#1760

Transformations to Symmetry and Homoscedasticity

David Ruppert[1] and Brian Aldershof[2]

[1] David Ruppert is a professor of Operations Research and Industrial Engineering at Cornell University.

[2] Brian Aldershof is a graduate student in the Department of Statistics at the University of North Carolina.

This research was supported by NSF Grants DMS–8400602 and DMS–8701201.

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 varibles 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 an example and discuss computation of the estimators.

Key Words Transform–both–sides, skewness, tests for heteroscedasticity, M–estimation.

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, 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 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

$$(1.1) \qquad h(y,\lambda) = y^{(\lambda)} = (y^{\lambda} - 1)/\lambda \qquad \text{if } \lambda \neq 0,$$

$$= \log(y) \qquad \text{if } \lambda \mp 0.$$

The Box and Cox assume that for some λ

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

where ϵ has a standard normal distribution and $x^{T}\beta$ is a simple linear model. They estimate λ, β, 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 $x^{T}\beta$) for the relationship between the covariates x and the response y.

A difficulty with model (1.1)–(1.2) is that since y is positive, ϵ cannot have a distribution whose support is $(-\infty, \infty)$ unless λ is zero. Model (1.1) – (1.2) can be inverted to

$$y = \{1 + \lambda(x^T\beta + \sigma\epsilon)\}^{1/\lambda}. \tag{1.3}$$

Since (1.3) will produce a positive y if $1/\lambda$ is an even integer, it is sometimes claimed that the Box–Cox model will hold with normal ϵ if $1/\lambda$ is an even integer. However, (1.1)–(1.2) cannot be inverted to (1.3) unless $\{1 + \lambda(x^T\beta + \sigma\epsilon)\}$ is non–negative with probability one, so the Box–Cox model with normal ϵ really requires that λ be zero.

It appears preferable to assume that ϵ is not normally distributed but rather has a symmetric distribution with finite support (and perhaps close to normal in some sense). One could use a parametric family of such distributions such as 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 where $x^T\beta = \mu$, a location parameter. 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, x_i), let $\hat{\beta}(\lambda)$ be the least–squares estimate of β when fitting (1.1)–(1.2) with λ fixed, and let

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

be the ith 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 the $\{\hat{\epsilon}_i(\lambda)\}$ is zero.

The advantage of the Hinkley and Taylor estimators is that when a transformation to symmetry does exist, that is, (1.2) holds for some λ with ϵ symmetrically distributed, then these estimators will consistently estimate the correct λ. Except when $\lambda = 0$, the Box–Cox 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 relationship between $y^{(\lambda)}$ and x or from the mean–variance relationship of $y^{(\lambda)}$.

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 x, say that the median of y given x is

$$f(x,\beta),$$

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. $y = f(x,\beta) + \text{error}$. In this situation, Carroll and Ruppert (1984) propose "transforming both sides" of the regression equation, that is, using the model

$$y_i^{(\lambda)} = f^{(\lambda)}(x_i,\beta) + \epsilon_i, \tag{1.4}$$

where $\epsilon_1,\ldots,\epsilon_n$ are i.i.d. F. Model (1.4) 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 (1.4) implies that the median of y given x is $f(x,\beta)$, regardless of the value of λ. Thus, the value of λ is not determined by the conditional median of y given x. Rather, it is the asymmetry and the mean–variance relationship of y that determines λ. If (1.4) 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,

$$\text{(1.5)} \qquad \mathrm{var}(y^{(\lambda)}) \cong \sigma_\epsilon^2 \, \{f(x,\beta)\}^{2(\lambda - \lambda_0)}.$$

In this paper we apply the transform–to–symmetry estimators of Hinkley and Taylor to the transform–both–sides model, (1.4). 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) $\sigma^2(f(x_i,\beta)) \to 0$, where $\sigma^2(f(x_i,\beta))$ is the sample variance of $\{f(x_i,\beta): i = 1,...,n\}$, and (2) $\sigma \to 0$, 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 the their distribution. In case (1) 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) the transformation $y_i \to y_i^{(\lambda)}$ is nearly linear for fixed x_i; therefore $y_i^{(\lambda)}$ and y_i have the same distribution except for a location and scale change (the scale change depends on x_i), 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.

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 $y_1,...,y_n$ 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 $y_1,...,y_n$ as

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

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 transform–both–sides model as follows. For fixed λ, let $\hat{\epsilon}_i(\lambda) = y_i^{(\lambda)} - f^{(\lambda)}(x_i,\hat{\beta}(\lambda))$ 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

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

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

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

If λ equals λ_0, the true value of this parameter, then $\hat{\mu}(\lambda)$ will converge to zero as $n \to \infty$, 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

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

Let

$$r_i(\theta) = [y_i^{(\lambda)} - f^{(\lambda)}(x_i,\beta)]/\sigma \tag{2.2}$$

and

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

The superscript "c" is to indicate that the residual has been centered. Having μ in r_i^c allows us to center $f^{(\lambda)}(x_i,\beta)$ 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

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

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

$$(2.4)\qquad \sum_{i=1}^{n} \Psi_s(r_i(\hat{\theta}),r_i^c(\hat{\theta}),x_i,\hat{\theta}) = 0.$$

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 $\omega(\theta,x)$ be a weight function taking values in [0, 1]; for robustness

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

should be bounded as a function of θ and x. Then replace (2.3) by

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

and let $\hat{\theta}$ 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 $\hat{\theta}$ can be investigated using Huber's (1967) results on M–estimation or related techniques. Here 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

$$A_s = E \sum_{i=1}^{n} \Psi_s(r_i(\theta_0),r_i^c(\theta_0),x_i,\theta_0)\ \Psi_s^T(r_i(\theta_0),r_i^c(\theta_0),x_i,\theta_0) \tag{2.5}$$

and

$$B_s = -\,(\partial/\partial\theta)\ E \sum_{i=1}^{n} \Psi_s(r_i(\theta_0),r_i^c(\theta_0).x_i,\theta_0). \tag{2.6}$$

V_s can be consistently estimated by replacing θ_0 with $\hat{\theta}$ and replacing the expected sums in (2.5) and (2.6) 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 (1.4) holds with ϵ symmetrically distributed and that $\hat{\theta}$ is asymptotically normal with variance matrix V_s, or more precisely,

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

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*) T*he 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) where 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. TRANSFORMATION TO HOMOSCEDASTICITY

As in section 2, we assume that model (1.4) 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 by (1.5) 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 to test if $\lambda = \lambda_0$, and estimating λ by the value giving zero correlation. 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 $r_i(\theta)$ defined by (2.2) does not depend on μ, we now write $r_i(\Delta)$ instead of $r_i(\theta)$. Now we also assume that ω does not depend on μ and write $\omega(\Delta,x)$.

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

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

and

$$S(\Delta) = \sum_{i=1}^{n} \{b[f(x_i,\beta)] - \bar{b}(\beta)\} \; \chi[r_i(\Delta)].$$

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

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

Then $\hat{\Delta}$ solves

$$(3.2)\qquad \sum_{i=1}^{n} \Psi_h(r_i(\hat{\Delta}),x_i,\hat{\Delta}) = 0.$$

If (1.5) holds exactly then

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

as $\lambda \to \lambda_0$, and so the score test for heteroscedasticity uses $b(x) = \log(x)$; 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 $V_h = B_h^{-1} A_h B_h^{-T}$ where

$$(3.3)\qquad B_h = \partial/\partial\Delta\ E \sum_{i=1}^{n} \Psi_h(r_i(\Delta),x_i,\Delta),$$

and

$$(3.4) \qquad A_h = E \sum_{i=1}^{n} \Psi_h(r_i(\Delta),x_i,\Delta)\, \Psi_h^T(r_i(\Delta),x_i,\Delta).$$

Even when F is asymmetric, $\hat{\Delta}$ is consistent and asymptotically normal with variance V_h. However, the form of 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* β *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 anu 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*

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

and

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

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

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

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

The 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).

Let $\hat{\Delta}_{ML} = (\hat{\lambda}_{ML}, \hat{\sigma}_{ML}, \hat{\beta}_{ML})$ 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 $\sigma \to 0$. 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 $\sigma \to 0$, $\hat{\beta}_{\mathrm{ML}}$ has the same asymptotic variance as when λ is known, and theorem *3(a)* shows that $\hat{\Delta}$ has the same behavior. In fact, more is true:

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

The approximation (1.5) becomes exact as the scale parameter $\sigma \to 0$. For this reason, we could conjecture that as $\sigma \to 0$, 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 (1.4) 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 $\hat{\theta}_{\mathrm{s}} = (\hat{\lambda}_{\mathrm{s}}, \hat{\mu}_{\mathrm{s}}, \hat{\beta}_{\mathrm{s}}^{\mathrm{T}}, \hat{\sigma}_{\mathrm{s}})^{\mathrm{T}}$ be the transformation to symmetry estimator proposed in section 2, and let $\hat{\Delta}_{\mathrm{h}} = (\hat{\lambda}_{\mathrm{h}}, \hat{\sigma}_{\mathrm{h}}, \hat{\beta}_{\mathrm{h}}^{\mathrm{T}})^{\mathrm{T}}$ 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 θ_s be the limit of $\hat{\theta}_s$ and let Δ_h be the limit of $\hat{\Delta}_h$ as $n \to \infty$ (we assume that these limits exist). If $\lambda_s = \lambda_h$, then $\beta_s = \beta_h$ and $\sigma_s = \sigma_h$, but in general $\beta_s \neq \beta_h$ and $\sigma_s \neq \sigma_h$ if $\lambda_s \neq \lambda_h$.

To test

$$H_0: \lambda_s = \lambda_h \text{ versus } H_1: \lambda_s \neq \lambda_h, \tag{4.1}$$

we need the joint limiting distribution of $\hat{\lambda}_s$ and $\hat{\lambda}_h$. Let $\hat{\theta}_J = (\hat{\theta}_s^T, \hat{\Delta}_h^T)^T$ be the joint estimator of θ_s and Δ_h. Then $\hat{\theta}_J$ is an M–estimator solving (2.4) and (3.2) simultaneously. If we define $\Psi_J(\hat{\theta}_J,x_i)$ by stacking $\Psi_s(r_i(\hat{\theta}_s),r_i^c(\hat{\theta}_s),x_i,\hat{\theta}_s)$ and $\Psi_h(r_i(\hat{\Delta}_h),x_i,\hat{\Delta}_h)$, then the asymptotic variance of $\hat{\theta}_J$ is estimated by

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

where

$$\hat{B}_J = \partial/\partial\theta \sum_{i=1}^{n} \Psi_J(\hat{\theta}_J,x_i)$$

and

$$\hat{A}_J = \sum_{i=1}^{n} \Psi_J(\hat{\theta}_J,x_i)\, \Psi_J^T(\hat{\theta}_J,x_i).$$

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

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

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. 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

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

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 Δ.

Another possibility, one that we use in the example 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 θ by Δ 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

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

If w = 1, then $\hat{\theta}_{hs} = \hat{\theta}_s$. If w = 0, then $\hat{\theta}_{hs}$ equals Δ_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

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

Then

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

where

$$\hat{B}_{hs}(w) = w\hat{B}_s + (1-w)\ \hat{B}_h$$

and

$$\hat{A}_{hs}(w) = w^2\hat{A}_s + (1-w)^2\hat{A}_h + w(1-w)(\hat{A}_{s\cdot h} + \hat{A}_{s\cdot h}^T).$$

Define $\hat{C} = (\hat{A}_s, \hat{A}_h, \hat{A}_{s\cdot h}, \hat{B}_s, \hat{B}_h)$ and let $\hat{w}(\hat{C})$ be the value of w that minimizes $\hat{\sigma}^2(\hat{\lambda})$ of $\hat{V}_{hs}(w)$.

$\hat{C}$ and $\hat{\theta}_{hs}$ are both M–estimates, defined jointly by (2.5), (2.6), (3.3), (3.4), (4.4), and (4.3), where $w = \hat{w}(\hat{c})$ in (4.3). When estimating the asymptotic variance of $\hat{\theta}_{hs}$ it would be burdensome if we needed to examine the joint limiting distribution of $\hat{\theta}_{hs}$ and $\hat{C}$. Fortunately, this is not the case. The following theorem shows that for large enough sample sizes we can act as if $\hat{w}$ were fixed and use $\hat{V}_{hs}(\hat{w})$ to estimate the variance of $\hat{\theta}_{hs}$.

Theorem 5: The asymptotic variance of $\hat{\theta}_{hs}$ is the same as it would be if C where known.

5. AN EXAMPLE

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 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).

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 (1.3) using $(\lambda, \beta) = (\hat{\lambda}_{hs}, \hat{\beta}_{hs})$. The errors, ϵ, were generated by sampling with replacement from the "symmetrized residuals", $\{\hat{\epsilon}_1,\ldots\hat{\epsilon}_n,-\hat{\epsilon}_1,\ldots,-\hat{\epsilon}_n\}$, 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. 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 $f^{(\lambda)}(x_i,\beta)$, with $\lambda = \hat{\lambda}_{hs}$ treated as fixed. The estimated variance matrix is the

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

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 $\hat{\lambda}_{hs}$ from $\hat{V}_{hs}(\hat{w})$ agrees to two decimals with the bootstrap standard error. For samples of this size, treating $\hat{w}$ as fixed seems to be an adequate approximation. For $\hat{\beta}_{hs}$, the nonlinear least–squares standard errors are closer to the bootstrap standard errors than the standard errors from $\hat{V}_{hs}(\hat{w})$. 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 $\hat{V}_{hs}(\hat{w})$ is related to the jackknife estimate of variance that does not assume homoscedasticity; see Wu (1986).

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 $\hat{\beta}$, $\hat{\lambda}_{hs}$, 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 $\hat{\lambda}_{hs}$ was then found from $\hat{V}_{hs}(w)$.

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 λ until sufficient accuracy was obtained.

6. PROOFS

We will make repeated use of the following result.

Lemma 1: *Let A and B be* $q \times q$ *matrices such that*

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

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

Proof: This is a direct calculation using

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

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 $B^{-1}AB^{-T}$ for certain matrices B and 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 where 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 $B_{21} = 0$ is sufficient for $\hat{\theta}_2$ to have the same asymptotic variance as if θ_1 were known. If $B_{12} = B_{21} = 0$, 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 $A_{12} = A_{21}^T = 0$. For maximum likelihood estimation $B = B^T = A$, 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 B_s and A_s as in lemma 1 with $q_1 = p + 2$ and $q_2 = 1$. 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 $C = A_{s,11}$ and $D = B_{s,11}$. By part (a), the asymptotic variance of $(\hat{\lambda}, \hat{\beta}, \hat{\mu})$ is $D^{-1}CD^{-T}$. Partition C and D as in lemma 1 with $q_1 = 2$ and $q_2 = p$. To prove (b) by lemma 1 it suffices to prove that $D_{12} = 0$. Since

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

it follows that

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

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

and for the same reason

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

(2.7) and (2.8) prove that $D_{12} = 0$.

To prove (c), note that in general $D_{21} \neq 0$ and $C_{12} = C_{21}^T \neq 0$ 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. □

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 $B_{h,12} = 0$, where

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

Now,

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

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

since

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

and $E\dot{\chi}(\epsilon/\sigma) = 0$. 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 $D = B_{h,11}$ and $C = A_{h,11}$ where B and A are partitioned as in the proof of (a). Then partition D and C as in lemma 1 with $q_1 = q_2 = 1$. Then the proof follows from lemma 1 since

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

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

Proof of Theorem 3: (*a*) Partition B_h and 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 $B_{h,21} \to 0$ as $\sigma \to 0$; see the remarks after lemma 1. Now $B_{h,21} = [B_{h,21}^{(1)}\ \ B_{h,21}^{(2)}]$ where

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

and

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

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

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

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

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

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

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

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 $f = f(x_i,\beta)$,

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

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

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

so

$$(3.8) \qquad \partial r_i(\Delta)/\partial\lambda \cong \sigma^{-1}\{\lambda^{-2}[1 + \log(f^\lambda)][y^\lambda - f^\lambda] - \lambda^{-2}[y^\lambda - f^\lambda]\} =$$

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

Therefore, as $\sigma \to 0$

$$(3.9) \qquad B_\lambda \cong E\ [(\epsilon/\sigma)\ \dot{\chi}(\epsilon/\sigma)] \sum_{i=1}^{n} \log[f(x_i,\beta)]\ b_i^c(\beta).$$

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

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

Proof of Theorem 4: The log–likelihood is

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

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

(3.10) $$\sum_{i=1}^{n} r_i(\Delta)\; f_\beta^{(\lambda)}(x_i,\beta) = 0,$$

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

and

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

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

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

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

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

$$E\left[\partial/\partial C \; \{w(C) \; \Psi_s(r_i(\theta),r_i^c(\theta),x_i,\theta) + [1-w(C)] \; \Psi_h(r_i(\theta),r_i^c(\theta),x_i,\theta)\}\right]$$

$$= E\left[\Psi_s(r_i(\theta),r_i^c(\theta),x_i,\theta) - \Psi_h(r_i(\theta),r_i^c(\theta),x_i,\theta)\right] \; \partial w(C)/\partial C = 0,$$

since $E \; \Psi_s(r_i(\theta),r_i^c(\theta),x_i,\theta) = \Psi_h(r_i(\theta),r_i^c(\theta),x_i,\theta) = 0$. □

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

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*, Chapman and Hall. (to appear).

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. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1*, 221–233.

Huber, P. J. (1981), *Robust Statistics*, John Wiley: New York.

Li, G., (1985), "Robust Regression," in *Exploring Data Tables, Trends, and Shapes*, Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors. John Wiley: New York.

Mallows, C. L. (1975), "On Some Topics in Robustness," unpublished manuscript.

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, M. Mangel, editor, New York: Springer Verlag.

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. (1970), "Efficiency robustness of estimators," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 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. Bootstrap expectations, asymptotic standard deviations, and bootstrap standard deviations.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\lambda}$	$\hat{w}$
Original Estimates = values generating bootstrap data	3.8	-9.7×10^{-5}	−.26	.85
n = 27				
Bootstrap expectation	3.8	-9.3×10^{-5}	−.29	.72
Std. dev. from $\hat{V}_{hs}(\hat{w})$	.61	3.5×10^{-5}	.28	——
Std. dev. from NL regression	.67	3.0×10^{-5}	——	——
Bootstrap Std. dev.	.81	3.3×10^{-5}	.45	.22
n = 81				
Bootstrap expectation	3.8	-9.6×10^{-5}	−.22	.81
Std. dev. from $\hat{V}_{hs}(\hat{w})^*$	.35	2.0×10^{-5}	.17	——
Std. dev. from NL regression*	.39	1.7×10^{-5}	——	——
Bootstrap std. dev.	.41	1.8×10^{-5}	.17	.083

*Obtained by dividing the entries for n = 27 by $\sqrt{3}$.