	$\hat{\beta}_{hs,2}$	$\hat{\lambda}_{hs}$	$\hat{w}$
Original Estimates = values generating bootstrap data	3.8	-9.7×10^{-4}	-.26	.85
<u>n = 27</u>				
Bootstrap expectation	3.8	-9.3×10^{-4}	-.29	.72
Std. dev. from $\hat{V}_{hs}(\hat{w})$	.61	3.5×10^{-4}	.28	—
Std. dev. from NL regression	.67	3.0×10^{-4}	—	—
Bootstrap Std. dev.	.81	3.3×10^{-4}	.45	.22
<u>n = 81</u>				
Bootstrap expectation	3.8	-9.6×10^{-4}	-.22	.81
Std. dev. from $\hat{V}_{hs}(\hat{w})^a$	.35	2.0×10^{-4}	.17	—
Std. dev. from NL regression ^a	.39	1.7×10^{-4}	—	—
Bootstrap std. dev.	.41	1.8×10^{-4}	.17	.083

^aObtained by dividing the entries for n = 27 by $\sqrt{3}$.