NO. 2004

TRANSFORMATIONS TO SYMMETRY IN THE TRANSFORM-BOTH-SIDES REGRESSION MODEL

by

Miguel Nakamura

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics

Chapel Hill

1989

Approved by:

David Ruppert
Advisor

E. Carlstein
Reader

Gordon Simons
Reader

MIGUEL NAKAMURA. Transformations to Symmetry in the Transform-Both-Sides Regression Model. (Under the direction of DAVID RUPPERT.)

ABSTRACT

The Transform-Both-Sides (TBS) regression model of Carroll and Ruppert (1984) is practical when there exists a conjectured relationship between the dependent variable and the independent variables, although the exact way in which randomness affects the relationship is unknown. The object is to make prediction more precise while respecting the pre-specified relationship. This dissertation is concerned with estimation methods which are alternate to the normal-theory Maximum Likelihood Estimate (MLE), and whose properties rely solely on symmetry of errors.

First, consistency and asymptotic normality of the skewness estimator proposed by Ruppert and Aldershof (1989) are proved. For consistency, an asymptotically equivalent estimator is constructed in order to attain convergence.

Secondly, a new estimator is proposed based on the notion of a weighted Minimum Distance (MD). This estimate is seen to be a special case of Generalized M-Estimation, a general class for which consistency and asymptotic normality are then proved. The MD estimate is shown to be consistent for Box-Cox modified power transformations, and modified power transformations with shift with the power fixed. Some examples are provided which tend to show that the original parameterization in a shifted power transformation is inappropriate.

Finally, different features are explored for MD in a simulation study. For a power transformation parameter and normal errors, a weight which appears suitable is produced. Comparisons of MD with other estimators in TBS are also provided; in particular, it is found that the loss of efficiency with respect to the MLE in the normal case is not extreme.

ACKNOWLEDGMENTS

First, I wish to express my sincere appreciation to my advisor, David Ruppert, for the guidance and patience which he displayed through all stages of work on this dissertation.

I thank E. Carlstein, I. M. Chakravarti, G. Simons, and W. L. Smith, the members of my committee, for reviewing the manuscript.

I am grateful to the staff of the Department of Statistics for providing a fruitful environment for learning and where help could always be found if needed, with special thanks to June Maxwell.

For the warm friendship and hospitality during my "year abroad", I am grateful to the '87-'88 dwellers of 357 Upson Hall in Cornell University: Andrew, Karen, Maika, and Mary.

I also wish to thank my fellow graduate students and friends in Chapel Hill for their fellowship and words of encouragement. To construct an exhaustive list of names for this would be a nearly impossible task; I duly thank all.

Likewise, I thank my family for their constant support throughout the years.

For financial support, I thank the Department of Statistics at UNC and the Consejo Nacional de Ciencia y Tecnologia (Mexico).

This work was partially supported by The National Science Foundation under grants DMS-8701201 and DMS-8902973.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.0 The Regression Setting

Broadly speaking, we may say that the object of regression analysis is to study relationships that may be present between a dependent variable Y and an independent variable X. Most of the time, we are restricted to only being able to record occurrences of X and Y which include unobservable random perturbations, due to random variation and measurement error. Regression methods focus on extracting characteristics of the underlying relationship which are masked by the observed randomness. These methods are means for obtaining a better understanding of the connections between X and Y and for making predictions on Y based on X. The accuracy of these procedures depends on a few assumptions which may be violated in given situations, in particular, ones that involve properties of the probability distributions of the perturbations. The present work deals with the application of transformations to data that arise in the regression setting when some key distributional properties of the perturbations are not verified. The ultimate purpose behind transformation is to increase the efficiency and precision of the inferences that are drawn.

When observing pairs of variables $(X_1,Y_1),\ldots,(X_n,Y_n)$, frequently a model of the following form is adopted. For $1 \leq i \leq n$ let

(1.1) $$Y_i = f(X_i, \beta) + \sigma\epsilon_i$$

where the following notation and assumptions are employed:

β is a p-dimensional vector of regression parameters,

$X_i = (X_{i1}, \ldots, X_{ik})^T$ is a k-dimensional random vector,

$\sigma > 0$ is a parameter to model the extent of variability,

ϵ_i for $1 \leq i \leq n$ are independent having a distribution with mean zero and variance one, and

$f : \mathbb{R}^{k+p} \longrightarrow \mathbb{R}$ is of a given (known) form.

As stated, the term $f(X_i, \beta)$ models the conditional expectation of Y_i given X_i, while the term $\sigma\epsilon_i$ can be interpreted as the unobserved random perturbation introduced by nature. When considering the above model, the implicit presumption is that this perturbation is additive. In the case that the probability distribution of the term ϵ_i is completely known, model (1.1) provides a description of the conditional distribution of Y_i given X_i.

When the primary objective is to estimate the parameter β, several estimation procedures can be applied assuming model (1.1). Least squares and M-estimation are examples of these procedures; these methods are generally efficient and/or consistent when the ϵ_i are (at least) symmetric with tails no heavier than the normal distribution and have constant variance (see Hampel *et al.* (1986), chapter 6). Jennrich (1969), Malinvaud (1970), and Wu (1981), discuss the properties of least squares estimators. When *heteroscedasticity* (that is, the lack of constant variance across the ϵ_i) is present, and therefore estimation procedures are deemed non-optimal, alternate techniques such as weighting or heteroscedastic regression may be applied (Carroll and

Ruppert (1988), chapters 1-3).

Transformation is another way that investigators have proposed to counteract the negative effect of lack of symmetry or lack of homoscedasticity in errors. Although transformations applied to random variables can be traced back to early work and in different settings (e.g. Bartlett (1947)), it is primarily regression models which we have in mind for this work. We can then speak of transforming either the response variable (Y) or the explanatory variable (X), or both, and hope to produce one of several favorable effects. Transformations of independent variables and transformations that convert an intrinsically linear model to explicit linearity in the parameters will not be of primary concern here. When the latter is done, some non-normality or heteroscedasticity can be induced in the errors, as shown by the examples in Carroll and Ruppert (1984) or Box and Hill (1974). Some of the work in transformations has focused on searching for the form of f either nonparametrically or parametrically through families of transformations (Box and Tidwell (1962)). In the present work, we are concerned with transformations that intentionally alter the distribution of the errors in the response.

We can see the significance of a general transformation by noting the following remark (Carroll and Ruppert (1988)): If $\varphi(y)$ is a strictly convex function, its effect on the distribution of a random variable Y is to increase right skewness. If the function φ is increasing, this effect can be explained by noting that the derivative of a convex function is greater on the right tail of the distribution than on the left. Thus, values on the right will be spread out more than those on the left. The argument still holds if the function φ is decreasing, since the left tail of Y becomes the right tail of $\varphi(Y)$. Similarly, a concave transformation will decrease right skewness. In a general discussion on convex transformations, van Zwet (1964) makes this point into a precise statement by proving the following result: If φ is convex and $\gamma(X)$ is the standard coefficient of skewness of a random variable X given by $\gamma(X) = E(X - EX)^3/SD^3(X)$, then $\gamma(\varphi(X)) \geq \gamma(X)$. The coefficient of skewness thus defined recurs in work that concerns symmetry. In Chapter II, we will notice an example of this observation.

A transformation may have effect not only on skewness, but also on heteroscedasticity. Bartlett (1947) noted that if random variables Y_i have means μ_i and standard deviations σ_i related by $\sigma_i = g(\mu_i)$, then $\varphi(Y_i)$ will have approximately constant variance if $(d/dy)\varphi(y)$ is proportional to $[g(y)]^{-1}$.

Box and Cox (1964) proposed one possible approach for incorporating a transformation in regression: Extend model (1.1) by assuming that it is not the Y_i's themselves which possess symmetric distributions with constant variance (given X_i's), but rather transformed values $h(Y_i)$ for some function $h(\cdot)$ with reasonable properties (e.g. monotone). Considering a family of monotonic functions in y, $\{ h(y,\lambda) : \lambda \in \Lambda \subset \mathbb{R}^q \}$, provides a parameterized version of the idea. We regard λ as an additional parameter and assume that for some λ,

$$h(Y_i,\lambda) = f(X_i,\beta) + \sigma\epsilon_i \tag{1.2}$$

where the ϵ_i are independent, identically distributed, and symmetric random variables. If $h(y,\lambda) = y$ for some λ, then (1.2) includes (1.1) as a special case.

We must now estimate λ and β simultaneously. Although greater efficiency in estimating β will be attained as a consequence, the estimation of λ by itself may also be of interest since it may provide insight into the relationship between X and Y. Furthermore, we would be in a better position to model the complete conditional distribution of Y given X if we were able to recognize a simple distributional form for the errors ϵ.

Non-parametric approaches to the transformation problem are also possible, although we will not pursue them any further. In this direction, we point out the work of Breiman and Friedman (1985) in which the following model is a special case:

$$\theta(Y_i) = \sum_{j=1}^{k} \phi_j(X_{ij}) + \sigma\epsilon_i$$

where θ, ϕ_j, j=1,..., k are (nonparameterized) functions, ϵ_i are standard normal, and $\{ \phi_j(X_{ij}) : j=1,\ldots, k \}$ are jointly normal. These authors provide a method for estimating θ, ϕ_j j=1,..., k using smoothing techniques, and call it the ACE algorithm (after Alternating Conditional Expectations). They obtain this algorithm as a consequence of maximizing the correlation between $\theta(Y)$ and

$$\phi(X_1,\ldots, X_p) = \sum_{j=1}^{k} \phi_j(X_j).$$

They state that if the $\phi_j(X_j)$ are not normal, then the estimates are approximately correct. In a different approach, Hastie and Tibshirani (1986) treat the problem of estimating $E(Y|X)$ when the assumed form is

$$E(Y|X) = \sum_{j=1}^{k} \varphi_j(X_j).$$

Note that this latter view deals with transformations of the independent variables but not the response.

In section 1.1, we review the transformation literature, in particular, that of the Box-Cox transformation which deals with the case when only the response is transformed. We will emphasize the parametric point of view of a transformation, as in equation (1.2). In section 1.2 we take a different approach to transformations when we consider the Transform-Both-Sides model of Carroll and Ruppert (1984). We shall briefly discuss some properties of these transformations as well as the maximum likelihood estimators. We also make some comparisons with the Box-Cox transformation model in this section. In section 1.3, we introduce some alternative estimators of transformation parameters, which will later be the subject of study in Chapter II.

1.1 Transformations

Box and Cox (1964) proposed the following family of transformations and a corresponding special case of model (1.2):

$$y^{(\lambda)} = h(y,\lambda) = \begin{cases} (y^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}. \tag{1.3}$$

We shall refer to this family of transformations as the *Modified Power Transformation*, which is valid for positive y. We notice that model (1.2) with $\lambda = 1$ corresponds to the "no Transformation" in model (1.1).

A transformation in the family (1.3) is either concave or convex according to $\lambda < 1$ or $\lambda > 1$. Thus, the effect on the skewness of Y is intuitively clear according to the heuristic explanation mentioned in section, 1.0: $\lambda < 1$ decreases right skewness; $\lambda > 1$ increases right skewness.

Box and Cox (1964) assume that for some λ,

$$(1.4) \qquad Y_i^{(\lambda)} = X_i^T\beta + \sigma\epsilon_i$$

where the ϵ_i are assumed to be independent and identically distributed random variables having the standard normal distribution with density ϕ. Furthermore, they propose to estimate λ by maximum likelihood. We shall denote this estimator by $\hat{\lambda}_{BC}$. Because of the form of the postulated model, we notice that this transformation attempts to simultaneously attain three properties: normality, simplicity in structure for $E(Y|X)$, and homoscedasticity.

Various authors return to the study of the Box and Cox transformation in later papers. Draper and Cox (1969) obtain an approximate expression for the asymptotic variance of $\hat{\lambda}_{BC}$, while Andrews (1971) and Atkinson (1973) consider the problem of testing and constructing confidence intervals for the parameters of the model. Hernandez and Johnson (1980) show that $\hat{\lambda}_{BC}$ converges to a value that minimizes the Kullback-Leibler information between the underlying density of Y in model (1.2) and a normal density with the same mean and variance. They provide an interpretation of the estimate, which is that $\hat{\lambda}_{BC}$ transforms Y to the "nearest" possible normal density, even if the transformed Y does not appear normal at all. Bickel and Doksum (1981) consider the Box-Cox model with a general density g for the ϵ_i instead of ϕ. They study consistency and consider the estimate from the robustness point of view. They also verify the fact that $\hat{\lambda}_{BC}$ is not consistent for λ except in the case $g = \phi$ or $\lambda = 0$ (Bickel and Doksum (1981), Hinkley (1975)).

The problem of estimating the regression parameter β when λ has also been estimated is an issue of controversy. Box and Cox (1964) advocate first estimating λ, and then treating the estimate $\hat{\lambda}$ as fixed for computation of the covariance matrices of $\hat{\beta}$. In their case, estimation is accomplished using linear least squares and then "standard analysis" on the estimated transformation is employed. Part of the reason for the controversy arises from the fact that λ is treated as unknown and has to be estimated. The covariance matrix of the β parameter suffers great modification, when compared to the same matrix when λ is known and only β has to be estimated. We refer to the following references for further details: Bickel and Doksum (1981); Draper and Cox (1969); Carroll and Ruppert (1981); Box and Cox (1982); and Hinkley and Runger (1984). We shall only mention that the problem has its origin in the fact that $\hat{\lambda}$ and $\hat{\beta}$ are highly correlated.

On the other hand, when the goal is to perform inference about Y in the original scale of observations, the effect of not knowing λ and having to estimate it is only moderate. Carroll and Ruppert (1981) discuss estimation of the conditional median of Y given X, and Taylor (1986) considers a pair of methods for estimating the conditional mean. In both conditioning situations, they discovered that there is only a small to moderate cost due to estimating λ, when cost is measured by the Mean Squared Error of estimation.

Several authors have proposed other estimators for λ and β which are based on ideas that are different from maximum likelihood. Most of these estimators have their roots in the one-sample model $Y_i^{(\lambda)} = \mu + \sigma\epsilon_i$, where the density of the ϵ_i is denoted by g. Some of the methods can be extended heuristically to the general case. We now briefly describe some of the approaches:

(a) Hinkley (1975) assumes that g is a symmetric distribution and proposes finding λ to solve

$$\hat{\xi}_p^{(\lambda)} + \hat{\xi}_{1-p}^{(\lambda)} - 2\hat{\xi}_{½}^{(\lambda)} = 0$$

for fixed $p \in (0,1/2)$ where $\hat{\xi}_q^{(\lambda)}$ is the sample q-percentile of $\{ Y_i^{(\lambda)}\}$ i=1,..., n.

(b) In a later paper, Hinkley (1977) solves for

$$\left[\bar{Y}^{(\lambda)} - \hat{\xi}_{.5}^{(\lambda)}\right] / \hat{\sigma}^{(\lambda)} = 0$$

where

$$[\hat{\sigma}^{(\lambda)}]^2 = (1/n) \sum_{i=1}^{n} (Y_i^{(\lambda)} - \bar{Y}^{(\lambda)})^2$$

and

$$\bar{Y}^{(\lambda)} = (1/n) \sum_{i=1}^{n} Y_i^{(\lambda)}.$$

(c) Bickel and Doksum (1981) drop the assumption that g is normal by modifying the likelihood equation used in the estimation of β. They extend the estimating procedure to include some estimators that have some robust properties, and they consider the simplifying assumption $\sigma \longrightarrow 0$ in addition to $n \longrightarrow \infty$ for asymptotic results. This small σ approach also appears in the work of Draper and Cox (1969), and later in Carroll and Ruppert (1984) and Kettl (1987).

(d) Some estimators for the transformation parameter λ that have some robust properties have also been considered by Carroll (1980) and Bickel and Doksum (1981). Taylor (1985) compares these estimators with the estimators of Hinkley.

(e) In the case of a single sample and symmetric errors, Taylor

(1985) considers estimators that solve the equation

$$(1.5) \qquad (1/n) \sum_{i=1}^{n} \Psi\{(Y_i^{(\lambda)} - \bar{Y}^{(\lambda)})/\hat{\sigma}^{(\lambda)}\} = 0$$

where Ψ is an odd function. He is able to obtain consistency and asymptotic normality results, as well as to give an optimal choice for the function Ψ. When the errors are normally distributed, the optimal choice for Ψ is $\Psi(u) = u^3$; thus, the left hand side of (1.5) becomes the standard coefficient of skewness.

1.2 The Transform-Both-Sides (TBS) Model

Carroll and Ruppert (1984) introduced a model which they describe as follows: Assume there exists a theoretical model that relates Y and X given by

$$(1.6) \qquad Y_i = f(X_i, \beta)$$

for a known function f and an unknown p-vector of parameters β. The form of the function f might be postulated by physical or scientific considerations, for instance by chemical theory. Snee (1986); Carroll and Ruppert (1984, 1987, 1988); Ruppert and Carroll (1985); Kettl (1987); Bates, Wolf and Watts (1986); and Ruppert, Cressie, and Carroll (1989) provide many examples of f with applications of the TBS methodology from several fields of study.

Errors in measurement and random variation will cause the relation (1.6) not to be observed exactly. Fitting model (1.1) by least squares is a conceivable solution, but after a fit, the residuals may show skewness and/or heteroscedasticity. This aspect is an indication that the optimal properties of the least-squares estimator are spoiled;

therefore a transformation of Y to h(Y) may be suggested in order to remove skewness, induce homoscedasticity, or both. If the transformation is monotonic, it is possible to preserve the theoretical relationship (1.6) between X and Y by transforming $f(X,\beta)$ to $h(f(X,\beta))$ as well.

If we let $\{ h(y,\lambda), \lambda \in \Lambda \subset \mathbb{R}^q\}$ denote a family of monotonic transformations, the TBS model states that for some values of (λ,β,σ),

$$h(Y_i,\lambda) = h(f(X_i,\beta),\lambda) + \sigma\epsilon_i \tag{1.7}$$

where the ϵ_i are i.i.d. symmetric (mean zero) random variables with unit variance. The modified power transformation of Box and Cox is a special case for the family $\{ h(y,\lambda)\}$. The complete designation for this choice would be TBS with Box-Cox transformation, but frequently we shall simply write TBS.

The TBS model (1.7) allows randomness to affect the relationship $Y = f(X,\beta)$ in different ways, not necessarily in an additive fashion. It is also capable of explaining skewness and heteroscedasticity that may emerge when Y is fitted to $f(X,\beta)$. We maintain the relationship (1.6) in the sense that $Y = f(X,\beta)$ in the absence of error, because we apply the same transformation to "both sides" of the theoretical equation (1.6) and $h(y,\lambda)$ is monotonic. This concept is the main difference from the Box-Cox model (1.4), in which only the response Y is transformed using a modified power transformation, and $f(X,\beta)$ is postulated as linear in order to obtain a simple form for the regression. In other words, the choice $f(X,\beta) = X^T\beta$ in the Box-Cox model does not arise from a scientific theory; the form $X^T\beta$ is chosen instead in order to obtain a model that allows for "simplicity of structure for E(Y)" (Box and Cox (1964)). Carroll and Ruppert (1984, 1988) emphasize that the TBS model (1.7) is not to be considered a replacement or generalization for the model (1.4), since each model is directed toward achieving different goals. They also note that within the TBS formulation, $f(X,\beta)$ can be interpreted as the conditional median of Y given X, regardless of the real value of λ. In contrast, in the Box-Cox model, $f(X,\beta)$ is the median (and mean) of $Y^{(\lambda)}$ for the true value of λ, but $f(X,\beta)$ has no physical interpretation otherwise.

We will now deal with the question of estimation in the TBS model. Carroll and Ruppert (1988, chapter 4) discuss estimation of (β,λ,σ) by maximum likelihood in the case of normal errors and the modified power transformation. They show that the MLE estimate is found by minimizing

$$\sum_{i=1}^{n} (e_i(\lambda,\beta)/\dot{Y}^{\lambda-1})^2$$

where $\dot{Y}$ is the geometric mean of $Y_1,\dots, Y_n$ and each $e_i(\lambda,\beta)$ is the residual defined by

$$e_i(\lambda,\beta) = Y_i^{(\lambda)} - f^{(\lambda)}(X_i,\beta). \tag{1.8}$$

There is a significant feature that TBS possesses; in contrast to the transform-the-response model discussed in the previous section, Carroll and Ruppert (1984) found that in TBS, $\hat{\lambda}$ and $\hat{\beta}$ are only weakly related. Thus, the controversy between treating $\hat{\lambda}$ as fixed and acknowledging that it is random does not exist as it does in the Box-Cox transformation model.

Its sensitivity to outliers (Carroll and Ruppert (1985)) and its inconsistency under lack of normality of the errors (Carroll and Ruppert (1988)), are two disadvantages of the maximum likelihood estimator. We note, however, that since the responses are assumed positive, the distribution of ϵ in models (1.4) and (1.7) cannot have infinite support for all λ. It is more desirable to assume only approximate normality (perhaps up to truncation) or symmetry on a compact set, and to develop estimation procedures that could be applied under these circumstances. Examples of methods that follow this trend are the estimates of Hinkley and Taylor mentioned above, and the Ruppert-Aldershof estimates to be defined in section 1.3.

In the context of maximum likelihood estimation, several methods exist to obtain large sample estimates of the covariance matrix of $(\hat{\beta},\hat{\lambda},\hat{\sigma})$; Carroll and Ruppert (1988) recognize at least six methods and give some guidelines in connection with their use. Once the covariance matrix is estimated, we would then apply standard procedures to perform inference on the parameters. For example, the asymptotic normality of the MLE can be used for the construction of Wald-type confidence intervals and tests; another alternative is likelihood-ratio confidence regions and tests.

Sometimes inference in the original untransformed scale of the Y observations will be required. Carroll and Ruppert (1988, chapter 4) first notice that TBS provides the following model for the observations:

$$Y = h^{-1}\{[h(f(X,\beta),\lambda) + \epsilon],\lambda\}$$

where $h^{-1}(y,\lambda)$ is the inverse of the transformation $h(y,\lambda)$ as a function of y. If the distribution of the ϵ_i is denoted by F, then the q-th quantile of Y given X is

$$Q_q(Y|X) = h^{-1}\{[h(f(X,\beta),\lambda) + F^{-1}(q)],\lambda\} \tag{1.9}$$

for $q \in (0,1)$, and the conditional mean of Y given X is

$$E(Y|X) = \int h^{-1}\{[h(f(X,\beta),\lambda)+\epsilon],\lambda\}\, dF(\epsilon). \tag{1.10}$$

Through (1.9) and (1.10), we gain partial information on the conditional distribution of Y given X. One way to obtain estimates of the above two quantities is by substituting β, λ, and $F(\epsilon)$ for estimates $\hat{\beta}$, $\hat{\lambda}$, and $\hat{F}(\epsilon)$. A few different possibilities exist for estimating F. For instance, if F is assumed to be $N(0,\sigma^2)$, then an estimate for F is the $N(0,\hat{\sigma}^2)$ distribution. If we estimate F using the empirical distribution $\hat{F}_n$ of the residuals $\{ e_i(\lambda,\beta) \}$, we obtain the estimator for $E(Y|X)$ which Duan (1983) called the "smearing estimator".

Carroll and Ruppert (1987) also address the problem of detecting outlying observations and assessing their influence on the estimate of (β,λ) in the TBS model. They develop a case-deletion diagnostic that avoids recomputation of estimates for β and λ, in order to evaluate the effect of an observation on the estimates. They also propose a robust estimator by solving

$$\sum_{i=1}^{n} w_i(Y_i,\lambda,\beta)\nabla\ell_i(Y_i,\lambda,\beta,\hat{\sigma}(\lambda,\beta)) = 0$$

where ℓ_i is the log-likelihood, w_i is a weighting function, and

$$\hat{\sigma}^2(\lambda,\beta) = \{\sum_{i=1}^{n} w_i(Y_i,\lambda,\beta)e_i^2(\lambda,\beta)\}/\sum_{i=1}^{n} w_i(Y_i,\lambda,\beta)$$

is a weighted variance estimator.

1.3 Transformations to Symmetry and Homoscedasticity

When considering the transform-both-sides model (1.7), authors have proposed and studied other methods for estimating λ based on different ideas. The following estimates provide examples. Ruppert and Aldershof (1989) define the "skewness estimator" (or *transformation to symmetry*) as $\hat{\lambda}_{sk}$ the value of λ that solves

$$(1.11) \qquad (1/n) \sum_{i=1}^{n} \{e_i(\lambda, \hat{\beta}(\lambda)) / \hat{\sigma}(\lambda)\}^3 = 0$$

where

$$\hat{\sigma}^2(\lambda) = (1/n) \sum_{i=1}^{n} e_i^2(\lambda, \hat{\beta}(\lambda))$$

and $\hat{\beta}(\lambda)$ is the least-squares estimator for fixed λ. This proposal was based on the work of Hinkley (1975) and Taylor (1985) (see equation (1.5)). This estimator $\hat{\lambda}_{sk}$ then defines the estimate $\hat{\beta}_{sk} = \hat{\beta}(\hat{\lambda}_{sk})$. More generally, the Ruppert-Aldershof skewness estimator would solve the equation

$$(1.12) \qquad (1/n) \sum_{i=1}^{n} \Psi(e_i(\lambda, \hat{\beta}(\lambda)) / \hat{\sigma}(\lambda)) = 0$$

where Ψ is a function such that

$$\sum_{i=1}^{n} \Psi(e_i(\lambda, \hat{\beta}(\lambda)) / \hat{\sigma}(\lambda)) = 0$$

if the residuals $\{ e_i \}$ are symmetric (or $E\Psi(Y) = 0$ if Y is a symmetric random variable). If a value of λ exists for which the TBS model holds with symmetric errors, this procedure will estimate it consistently. In Chapter II, we will prove some asymptotic properties of this estimator.

In a different approach, Ruppert and Aldershof (1989) define the "heteroscedasticity estimator" (or *transformation to homoscedasticity*) by $\hat{\lambda}_{het}$ the value of λ that solves the equation

$$(1.13) \quad \sum_{i=1}^{n} (e_i^2(\lambda,\hat{\beta}(\lambda))/\hat{\sigma}^2(\lambda))[\log f(X_i,\hat{\beta}(\lambda)) - \overline{\log f(X,\hat{\beta}(\lambda))}] = 0$$

where $$\overline{\log f(X,\hat{\beta}(\lambda))} = (1/n) \sum_{i=1}^{n} \log f(X_i,\hat{\beta}(\lambda)).$$

We can express this approach in general form as follows: For fixed λ, let $H_n(\lambda)$ be some measure of heteroscedasticity that depends on $\{e_i(\lambda,\hat{\beta}(\lambda)\}$ and $\{f(X_i,\hat{\beta}(\lambda)\}$; an example is the measure given by the Pearson Correlation between $\{e_i^{\,2}(\lambda,\hat{\beta}(\lambda)\}$ and $\{\log f(X_i,\hat{\beta}(\lambda))\}$. We require that $H_n(\lambda)$ satisfies that $H_n(\lambda) = 0$ if there is no heteroscedasticity, $H_n(\lambda) > 0$ if variance is an increasing function of the mean, and $H_n(\lambda) < 0$ if it is decreasing. In the general setting, a heteroscedasticity estimator would solve

$$(1.14) \qquad H_n(\lambda) = 0.$$

Carroll and Ruppert (1988) also discuss these general heteroscedasticity estimators. They anticipate consistency of $\hat{\lambda}_{het}$ under the TBS model assuming homoscedastic but not necessarily symmetric errors. In their work, Ruppert and Aldershof (1989) study the form of asymptotic covariances and find that estimating the nuisance parameter σ does not affect asymptotic covariances of (β,λ). In addition, the asymptotic distribution of the estimate for λ does not depend on whether β is known or not. For small σ, the estimating equation (1.13) is optimal among families where $b(f(X_i,\hat{\beta}(\lambda))$ replaces $\log f(X_i,\hat{\beta}(\lambda))$ for monotonically increasing functions b.

Since the estimator $\hat{\lambda}_{sk}$ attempts to transform to symmetry and $\hat{\lambda}_{het}$ tries to eliminate heteroscedasticity, a natural question arises: The possibility that a value of the transformation parameter exists that attains both distinctions. In this case both of the estimators defined would estimate the same quantity. Ruppert and Aldershof (1989) propose a possible test of this hypothesis based on asymptotic normality, and explore the possibility of improving the estimate for λ by a combination of $\hat{\lambda}_{sk}$ and $\hat{\lambda}_{het}$. Thus, they propose a third estimator $\hat{\lambda}_{hsk}$, which combines the ideas of the two previous estimators. They state that it is natural to consider the linear combination

$$\hat{\lambda}_{hsk} = \hat{w}\hat{\lambda}_{het} + (1 - \hat{w})\hat{\lambda}_{sk}$$

where the weight $\hat{w}$ would be chosen to achieve minimum asymptotic variance. They examine another view which solves a linear combination with weights w and (1 - w) of the estimating equations (1.12) and (1.13) set to zero. The additional equation

$$\frac{\partial}{\partial w}\, \mathrm{var}(\hat{\lambda}_{hsk}) = 0$$

must be introduced to specify the weight. If a transformation that induces both symmetry and homoscedasticity exists, $\hat{\lambda}_{hsk}$ should produce a better estimate than either of the two components. Ruppert and Aldershof (1989) inspect some applications of all the stated estimators using real data sets.

1.4 Summary of Work

In Chapter II we will establish consistency and asymptotic normality for the Ruppert-Aldershof estimate of the transformation to symmetry. As an alternative to this approach and others described above, in this work we propose and study a new estimator of a transformation based on the notion of minimum distance. One advantage of this estimator is that it will be applicable in situations where other procedures fail to be consistent, while assuming only a symmetric distribution on the errors. In Chapter III we define this estimator, show that it belongs to a certain class of estimators (Generalized M-Estimators), and then prove consistency for this class. Asymptotic normality for estimators in this class is the subject of Chapter IV. In Chapter V we address the problem of increasing the efficiency of the estimator, and we also report the results of a Monte Carlo study for comparing the new estimate to the maximum likelihood estimate and its performance under different error distributions.

CHAPTER II

TRANSFORMATIONS TO SYMMETRY

2.0 Introduction

In Chapter I, we were introduced to the following Transform-Both-Sides regression model of Carroll and Ruppert (1984):

$$(2.1) \qquad h(Y_i,\lambda) = h(f(X_i,\beta),\lambda) + \sigma\epsilon_i$$

where $f(X,\beta)$ is of an assumed known form, β is an unknown regression parameter, $h(Y,\lambda)$ is a one-to-one transformation of Y for each fixed λ, and λ is an unknown transformation parameter that may be, in general, vector-valued. The motivation behind this specification was to induce symmetric and homoscedastic errors while preserving a theoretical relationship between X and the conditional median of Y, given X. The advantages are to increase efficiency in parameter estimation, and to make prediction intervals more accurate. Also, the model provided a way of incorporating random variation into a pre-specified relationship. When there is no prior knowledge on the structure of variability, this is an important feature.

Carroll and Ruppert (1988) explored estimation by maximum likelihood. The purpose of this chapter is to establish some asymptotic properties of one of the alternate estimating procedures introduced by Ruppert and Aldershof (1989) for the parameters in model (2.1). As we reviewed in Chapter I (section 1.3), two main procedures were proposed in that paper for estimating the transformation parameter. The first of these methods is an extension of the transformations to symmetry of Taylor (1985), and uses a measure of skewness that is applied to residuals. When errors are normally distributed, we obtain an estimator with optimal properties when the measure of skewness is chosen to be the standardized coefficient of skewness. The second proposed estimator employs a certain measure of heteroscedasticity, and was shown to be approximately optimal under small σ asymptotics. In either case, the estimates were members of the class of M-estimators, which enabled theoretical analysis. In this chapter we are concerned with the first of the two approaches.

We will assume that for $1 \leq i \leq n$, the errors ϵ_i are independent and identically distributed, and that the X_i are random variables also with a common probability distribution. Thus, the pairs $Z_i = (Y_i, X_i)$ are i.i.d. random vectors, with values in some space $\mathcal{X}$. Following Ruppert and Aldershof (1989), we introduce an additional parameter μ in order to center residuals at their mean in the definition of their transformation to symmetry. The vector of all parameters is $\theta = (\beta, \lambda, \mu)^T$, and it will be assumed to be an element of Θ, a compact subset of Euclidean space.

For convenience of later reference we now introduce the following notation.

Notation 2.1:

N1. $\theta = (\beta, \lambda, \mu) \in \Theta \subset \mathbb{R}^{p+2}$, $\beta \in \mathbb{R}^p$

N2. $r(\theta, Z) = h(Y, \lambda) - h(f(X, \beta), \lambda)$

N3. $r_i(\theta) = r(\theta, Z_i)$ $\quad 1 \le i \le n$

N4. $Z = (Y, X)$

N5. $Z_i = (Y_i, X_i)$ $\quad 1 \le i \le n$

N6. $\mathcal{X}_n = (Z_1, Z_2, \ldots, Z_n)$

N7.
$$\Psi_j(\theta, Z) = \begin{cases} r(\theta,Z)(\partial/\partial\theta_j)r(\theta,Z) & 1 \le j \le p \\ [r(\theta,Z) - \mu]^3 & j = p+1 \\ r(\theta,Z) - \mu & j = p+2 \end{cases}$$

N8. $G_j(\theta) = E\Psi_j(\theta, Z)$ $\quad 1 \le j \le p+2$

N9. $\Psi_j^{(n)}(\theta, \mathcal{X}_n) = (1/n) \sum_{i=1}^{n} \Psi_j(\theta, Z_i)$ $\quad 1 \le j \le p+2$

N10. $\Psi^{(n)} : \mathbb{R}^{p+2} \times \mathcal{X}^n \longrightarrow \mathbb{R}^{p+2}$ is the function with coordinates $\Psi_j^{(n)}$

N11. $H(\theta) = \left[E[\frac{\partial}{\partial\theta_\ell} \Psi_j(\theta, Z)]\right]$ (a p+2 square matrix).

■

Definition 2.2: (Ruppert and Aldershof (1989))

The *Ruppert-Aldershof Skewness Estimator* $\hat{\theta}_n$ is defined as a solution to the following equation:

$$\Psi^{(n)}(\theta, \mathcal{X}_n) = 0. \tag{2.2}$$

■

2.1 General Results About Consistency

The first property to be investigated about this estimator is the existence of a sequence of solutions to equation (2.2) that is consistent. This existence, as well as asymptotic normality of a consistent solution, will be the topics of section 2.2. In this section, we give some useful lemmas and develop some foundations. An indexed set of functions on T to $\mathbb{R}^q$, $\mathcal{F} = \{ f(t,x) : x \in A \}$, is said to be *equicontinuous in t for* $x \in A$, if for each $\epsilon > 0$ there is a $\delta(\epsilon) > 0$ such that if

$$t, t' \in T \text{ and } | t - t' | < \delta,$$

then

$$\| f(t,x) - f(t',x) \| < \epsilon$$

for all $x \in A$.

Lemma 2.3: (Rubin (1956))

Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed random variables with values in an arbitrary space X, having common distribution P. Let T be a compact topological space, and let f be a function on TxX, measurable for each $t \in T$. Define

$$H_n(t) = (1/n) \sum_{j=1}^{n} f(t, X_j) \text{ and}$$

$$H(t) = \int f(t,x)\, dP(x).$$

If there is an integrable g such that $| f(t,x)| < g(x)$ for all $t \in T$ and $x \in X$, if there is a sequence S_i of measurable sets such that

$$P(X - \bigcup_{i=1}^{\infty} S_i) = 0,$$

and for each i, $f(t,x)$ is equicontinuous in t for $x \in S_i$, then with probability one,

$$\sup_{t \in T} | H_n(t) - H(t)| \longrightarrow 0$$

as $n \longrightarrow \infty$ and the limit function $H(t)$ is continuous.

■

Two ways exist which are used to define M-estimators: By minimizing a sample-based function, or by solving a system of equations (appropriately called the *estimating equations*). In the next lemma, we see how consistency of M-estimation follows. Condition (2.3) below is strong, but will be satisfied for the cases that we are concerned with in this chapter; in fact, (2.3) will be implied by Lemma 2.3. This condition may be replaced by other assumptions (Huber (1967)), but in its present form it also serves well to illustrate the principle of M-estimation.

Lemma 2.4: Let $(T, \mathcal{T}, P_T)$ be a probability space, and let $H_n(t)$ and $H(t)$ be defined as in Lemma 2.3. If

$$\sup_{t \in T} | H_n(t) - H(t)| \xrightarrow{a.s.} 0 \qquad (2.3)$$

and either

(a) $\hat{t}_n$ is defined by

$$\inf_{t \in T} H_n(t) = H_n(\hat{t}_n)$$

and the function $H(t)$ has a unique minimum at $t = t_o \in T$

or,

(b) $\hat{t}_n$ is defined by

$$H_n(\hat{t}_n) = 0$$

and the equation $H(t)$ has a unique root at $t = t_o \in T$,

then $\hat{t}_n \longrightarrow t_o$ with P_T-probability 1.

proof: Fix $\epsilon > 0$ and define

$$A_\epsilon = T \cap \{ t: | t - t_o| \geq \epsilon \}.$$

We wish to show that

$$P_T(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{ | \hat{t}_n - t_o| < \epsilon \}) = 1$$

or equivalently

$$P_T(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{ \hat{t}_n \notin A_\epsilon \}) = 1.$$

Case (a): Since A_ϵ is compact,

$$\inf_{t \in A_\epsilon} H(t)$$

is attained, and by assumption, strictly greater than $H(t_o)$, say greater than $H(t_o) + 3\delta$ for some $\delta > 0$.

If (2.3) holds, then with probability one there exists an integer N_o such that for all $n > N_o$,

$$\sup_{t \in T} | H_n(t) - H(t)| < \delta$$

and therefore with probability one

$$H(t) - 2\delta < H(t) - \delta \leq H_n(t) \leq H(t) + \delta$$

for all t and all $n > N_0$.

This implies

$$\inf_{t \in A_\epsilon} H_n(t) \geq \inf_{t \in A_\epsilon} H(t) - 2\delta$$

and

$$\inf_{t \notin A_\epsilon} H_n(t) \leq \inf_{t \notin A_\epsilon} H(t) + \delta = H(t_0) + \delta$$

and thus for all $n > N_0$,

$$\inf_{t \in A_\epsilon} H_n(t) \geq \inf_{t \in A_\epsilon} H(t) - 2\delta > H(t_0) + \delta \geq \inf_{t \notin A_\epsilon} H_n(t),$$

thus showing that $H_n(t)$ is not minimized on A_ϵ, that is, $\hat{t}_n \notin A_\epsilon$.

Case (b): Since H(t) is continuous, there exists $\delta > 0$ such that if $t \in T$ and $|t - t_0| \geq \epsilon$, then $|H(t)| > \delta$. If (2.3) holds, then there exists an integer N_0 such that with probability one,

$$\sup_{t \in T} |H_n(t) - H(t)| < \delta/2$$

for all $n > N_0$.

Since

$$|H(t)| - |H_n(t) - H(t)| \leq |H_n(t)|$$

we have

$$\inf_{t \in A_\epsilon} |H(t)| - \sup_{t \in A_\epsilon} |H_n(t) - H(t)| \leq \inf_{t \in A_\epsilon} |H_n(t)|.$$

Thus, with probability one, for all $n > N_0$,

$$0 < \delta - \delta/2 \leq \inf_{t \in A_\epsilon} |H_n(t)|,$$

showing that $H_n(t)$ has no roots on A_ϵ, that is $\hat{t}_n \notin A_\epsilon$.

■

It is convenient at this point to establish some preliminary results. The setting shall be more general than the situation just described for the Ruppert-Aldershof transformation to symmetry. We will make provisions for estimators defined by other sets of M-estimating equations as well, for example, the equations that define the Ruppert-Aldershof heteroscedasticity estimator. The dimension of the parameter that is being estimated need not be equal to one.

<u>Lemma 2.5</u>: Let $X_1, X_2, \ldots$ be i.i.d. random vectors on some probability space $(\mathcal{X}, \mathcal{F}, P)$, and $\Theta \subset \mathbb{R}^p$ a compact parameter space. Denote $\mathcal{X}_n = (X_1, \ldots, X_n)$, and assume a given set of p functions $\Psi_j(\theta, X) : \Theta \times \mathcal{X} \longrightarrow \mathbb{R}$, for $1 \leq j \leq p$. Define

$$G_j(\theta) = E\Psi_j(\theta, X),$$

$$\Psi_j^{(n)}(\theta, \mathcal{X}_n) = (1/n) \sum_{i=1}^{n} \Psi_j(\theta, X_i),$$

$\Psi^{(n)}: \mathbb{R}^p \times \mathcal{X}^n \longrightarrow \mathbb{R}^p$ is the (random) function with coordinates $\Psi_j^{(n)}(\theta)$, and

$$H(\theta) = E\left[\frac{\partial}{\partial \theta} \Psi^{(n)}(\theta, \mathcal{X}_n) \right].$$

If the following assumptions hold:

<u>Assumptions 2.6</u>:

A1. There exists a $\theta_o \in \Theta$ such that $G_j(\theta_o) = 0$ for $1 \leq j \leq p$. Without loss of generality we may assume $\theta_o = 0$.

A2. The support of X is bounded.

A3. $H(\theta_o)$ is nonsingular.

A4. Third order partials of $\Psi_j(\theta, X)$ with respect to θ exist and are a.s. continuous, $1 \leq j \leq p$.

Then, with probability one, for $\epsilon > 0$ sufficiently small, there exists N_ϵ and a sequence $\hat{\theta}_n$ such that for all $n > N_\epsilon$,

$$\Psi^{(n)}(\hat{\theta}_n, \mathcal{X}_n) = 0 \tag{2.4}$$

and such that

$$\| \hat{\theta}_n - \theta_0 \| < \epsilon.$$

<u>proof</u>: A key step in the proof will utilize Brouwer's Fixed Point Theorem from analysis, which states that any continuous function f with domain $D = \{ x \in \mathbb{R}^p : \| x \| \leq B \}$ for some $B > 0$ and range contained in D has at least one point $u \in D$ such that $f(u) = u$.

The system of equations (2.4) is equivalent to

$$\Gamma^{(n)}(\theta, \mathcal{X}_n) =: H^{-1}(0)\Psi^{(n)}(\theta, \mathcal{X}_n) = 0. \tag{2.5}$$

We have assumed enough conditions so that Lemma 2.3 can be applied to insure that

$$\sup_{\theta} \left\| \Psi_j^{(n)}(\theta, \mathcal{X}_n) - G_j(\theta) \right\| \xrightarrow{a.s.} 0 \tag{2.6}$$

as $n \longrightarrow \infty$, for all $1 \leq j \leq p$, and

$$H(\theta) = \left[\frac{\partial}{\partial \theta_\ell} G_j(\theta) \right],$$

that is, differentiation and expectation do interchange.

Let $H = H(0)$, and denote the (j, ℓ)-th entry of H^{-1} by $h^{(j\ell)}$. Also, for $1 \leq j \leq p$, let

$$\Gamma_j(\theta) = \sum_{\ell=1}^{p} h^{(j\ell)} G_\ell(\theta)$$

and

$$\Gamma_j^{(n)}(\theta, \mathcal{X}_n) = \sum_{\ell=1}^{p} h^{(j\ell)} \Psi_\ell^{(n)}(\theta, \mathcal{X}_n).$$

By (2.6), we conclude with probability one,

$$\sup_\theta \left\| \Gamma_j^{(n)}(\theta, \mathcal{X}_n) - \Gamma_j(\theta) \right\| \xrightarrow{a.s.} 0, \ 1 \le j \le p. \tag{2.7}$$

Expanding to first order about $\theta = 0$, we obtain, as $\| \theta \| \longrightarrow 0$,

$$\Gamma_j(\theta) = \sum_{\ell=1}^{p} h^{(j\ell)} G_\ell(\theta) = \sum_{\ell=1}^{p} h^{(j\ell)} \{ G_\ell(0) + \sum_{k=1}^{p} \theta_k [\frac{\partial}{\partial \theta_k} G_\ell(\theta)]_{\theta=0}$$

$$+ o(\| \theta \|) \}$$

$$= \theta_j + \sum_{\ell=1}^{p} h^{(j\ell)} o(\| \theta \|) = \theta_j + o(\| \theta \|),$$

using the fact that $G(0) = 0$.

This implies that $\| \Gamma(\theta) - \theta \| = o(\| \theta \|)$, so there exists a $\delta > 0$ such that $\| \theta \| < \delta$ implies $\| \Gamma(\theta) - \theta \| < \| \theta \|/2$. Restrict ϵ to be smaller than δ. Then $\| \theta \| < \epsilon$ implies

$$\| \Gamma(\theta) - \theta \| < \| \theta \|/2 < \epsilon/2. \tag{2.8}$$

Since (2.7) implies

$$\sup_\theta \left\| \Gamma^{(n)}(\theta, \mathcal{X}_n) - \Gamma(\theta) \right\| \xrightarrow{a.s.} 0,$$

with probability one there exists an N_ϵ such that for all θ

$$\left\| \Gamma^{(n)}(\theta, \mathcal{X}_n) - \Gamma(\theta) \right\| < \epsilon/2 \quad \text{for all } n > N_\epsilon. \tag{2.9}$$

Using the inequality

$$\left\| \theta - \Gamma^{(n)}(\theta,\mathcal{X}_n) \right\| \leq \left\| \Gamma^{(n)}(\theta,\mathcal{X}_n) - \Gamma(\theta) \right\| + \left\| \Gamma(\theta) - \theta \right\|,$$

as well as inequalities (2.8) and (2.9), we may conclude that with probability one, there exists an integer N_ϵ such that

$$\left\| \theta - \Gamma^{(n)}(\theta,\mathcal{X}_n) \right\| \leq \epsilon \text{ for all } n > N_\epsilon \text{ on the set } \| \theta \| < \epsilon.$$

By Brouwer's Fixed Point Theorem mentioned above, for all $n > N_\epsilon$, there exists $\hat{\theta}_n$ such that $\| \hat{\theta}_n \| < \epsilon$ and $\hat{\theta}_n - \Gamma^{(n)}(\hat{\theta}_n,\mathcal{X}_n) = \hat{\theta}_n$, that is, $\Gamma^{(n)}(\hat{\theta}_n,\mathcal{X}_n) = 0$. Equation (2.4) now follows (2.5).

■

Under the same notation and assumptions, we have the following Lemma, which amounts to the proof of the existence of a consistent sequence of solutions to a given set of M-estimating equations.

Lemma 2.7: Suppose Assumptions 2.6 hold. Then, with probability 1, there exists a sequence $\{ \hat{\theta}_n \}$ such that $\hat{\theta}_n \longrightarrow \theta_o$ and $\psi^{(n)}(\hat{\theta}_n,\mathcal{X}_n) = 0$.

proof: We denote the underlying probability space by (Ω, A, P). For fixed ϵ in Lemma 2.5, there is a set Ω_ϵ such that $P(\Omega_\epsilon) = 1$, and such that for all $\omega \in \Omega_\epsilon$, there is an N_ϵ and a solution $\hat{\theta}_n$ such that

$$\psi^{(n)}(\hat{\theta}_n,\mathcal{X}_n) = 0 \text{ and } \| \hat{\theta}_n - \theta_o \| < \epsilon$$

for all $n > N_\epsilon$.

Denote the explicit dependence of $\hat{\theta}_n$ on ω and ϵ, by $\hat{\theta}_{n\epsilon}(\omega)$. Define

$$\Omega_o = \bigcap_{j=1}^{\infty} \Omega_{1/j}$$

and notice that $P(\Omega_0) = 1$. Consider $\omega \in \Omega_0$. Without loss of generality, the $N_{1/j}$ which exist by Lemma 2.5 can be chosen to satisfy

$$N_1(\omega) \le N_{1/2}(\omega) \le N_{1/3}(\omega) \le \cdots$$

If $\omega \in \Omega_0$, define

$$\hat{\theta}_n(\omega) = \begin{cases} \hat{\theta}_{n,1/k}(\omega) & \text{if } N_{1/k}(\omega) \le n < N_{1/(k+1)}(\omega) \quad k=1,2,\ldots \\ 0 & \text{if } n < N_1(\omega) \end{cases}$$

and if $\omega \notin \Omega_0$, $\hat{\theta}_n(\omega) = 0$ for all n.

The sequence of random variables $\hat{\theta}_n$ achieves the conclusion of the lemma.

■

2.2 Consistency and Asymptotic Normality of the Transformation to Symmetry

We now establish that if the TBS model (2.1) holds for some parameter values and symmetric errors, then with probability one, the Ruppert-Aldershof estimating equations based on skewness admit a solution $\hat{\theta}_n$ that is consistent. A problem which we will have to address after this result is obtained, is how to identify the consistent solutions in practice since the estimating equations may have multiple roots. Lemma 2.7 only states that a consistent solution exists.

The next theorem makes use of the Notation 2.1 and the definition of the Ruppert-Aldershof estimator via the estimating equation (2.2).

Theorem 2.8: Consider the TBS model (2.1) and the conventions of Notation 2.1, in particular let the function Ψ and the matrix H be defined by N7 and N11, respectively. Assume there exists a parameter value $\theta_0 \in \Theta$ such that the TBS model (2.1) holds. Suppose $Z_i = (Y_i, X_i)$,

$1 \leq i \leq n$ are i.i.d. observations obeying the model, and that the following assumptions hold:

B1. The distribution of ϵ is symmetric on a bounded support.

B2. The support of the random variable X is bounded.

B3. The matrix $H(\theta_o)$ is nonsingular.

B4. The third order partial derivatives of $f(X,\beta)$ with respect to β all exist and are a.s. continuous.

B5. The third order partial derivatives of $h(Y,\lambda)$ with respect to λ all exist and are a.s. continuous.

B6. The parameter space Θ is compact.

B7. For each i, the error ϵ_i and the X_i are independent.

Then, with probability one, there exists a sequence of solutions $(\hat{\beta}_n,\hat{\lambda}_n)$ to the Ruppert-Aldershof equations (2.2) such that

$$(\hat{\beta}_n,\hat{\lambda}_n) \longrightarrow (\beta_o,\lambda_o).$$

proof: The theorem will be a direct consequence of Lemma 2.7, once we identify that the Ruppert-Aldershof equations (2.2) are a special case through the particular form of the function Ψ given by N7 of Notation 2.1.

Since a parameter value θ_o such that the model holds exists, we check that $G_j(\theta_o) = E\Psi_j(\theta_o,Z) = 0$ for all $j=1,\ldots,p+2$, as follows. In view of the fact that under the assumed model,

$$r(\theta_o,Z) = \sigma_o\epsilon$$

and

$$\frac{\partial}{\partial\beta} r(\theta,Z) = -h'(f(X,\beta),\lambda) \frac{\partial}{\partial\beta} f(X,\beta)$$

is independent of ϵ, we have by B1

$$Er(\theta_o,Z) = Er^3(\theta_o,Z) = 0$$

and by B7,

$$E\left[r(\theta_o,Z) \frac{\partial}{\partial\beta} r(\theta_o,Z)\Big|_{\beta=\beta_o}\right] = -\sigma_o E[\epsilon]\, E\left[\frac{\partial}{\partial\beta} r(\theta_o,Z)\Big|_{\beta=\beta_o}\right] = 0.$$

Thus A1 of Assumptions 2.6 holds. The remaining assumptions of Lemma 2.7 are satisfied through assumptions B1-B7. In particular, the required smoothness properties of the functions $\Psi_j(\theta,Z)$ for $1 \le j \le p+2$ are a consequence of the smoothness of the functions $r(\theta,Z)$, $f(X,\beta)$, and $h(Y,\lambda)$.

■

Next we consider a consistent solution $\hat{\theta}_n$. Asymptotic normality under i.i.d. observations is established in the following result, which is set forth by Ruppert and Aldershof (1989), but is stated here for completeness.

<u>Theorem 2.9:</u> Under the assumptions of Theorem 2.8, any sequence of solutions to (2.2) such that $\hat{\theta}_n \xrightarrow{P} \theta_o$, satisfies

$$n^{1/2}(\hat{\theta}_n - \theta_o) \xrightarrow{d} N(0,\ A^{-1}BA^{-T})$$

where

$$A = \frac{\partial}{\partial\theta} E\psi(\theta,Z)\Big|_{\theta=\theta_o},$$

$$B = E[\psi(\theta_o,Z)\psi^T(\theta_o,Z)],$$

$$A^{-T} = (A^{-1})^T,$$

$$\psi(\theta,Z) = (\Psi_1(\theta,Z), \ldots, \Psi_{p+2}(\theta,Z))^T,$$

and $\Psi_j(\theta,Z)$, $j=1,\ldots, p+2$ are the functions defined by N7 in Notation 2.1.

proof: The Ruppert-Aldershof estimator $\hat{\theta}_n$ satisfies

$$(1/n) \sum_{i=1}^{n} \psi(\hat{\theta}_n, Z_i) = 0$$

and thus it is an M-Estimator. Asymptotic normality will follow from an application of Huber's (1967) Theorem 3, on general M-Estimation theory.

■

2.3 Identification of a Consistent Solution in the Transformation to Symmetry

We now turn to the problem of identifying a sequence of solutions to the Ruppert-Aldershof equations that is consistent. The problem exists because of fact that several solutions might exist, all of which may not necessarily converge to the unknown parameter value.

We will describe a device which will enable us to compute a certain estimate that will possess the same asymptotic normal distribution as stated in Theorem 2.9. A preliminary step shall be to estimate the regression parameter β consistently, even in the case that the transformation parameter λ is unknown. This can be achieved under the TBS formulation by using the Least Absolute Deviations estimate for β, to be described below.

Once this preliminary estimate for β is obtained, a consistent estimate for λ will be constructed, and furthermore, a one-step procedure will be applied in order to produce an estimate of (β, λ) with the desired asymptotic properties. In Theorem 2.13, we will state this stepwise estimating procedure with greater detail. We first define the Least Absolute Deviation (LAD) estimator for β.

<u>Definition 2.10</u>: Consider a general non-linear regression model

$$Y_i = f(X_i, \beta_o) + \epsilon_i \qquad 1 \leq i \leq n$$

where each ϵ_i has median equal to zero (they need not be symmetrically distributed).

The *Least Absolute Deviation* (LAD) estimator for β based on observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, denoted by $\hat{\beta}_{LAD}$, is defined by

$$(1/n) \sum_{i=1}^{n} | Y_i - f(X_i, \hat{\beta}_{LAD}) | = \inf_{\beta} (1/n) \sum_{i=1}^{n} | Y_i - f(X_i, \beta) |.$$

■

This estimator has desirable properties that are a consequence of the fact that the conditional median of Y given X is $f(X, \beta_o)$ and the fact that $E| Y - f(X, \beta) |$ as a function of β is minimized at $\beta = \beta_o$. Bassett and Koenker (1978) investigate the least absolute deviation estimator in the linear regression case. They find that the LAD and the Least Squares estimators compare in much the same way as the sample median and the sample mean compare when estimating location in a single sample. In the nonlinear case, Oberhofer (1982) investigates general sufficient conditions for consistency to the true value of β_o.

A distinguishing property of the TBS regression model with median-zero errors, is that the conditional median of Y given X is still

$f(X,\beta_o)$, regardless of the value of λ. This makes LAD estimation a likely method for estimating β consistently even if λ is unknown, although it may not be fully efficient.

Lemma 2.11: Consider the TBS regression model with errors that have a unique median at zero, and the assumptions of Theorem 2.8. If the regression function $f(X,\beta)$ is such that $\beta \neq \beta_o$ implies $f(X,\beta) \neq f(X,\beta_o)$ on a set of positive measure, then the estimator $\hat{\beta}_{LAD}$ of Definition 2.10 satisfies

$$\hat{\beta}_{LAD} \xrightarrow{a.s.} \beta_o \quad \text{and} \quad n^{1/2}(\hat{\beta}_{LAD} - \beta_o) \xrightarrow{d} N(0,\Sigma)$$

for some matrix Σ.

proof: Let $Z = (X,Y)$, $\rho(Z,\beta) = |Y - f(X,\beta)|$, and $\gamma(\beta) = E\rho(Z,\beta)$. Then $\hat{\beta}_{LAD}$ minimizes

$$(1/n) \sum_{i=1}^{n} \rho(Z_i,\beta)$$

and is therefore an M-Estimate. The proof of consistency will follow from an application of Huber's (1967) Theorem 1. An assumption that is critical for the result is that $\beta \neq \beta_o$ implies $\gamma(\beta) > \gamma(\beta_o)$. This can be verified as follows: Denote by $h^{-1}(x,\lambda)$ the inverse of the function $h(y,\lambda)$ in the TBS model. Then

$$Y = Y(X,\epsilon) = h^{-1}(h(f(X,\beta_o),\lambda_o) + \sigma_o\epsilon,\ \lambda_o)$$

and the conditional median of Y given X is $f(X,\beta_o)$, since

$$\begin{aligned}\tfrac{1}{2} = P\{ Y \leq f(X,\beta_o)\} &= P\{ h(f(X,\beta_o),\lambda_o) + \sigma_o\epsilon \leq h(f(X,\beta_o),\lambda_o)\} \\ &= P\{ \epsilon \leq 0\}.\end{aligned}$$

Due to independence between ϵ and X, we can write

$$E| Y - f(X,\beta)| = E_X E_\epsilon | Y - f(X,\beta)|.$$

Now, $E_\epsilon | Y - f(X,\beta)|$ for each fixed X is minimized uniquely when $f(X,\beta) = f(X,\beta_o)$, thus if $\beta \neq \beta_o$, there is a set of positive X-measure such that $f(X,\beta) \neq f(X,\beta_o)$ and hence

$$E_\epsilon | Y - f(X,\beta)| > E_\epsilon | Y - f(X,\beta_o)|$$

on this set. It follows that $\gamma(\beta) > \gamma(\beta_o)$.

The remaining assumptions of Huber's Theorem 1 are verifiable because of the compactness of Θ, the boundedness of the support of X, and the continuity of the function $\rho(Z,\beta)$. This proves consistency of $\hat{\beta}_{LAD}$ under the assumptions we have adopted.

For asymptotic normality, we note that $\hat{\beta}_{LAD}$ satisfies

$$(1/n) \sum_{i=1}^{n} \Psi(Z_i,\hat{\beta}_{LAD}) = 0$$

where

$$\Psi(Z,\beta) = \begin{cases} -(\partial/\partial\beta)\ f(X,\beta) & \text{if } Y > f(X,\beta) \\ (\partial/\partial\beta)\ f(X,\beta) & \text{if } Y < f(X,\beta). \\ 0 & \text{if } Y = f(X,\beta) \end{cases}$$

Asymptotic normality will follow from an application of Huber's (1967) Theorem 3, as $E\Psi(Z,\beta)$ is differentiable in β.

■

The next result will make it possible to estimate λ consistently in the case that the regression parameter β_o is known. In turn, this capability will be a crucial fact in Theorem 2.13. We will consider the special case obtained when $h(y,\lambda)$ is the Box-Cox modified power transform. Other cases would have to be studied individually.

Lemma 2.12: In the TBS model (2.1) with the Box-Cox modified power transform

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

and assume that the value β_o is known. Define

$$r(\lambda) = Y^{(\lambda)} - f^{(\lambda)}(X,\beta_o).$$

Let the error ϵ possess a symmetric distribution $F_{\epsilon}(\cdot)$ with a symmetric density $f_{\epsilon}(\cdot)$, let F_X be the distribution of X, and assume X and ϵ are independent. Then the equation

$$E\{r^3(\lambda)\} = 0$$

has a unique solution at $\lambda = \lambda_o$.

proof: Suppose $\lambda_o > 0$. The proof for the case $\lambda_o \leq 0$ is similar. Since ϵ is symmetric and the TBS model holds, the equation

$$(2.10) \qquad Er^3(\lambda) = E_X E_{\epsilon}\{Y^{(\lambda)} - f^{(\lambda)}(X,\beta_o)\}^3 = 0$$

has a root at $\lambda = \lambda_o$. We wish to show that it is unique. Then it is sufficient to prove that for each fixed X,

$$(2.11) \qquad E_{\epsilon}\{Y^{(\lambda)} - f^{(\lambda)}(X,\beta_o)\}^3 \begin{cases} > 0 & \text{if } \lambda > \lambda_o \\ < 0 & \text{if } \lambda < \lambda_o \end{cases}.$$

For the modified power transform,

$$(2.12) \qquad Y^{(\lambda)} - f^{(\lambda)}(X,\beta_o) = [Y^{\lambda} - f^{\lambda}(X,\beta_o)] / \lambda$$

$$= [Y^{\lambda_o(\lambda/\lambda_o)} - f^{\lambda_o(\lambda/\lambda_o)}(X,\beta_o)] \,/\, \lambda$$

$$= [\{f^{\lambda_o}(X,\beta_o) + \lambda\, \epsilon\}^{\lambda/\lambda_o} - \{f^{\lambda_o}(X,\beta_o)\}^{\lambda/\lambda_o}] \,/\, \lambda.$$

Denote $f(X,\beta_o)$ by f_o. Using (2.12),

$$E_\epsilon r^3(\lambda) = (1/\lambda^3)\, E_\epsilon\{\, [f_o + \lambda_o\epsilon]^{\lambda/\lambda_o} - [f_o]^{\lambda/\lambda_o} \,\}^3$$

$$= (1/\lambda^3)[\int_0^\infty \{\, [f_o + \lambda_o u]^{\lambda/\lambda_o} - [f_o]^{\lambda/\lambda_o} \,\}^3 f_\epsilon(u)du$$

$$+ \int_{-\infty}^0 \{\, [f_o + \lambda_o u]^{\lambda/\lambda_o} - [f_o]^{\lambda/\lambda_o} \,\}^3 f_\epsilon(u)du \,].$$

Now, because of the symmetry of $f_\epsilon(\cdot)$, the change of variable $w = -u$ yields that the second integral above is equal to

$$\int_{-\infty}^0 \{\, [f_o - \lambda_o w]^{\lambda/\lambda_o} - [f_o]^{\lambda/\lambda_o} \,\}^3 f(w)dw$$

and thus

$$E_\epsilon r^3(\lambda) = \int_0^\infty \Big\{ [\, \{\, (f_o + \lambda_o u)^{\lambda/\lambda_o} - f_o^{\lambda/\lambda_o}\} \,/\, \lambda \,]^3 - [\, \{\, f_o^{\lambda/\lambda_o} - (f_o - \lambda_o u)^{\lambda/\lambda_o} \,\} \,/\, \lambda \,]^3 \Big\}\, f_\epsilon(u)\, du. \qquad (2.13)$$

Consider the function $\varphi(u) = (u^{\lambda/\lambda_o})/\lambda$, defined for $u > 0$. We find that

$$\varphi''(u) = (1/\lambda_o)(\lambda/\lambda_o - 1)\, u^{\lambda/\lambda_o - 2}.$$

We consider two cases.

<u>Case 1.</u> $\lambda > \lambda_o$.

In this case, since $\varphi''(u) > 0$, the function $\varphi(u)$ is strictly convex, and thus $\varphi(u_3) - \varphi(u_2) > \varphi(u_2) - \varphi(u_1)$ if

$$(2.14) \qquad u_3 - u_2 = u_2 - u_1 > 0.$$

Therefore, the integrand in (2.13) is positive for all X, so that

$$E_\epsilon[\, Y^{(\lambda)} - f^{(\lambda)}(X,\beta_o)]^3 > 0 \qquad \text{for all } X.$$

Case 2. $\lambda < \lambda_o$.

In this case the function $\varphi(u)$ is strictly concave, and thus $\varphi(u_3) - \varphi(u_2) < \varphi(u_2) - \varphi(u_1)$ under condition (2.14), the integrand in (2.13) is negative, and hence

$$E_\epsilon[\, Y^{(\lambda)} - f^{(\lambda)}(X,\beta_o)]^3 < 0 \qquad \text{for all } X.$$

Hence, (2.11) holds.

■

In Theorem 2.13 to follow, we will use a one-step algorithm in order to approximate the solution of a system of non-linear equations. We make reference to the Newton-Raphson procedure, which can be summarized as follows.

Let G be a function from $\mathbb{R}^p$ to $\mathbb{R}^p$ and assume a root to the equation $G(x) = 0$ exists at x_o. Let $x^{(0)}$ be an approximation to x_o, and define

$$x^{(1)} = x^{(0)} - [\, \frac{d}{dx} G(x) \Big|_{x=x^{(0)}}]^{-1} G(x^{(0)}).$$

This $x^{(1)}$ is called the one-step *Newton-Raphson* improvement towards the solution x_o. In the numerical analysis literature (e.g. Dahlquist, Björck and Anderson (1974), section 6.9.2), it is of common knowledge that this method is of "second order", that is, for each

$d > 0$ there is a constant $C = C(d)$ such that

$$\| x^{(0)} - x_o \| \leq d \text{ implies } \| x^{(1)} - x_o \| \leq C \| x^{(0)} - x_o \|^2.$$

Theorem 2.13 provides the method for obtaining an estimate that has the same asymptotic properties as the Ruppert-Aldershof transformation to symmetry. We are in the need for utilizing this procedure because solving the system (2.2) does not guarantee consistency. In the general case, the equation $G(\theta) = E\Psi(\theta, Z) = 0$ may have solutions such that $\theta \neq \theta_o$.

<u>Theorem 2.13</u>: Consider observations $\mathfrak{X}_n = ((Y_1, X_1), \ldots, (Y_n, X_n))$, according to the model (2.1) with the Box-Cox modified power transform, assume that the errors are symmetric, and the assumptions of Theorem 2.8. Let $\hat{\theta}_n = (\hat{\beta}_n, \hat{\lambda}_n)$ be the consistent solution to the Ruppert-Aldershof equations (2.2) which exists with probability one, by Theorem 2.8.

We recall the notation

$$r(\beta, \lambda) = Y^{(\lambda)} - f^{(\lambda)}(X, \beta),$$

$$r_i(\beta, \lambda) = Y_i^{(\lambda)} - f^{(\lambda)}(X_i, \beta), \ \theta = (\beta, \lambda), \text{ and}$$

$$\theta_o = (\beta_o, \lambda_o).$$

1. Define $\hat{\beta}_{LAD}$ by

$$\sum_{i=1}^{n} | Y_i - f(X_i, \hat{\beta}_{LAD}) | = \inf_{\beta} \sum_{i=1}^{n} | Y_i - f(X_i, \beta) |. \tag{2.15}$$

2. Let $\hat{\lambda}_{LAD}$ be the solution to

$$\sum_{i=1}^{n} r^3(\hat{\beta}_{LAD}, \lambda) = 0 \tag{2.16}$$

as a function of λ, and set

$$\hat{\theta}_{LAD} = (\hat{\beta}_{LAD}, \hat{\lambda}_{LAD}).$$

3. Let $\hat{\theta}_n^* = (\hat{\beta}_n^*, \hat{\lambda}_n^*)$ be a 1-step Newton-Raphson modification towards the solution to the Ruppert-Aldershof equations with starting value $\hat{\theta}_{LAD}$, that is, let

$$\hat{\theta}_n^* = \hat{\theta}_{LAD} - [\dot{\psi}^{(n)}(\hat{\theta}_{LAD}, \mathcal{X}_n)]^{-1} \psi^{(n)}(\hat{\theta}_{LAD}, \mathcal{X}_n). \tag{2.17}$$

Then, $\hat{\theta}_n^* \xrightarrow{a.s.} \theta_o$ and furthermore,

$$n^{1/2}(\hat{\theta}_n^* - \theta_o) \text{ and } n^{1/2}(\hat{\theta}_n - \theta_o)$$

have the same asymptotic distribution.

proof: The first thing to establish is that $\hat{\lambda}_{LAD} \xrightarrow{a.s.} \lambda_o$ as n tends to infinity. For this it will be useful to show that the function $r(\beta,\lambda)$ is equicontinuous in (β,λ) with respect to (Y,X). Equicontinuity was defined in section 2.1.

Since the function $r(\beta,\lambda) = Y^\lambda - f^\lambda(X,\beta)$ is continuously differentiable in both λ and β, and the parameter space is assumed to be compact, then we can find continuous functions $K_1(X,Y)$, $K_2(X,Y)$ such that

$$\| r(\beta,\lambda) - r(\beta',\lambda')\| \le K_1(X,Y) \| \beta - \beta'\| + K_2(X,Y) \| \lambda - \lambda'\|.$$

Since the support of both Y and X is bounded, then K_1 and K_2 can be bounded by some constant K and equicontinuity follows from the inequality that results:

$$(2.18) \qquad \sup_{X,Y} \| r(\beta,\lambda) - r(\beta',\lambda')\| \le K \| \beta - \beta'\| + K \| \lambda - \lambda'\|.$$

By Lemma 2.3, since the observations are i.i.d. and the parameter space has been assumed to be compact, (2.18) implies

$$(2.19) \qquad \sup_{\theta} \left| (1/n) \sum_{i=1}^{n} r_i^3(\beta,\lambda) - Er^3(\beta,\lambda) \right| \xrightarrow{a.s.} 0.$$

From Lemma 2.10 we know that $\hat{\beta}_{LAD} \xrightarrow{a.s.} \beta_o$. Therefore, because of uniform continuity of r_i in (β,λ), we obtain

$$(2.20) \qquad \sup_{\lambda} \left| (1/n) \sum_{i=1}^{n} r_i^3(\hat{\beta}_{LAD},\lambda) - (1/n) \sum_{i=1}^{n} r_i^3(\beta_o,\lambda) \right| \xrightarrow{a.s.} 0.$$

Now, since

$$\sup_{\lambda} \left| (1/n) \sum_{i=1}^{n} r_i^3(\hat{\beta}_{LAD},\lambda) - Er^3(\beta_o,\lambda) \right|$$

$$\le \sup_{\lambda} \left| (1/n) \sum_{i=1}^{n} r_i^3(\hat{\beta}_{LAD},\lambda) - (1/n) \sum_{i=1}^{n} r_i^3(\beta_o,\lambda) \right|$$

$$+ \sup_{\lambda} \left| (1/n) \sum_{i=1}^{n} r_i^3(\beta_o,\lambda) - Er^3(\beta_o,\lambda) \right|$$

then (2.19) and (2.20) imply

$$(2.21) \qquad \sup_{\lambda} \left| (1/n) \sum_{i=1}^{n} r_i^3(\hat{\beta}_{LAD},\lambda) - Er^3(\beta_o,\lambda) \right| \xrightarrow{a.s.} 0.$$

By Lemma 2.12, the unique root to $Er^3(\beta_o,\lambda) = 0$ is $\lambda = \lambda_o$. Consistency of $\hat{\lambda}_{LAD}$ then follows from uniform convergence (2.21), by using Lemma 2.4.

An application of Huber's M-Estimation theory will establish that $(\hat{\beta}_{LAD},\hat{\lambda}_{LAD})$ is asymptotically normal, but for the proof of the current theorem we will only need the fact that

(2.22) $$\| \hat{\theta}_{LAD} - \theta_o \| = o_p(n^{-1/4}).$$

The proof of the theorem now follows. The notation is that defined by Notation 2.1. Assuming the bounded supports, since Ψ is continuous, we have uniform continuity of $\Psi(\theta, Z)$ in (θ, Z) and since $\hat{\theta}_{LAD} \xrightarrow{a.s.} \theta_o$ we may conclude

(2.23) $$\Psi^{(n)}(\hat{\theta}_{LAD}, \mathcal{X}_n) = (1/n) \sum_{i=1}^{n} \Psi(\hat{\theta}_{LAD}, Z_i) \xrightarrow{a.s.} E\Psi(\theta_o, Z) = 0.$$

That $\hat{\theta}^*_n \xrightarrow{a.s.} \theta_o$, follows from (2.17) and (2.23).

To prove asymptotic normality, we begin by noting that we have assumed enough derivatives so that for $1 \le j \le p+2$ almost surely:

$$\Psi_j(\hat{\theta}_{LAD}, Z_i) = \Psi_j(\theta_o, Z_i) + \sum_{k=1}^{p+2} \frac{\partial}{\partial\theta_k} \Psi_j(\theta_o, Z_i)(\hat{\theta}^{(k)}_{LAD} - \theta_o{}^{(k)})$$

$$+ (1/2) \sum_{k=1}^{p+2} \sum_{k'=1}^{p+2} \frac{\partial^2}{\partial\theta_k \partial\theta_{k'}} \Psi_j(\overline{\theta}, Z_i)\ (\hat{\theta}^{(k)}_{LAD} - \theta_o{}^{(k)})\ (\hat{\theta}^{(k')}_{LAD} - \theta_o{}^{(k)})$$

for some $\overline{\theta} = \overline{\theta}(i,j)$ on the line between $\hat{\theta}_{LAD}$ and θ_o. There is a constant A such that the third term is bounded by $A(p+2) \| \hat{\theta}_{LAD} - \theta_o \|^2$ for all j, so the third term is $o_p(n^{-1/2})$.

Next, for $1 \le j \le p+2$,

(2.24) $$\Psi_j^{(n)}(\hat{\theta}_{LAD}, \mathcal{X}_n) =: (1/n) \sum_{i=1}^{n} \Psi_j(\hat{\theta}_{LAD}, Z_i)$$

$$= (1/n) \sum_{i=1}^{n} \Psi_j(\theta_o, Z_i) +$$

$$+ (1/n) \sum_{i=1}^{n} \sum_{k=1}^{p+2} \frac{\partial}{\partial\theta_k} \Psi_j(\theta_o, Z_i)(\hat{\theta}^{(k)}_{LAD} - \theta_o{}^{(k)})$$

$$+ (1/n) \sum_{i=1}^{n} \{o_p(n^{-1/2}) \text{ terms}\}$$

that is,

(2.25) $$\psi^{(n)}(\hat{\theta}_{LAD},\mathcal{X}_n) =$$

$$\psi^{(n)}(\theta_o,\mathcal{X}_n) + \dot{\psi}^{(n)}(\theta_o,\mathcal{X}_n)\ (\hat{\theta}_{LAD} - \theta_o) + o_p(n^{-1/2}).$$

Using (2.25) we find

(2.26) $$n^{1/2}(\hat{\theta}_n^* - \theta_o) = n^{1/2}(\hat{\theta}_{LAD} - \theta_o) + n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_{LAD})$$

$$= n^{1/2}(\hat{\theta}_{LAD} - \theta_o) - n^{1/2}[\dot{\psi}^{(n)}(\hat{\theta}_{LAD},\mathcal{X}_n)]^{-1}\ \psi^{(n)}(\hat{\theta}_{LAD},\mathcal{X}_n)$$

$$= n^{1/2}(\hat{\theta}_{LAD} - \theta_o) - n^{1/2}[\dot{\psi}^{(n)}(\hat{\theta}_{LAD},\mathcal{X}_n)]^{-1}\times$$

$$\{\ \psi^{(n)}(\theta_o,\mathcal{X}_n) + \dot{\psi}^{(n)}(\theta_o,\mathcal{X}_n)(\hat{\theta}_{LAD} - \theta_o) + o_p(n^{-1/2})\ \}$$

$$= n^{1/2}(\hat{\theta}_{LAD} - \theta_o) - n^{1/2}\{\ [\dot{\psi}^{(n)}(\theta_o,\mathcal{X}_n)]^{-1} + O_p(\|\ \hat{\theta}_{LAD} - \theta_o\ \|)\ \}\ \times$$

$$\{\ \psi^{(n)}(\theta_o,\mathcal{X}_n) + \dot{\psi}^{(n)}(\theta_o,\mathcal{X}_n)(\hat{\theta}_{LAD} - \theta_o) + o_p(n^{-1/2})\ \}$$

$$= \ -n^{1/2}[\dot{\psi}^{(n)}(\hat{\theta}_o,\mathcal{X}_n)]^{-1}\ \psi^{(n)}(\theta_o,\mathcal{X}_n) + o_p(1).$$

By the Central Limit Theorem,

(2.27) $$-n^{1/2}\psi^{(n)}(\theta_o,\mathcal{X}_n) \xrightarrow{d} N(0,B)$$

where $B = COV(\Psi(\theta_o,Z))$ and by the Strong Law of Large Numbers,

(2.28) $$\dot{\psi}^{(n)}(\theta_o,\mathcal{X}_n) =: (1/n)\sum_{i=1}^{n}\dot{\Psi}(\theta_o,Z_i) \xrightarrow{a.s.} A$$

where $A = E\dot{\Psi}(\theta_o,Z)$.

Equations (2.26), (2.27), and (2.28) imply

$$n^{1/2}(\hat{\theta}_n^* - \theta_0) \xrightarrow{d} N(0,\ A^{-1}BA^{-T}).$$

This distribution is precisely the limit distribution which was obtained in the conclusion of Theorem 2.9, and thus the proof is complete.

∎

CHAPTER III

CONSISTENCY OF A MINIMUM DISTANCE ESTIMATOR

3.0 Introduction

In this chapter we shall describe a new estimator for the transformation parameter in the transform-both-sides regression model, and investigate consistency. We discussed the TBS model in detail in section 1.3, where the definition was given by

$$h(Y,\lambda) = h(f(X,\beta),\lambda) + \sigma\epsilon. \tag{3.1}$$

In Chapter I we described several methods that researchers have proposed to estimate transformations, and we particularly stressed those techniques whose favorable properties were a consequence of the symmetry of the errors. Shortcomings of maximum likelihood are avoided in this way. The methods of Taylor (1985) or the Ruppert-Aldershof estimator studied in Chapter II are examples of this approach.

We assume throughout this chapter that the error ϵ does indeed possess a symmetric distribution $F_\epsilon(x)$. We will propose a method for estimating the parameter λ, which is based on a certain measure of departure of a probability distribution from symmetry. In section 3.1, we apply this criterion to the empirical distribution $F_n(x,\lambda,\beta)$ of the residuals, and then select as an estimate of λ a value which minimizes the criterion. Based on symmetry of the error distribution, the method will produce consistent estimates of transformation parameters in two important cases where other estimators such as maximum likelihood fail to converge: (i) The modified power transformations, and (ii) their extension to *shifted* power transformations. An extensive discussion of these examples is contained in section 3.7; in this section, we only point out some preliminary background.

To grasp the origin of the new estimate, we begin by letting $F_n(x)$ denote the empirical distribution function of a random sample of size n from a distribution F(x). We then define

$$Q_n(x) = F_n(x) + F_n(-x) - 1$$

and

$$V_n^2 = \int \{ Q_n(x) \}^2 \, dx.$$

If $F_n(x)$ is not symmetric, the number V_n^2 assesses the degree of departure from symmetry. Orlov (1972) and Rothman and Woodroofe (1972) suggested the statistic V_n^2 for testing the null hypothesis that F is symmetric; Koziol (1980) further elaborates upon the same statistic. Test statistics of this form are known in the literature as being of the *Cramer-von Mises-type*. Other methods based on $Q_n(x)$ for testing symmetry are also possible, for example, Butler (1969) considered

$$\sup_{x \leq 0} Q_n(x).$$

The notion of Cramér-von Mises-type statistics can be used for estimation purposes as well. Let $\Gamma = \{ F_\theta : \theta \in \Theta \}$ be a parameterized family of distributions, let $X_1, \ldots, X_n$ be a random sample from a distribution G, and let G_n be the empirical distribution function of the sample. Minimization over the parameter space Θ of the quantity

$$\mathrm{cvm}(G_n, F_\theta) = \int \{ G_n(x) - F_\theta(x) \}^2 w_\theta(x)\, dx$$

yields the so-called *Minimum Distance Estimator* of θ *(using the Cramer-von Mises discrepancy)*. If $G = F_{\theta_0}$ for some $\theta_0 \in \Theta$, then the quantity which is being estimated is clearly meaningful. Otherwise, we have to interpret the estimate as selecting a member of Γ which is close to G in the sense of the quantity $\mathrm{cvm}(G, F_\theta)$. For work on the minimum distance estimate in the problem of estimating location, see Boos (1981), and Parr and De Wet (1981), for example.

In section 3.1, we propose an estimate of the transformation parameter which is based on a discrepancy that is of the Cramér-von Mises type. Sections 3.2 through 3.5 deal with material that leads to the proof of consistency. In section 3.6 we prove consistency for the estimator of the transformation in selected families of transformations, including the Box-Cox modified power transformation and the modified power transformation with shift.

3.1 Proposed Estimator

We consider the situation of observing independent realizations $Z_i = (X_i, Y_i)$ for $1 \leq i \leq n$, that follow the regression model (3.1) where the errors ϵ_i have distribution $F_\epsilon(\cdot)$. The primary objective is to estimate the parameter $\theta = (\lambda, \beta, \sigma)$; in this section we propose an alternate method for achieving this purpose.

We make the following general assumptions with regard to the transform-both-sides model (3.1):

Assumptions 3.1:

A1. The parameter (β, λ, σ) lies in a compact set Θ.

A2. There exists $(\beta_0, \lambda_0, \sigma_0) \in \Theta$ such that (3.1) holds.

A3. The pairs (X_i, ϵ_i) $i=1, \ldots, n$ are independent.

A4. X is independent of ϵ.

A5. The random variables X and ϵ have compact supports $\mathfrak{X}$ and $\mathcal{E}$, respectively.

A6. The distribution $F_\epsilon(x)$ is symmetric.

A7. The function $f(x, \beta)$ is continuously differentiable with respect to β for each x.

A8. The function $h(y, \lambda)$ is continuously differentiable with respect to λ for each fixed y.

■

We shall first allow for a brief heuristic construction of the alternate estimate before proceeding on to the formal definitions. The idea is to utilize information contained in the residuals to discern if their distribution is symmetric or not. The criterion which we employ for this bears a resemblance with that of the Cramér-von Mises type, and the estimate is picked by minimization. We therefore adopt the qualifier "Cramer-von-Mises Minimum Distance" to designate the new estimate.

Putting $Z_i = (X_i, Y_i)$, we define the *residuals* by

$$r(\lambda, \beta, Z_i) = h(Y_i, \lambda) - h(f(X_i, \beta), \lambda) \tag{3.2}$$

and notice immediately that $r(\lambda_0, \beta_0, Z_i) = \sigma_0 \epsilon_i$, so that at the correct value of the parameter, the residuals are symmetric if the ϵ_i are symmetric. We now define the *standardized residuals* by

$$r^{*}(\theta, Z_i) = r(\lambda, \beta, Z_i) \,/\, \sigma. \tag{3.3}$$

Let $F_n(x,\theta)$ denote the empirical distribution function of the standardized residuals, that is

$$F_n(x,\theta) = (1/n) \sum_{i=1}^{n} 1(\, r(\lambda,\beta,Z_i) \leq \sigma x) \tag{3.4}$$

$$-\infty < x < \infty.$$

Suppose for heuristic purposes that the parameters β_0 and σ_0 are known, and thus $F_n(x,\theta)$ is a function of λ and x only. We could then define

$$F_n(x,\lambda) = (1/n) \sum_{i=1}^{n} 1(\, r(\lambda,\beta_0,Z_i) \leq \sigma_0 x),$$

and

$$Q_n(x,\lambda) = F_n(x,\lambda) + F_n(-x,\lambda) - 1.$$

We then define for each λ, the statistic

$$V_n^2(\lambda) = \int Q_n^2(x,\lambda)\; w(x)\; dx. \tag{3.5}$$

The quantity $V_n^2(\lambda)$ is a measure of the degree of symmetry of the empirical distribution function $F_n(x,\lambda)$. We introduce a weighting function w(x) to allow for some flexibility, as it will have an effect on the efficiency of the estimate. Later on, it will be of interest to investigate precisely what influence this weighting function has on the properties of the estimate. Without loss of generality, we may assume at this point that w(x) is symmetric because adopting any w(x) is equivalent to using $[w(x) + w(-x)]/2$. This equivalence holds because

$Q_n^2(x,\lambda)$ is symmetric in x, and thus

$$V_n^2(\lambda) = \int_0^\infty Q_n^2(x,\lambda)\ w(x)\ dx + \int_{-\infty}^0 Q_n^2(x,\lambda)\ w(x)\ dx$$

$$= \int_0^\infty Q_n^2(x,\lambda)\ w(x)\ dx + \int_0^\infty Q_n^2(x,\lambda)\ w(-x)\ dx$$

$$= \int_{-\infty}^\infty Q_n^2(x,\lambda)\ [w(x) + w(-x)]/2\ dx.$$

Since $r^*(\theta_0,Z_i) = \epsilon_i$ if the regression model truly holds for some value of the parameters, then $F_n(x,\lambda_0)$ estimates the distribution $F_\epsilon(x)$. Furthermore, if the latter distribution is symmetric, then $V_n^2(\lambda_0)$ estimates zero. It thus appears sensible to treat $V_n^2(\lambda)$ as a function of λ and to select as an estimate of λ_0 that value of λ which produces a minimum. We will, in fact, formally define the "Minimum Distance Estimator based on a Cramér-von Mises-type statistic", denoted by $\hat{\lambda}_{CVM}$, as that value of λ which minimizes $V_n^2(\lambda)$, that is

$$\inf_{\lambda\in\Lambda} V_n^2(\lambda) = V_n^2(\hat{\lambda}_{CVM}).$$

When β and σ are not known, some adjustments must be made to the preceding argument. Since the well-known optimal properties of the least-squares estimator for β in the presence of normal errors are desirable, we will employ least-squares for this purpose. We will denote by $\hat{\beta}(\lambda)$ and $\hat{\sigma}(\lambda)$ the least-squares estimators of β and σ for fixed λ, that is $\hat{\beta}(\lambda)$ satisfies

$$(1/n) \sum_{i=1}^n r^2(\lambda,\hat{\beta}(\lambda),Z_i) = \inf_\beta\ (1/n) \sum_{i=1}^n r^2(\lambda,\beta,Z_i)$$

and $\hat{\sigma}(\lambda)$ is then given by

$$\hat{\sigma}^2(\lambda) = (1/n) \sum_{i=1}^{n} r^2(\lambda, \hat{\beta}(\lambda), Z_i).$$

We make this estimation principle more precise in Definition 3.3. Some notation is convenient at this point. Suppose we observe (X_i, Y_i), $1 \le i \le n$, under the assumption that the transform-both-sides regression model (3.1) holds for a family of transformations $\{ h(y,\lambda) : \lambda \in \Lambda \}$. Throughout this chapter, we let $w(x)$ be a non-negative, fixed symmetric function.

Notation 3.2:

N1. $$Z = (X,Y), \quad Z_i = (X_i, Y_i)$$

N2. $$r(\lambda,\beta,Z) = h(Y,\lambda) - h(f(X,\beta),\lambda)$$

N3. $\hat{\beta}(\lambda)$ is the least-squares estimator for fixed λ,

that is, $\hat{\beta}(\lambda)$ minimizes $\sum_{i=1}^{n} \{r(\lambda,\beta,Z_i)\}^2$ as a function of β.

N4. $$\hat{\sigma}^2(\lambda) = (1/n) \sum_{i=1}^{n} r^2(\lambda, \hat{\beta}(\lambda), Z_i)$$

N5. $$r_i^*(\lambda) = r(\lambda, \hat{\beta}(\lambda), Z_i) \;/\; \hat{\sigma}(\lambda)$$

N6. $$F_n(x,\lambda) = (1/n) \sum_{i=1}^{n} 1(\; r_i^*(\lambda) \le x \;)$$

N7. $$Q_n(x,\lambda) = F_n(x,\lambda) + F_n(-x,\lambda) - 1$$

N8. $$V_n^2(\lambda) = \int Q_n^2(x,\lambda)\; w(x)\; dx$$

■

Definition 3.3: The *Minimum Distance Estimator* $\hat{\lambda}_{CVM}$ is defined by minimizing the quantity $V_n^2(\lambda)$ over λ:

(3.6) $$\inf_{\lambda \in \Lambda} V_n^2(\lambda) = V_n^2(\hat{\lambda}_{CVM}).$$

This also defines the following estimators of β and σ:

(3.7) $$\hat{\beta}_{CVM} = \hat{\beta}(\hat{\lambda}_{CVM})$$

(3.8) $$\hat{\sigma}^2_{CVM} = \hat{\sigma}(\hat{\lambda}_{CVM})$$

■

To study this estimator analytically, it is convenient to investigate the behavior of the estimate of λ when β_0 and σ_0 are known. In this case, the estimate $\hat{\lambda}_{CVM}$ is still defined by (3.6), but replacing $\hat{\beta}(\lambda)$ for β_0 and $\hat{\sigma}(\lambda)$ for σ_0 in the definition of $r_i^*(\lambda)$.

The first property we study for this estimator is consistency, in section 3.6. After demonstrating in section 3.2 that the estimator belongs to a broad class of estimators, we will then prove results for this general class in sections 3.3 through 3.5. We will discuss the asymptotic distribution of the estimate in Chapter IV.

3.2 Alternate Representation of the Estimator

It is useful to observe that the minimum distance estimator defined above can be represented in an alternate form, which shall be the basis for theoretical analysis. This framework provides motivation for the class of estimators to be studied in section 3.3, and it also allows for simplification in the computation of the estimate.

Lemma 3.4: We refer to Notation 3.2 assuming that β_0 and σ_0 are known. For each $\lambda \in \Lambda$, the quantity $V_n^2(\lambda)$ can be written in the form

$$(3.9) \qquad V_n^2(\lambda) = (1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} K_\lambda(Z_i, Z_j)$$

where for each λ the function $K_\lambda(\cdot,\cdot)$ is symmetric in the sense that $K_\lambda(x,y) = K_\lambda(y,x)$ for all (x,y).

proof: Letting

$$a_i(x,\lambda) = 1(r_i^*(\lambda) \le x) + 1(r_i^*(\lambda) \le -x) - 1$$

we note that

$$\begin{aligned} Q_n^2(x,\lambda) &= \{(1/n) \sum_{i=1}^{n} 1(r_i^*(\lambda) \le x) + (1/n) \sum_{i=1}^{n} 1(r_i^*(\lambda) \le -x) - 1 \}^2 \\ &= (1/n^2)\{ \sum_{i=1}^{n} [1(r_i^*(\lambda) \le x) + 1(r_i^*(\lambda) \le -x) - 1] \}^2 \\ &= (1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} a_i(x,\lambda)\, a_j(x,\lambda). \end{aligned}$$

Therefore,

$$(3.10) \qquad V_n^2(\lambda) = (1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} \int a_i(x,\lambda)\, a_j(x,\lambda)\, w(x)\, dx.$$

If β_0 and σ_0 are both known, then putting $r_i^*(\lambda) = r(\lambda,\beta_0,Z_i) \,/\, \sigma_0$ shows that for each λ, $V_n^2(\lambda)$ is of the form described in the Lemma, by setting

$$K_\lambda(Z_i,Z_j) = \int a_i(x,\lambda)\, a_j(x,\lambda)\, w(x)\, dx \quad \text{in equation (3.10).}$$

■

We have one additional result in this section, which is of interest from a computational point of view. It enables us to compute the criterion $V_n^2(\lambda)$ without having to actually perform integration, exploiting the fact that empirical distribution functions are step functions. This is an advantage because minimization algorithms often operate by multiple evaluations of the objective function. We note that each evaluation of $V_n^2(\lambda)$ for fixed λ already involves least squares (most likely non-linear), so this simplification reduces the task.

Proposition 3.5: Assume that the weighting function $w(x)$ that appears in the definition of $V_n^2(\lambda)$ is a symmetric function, and that there exists a (necessarily odd) nondecreasing function $\psi(x)$ such that $\psi'(x) = w(x)$ for all x. Let R_i be the rank of $\psi(r_i^*(\lambda))$ and Q_i the rank of $|\psi(r_i^*(\lambda))|$. We note since Ψ is nondecreasing, that R_i is just the rank of $r_i^*(\lambda)$. Then

$$V_n^2(\lambda) = (2/n^2) \left\{ 2 \sum_{k=1}^{n} c_k |\psi(r_k^*(\lambda))| + \sum_{k=1}^{n} |\psi(r_k^*(\lambda))| \right\}$$

where

$$c_i = \begin{cases} 2R_i - 2 - n + Q_i & \text{if } r_i^*(\lambda) < 0 \\ n - 2R_i + Q_i & \text{if } r_i^*(\lambda) \geq 0. \end{cases}$$

proof: Consider the expression for $V_n^2(\lambda)$ given by Lemma 3.4. For $1 \leq i \leq n$, use the notation r_i for $r_i^*(\lambda)$, and for $1 \leq i,j \leq n$ let

$$t_{ij} = \int_{-\infty}^{\infty} a_i(x,\lambda)\, a_j(x,\lambda)\, w(x)\, dx,$$

where

$$a_i(x,\lambda) = 1(r_i \leq x) + 1(r_i \leq -x) - 1.$$

Since we have assumed the transform-both-sides model with bounded supports for X and ϵ, as well as compact parameter spaces, then the integral above over $(-\infty, \infty)$ is actually over a bounded interval, say $[L, U]$, where $L = L(\lambda, i, j)$ and $U = U(\lambda, i, j)$. Thus

$$
\begin{aligned}
t_{ij} = \int_L^U \Big[& 1(r_i \le x)\, 1(r_j \le x) + 1(r_i \le x)\, 1(r_j \le -x) - 1(r_i \le x) \\
& + 1(r_i \le -x)\, 1(r_j \le x) + 1(r_i \le -x)\, 1(r_j \le -x) - 1(r_i \le -x) \\
& - 1(r_j \le x) - 1(r_j \le -x) + 1 \Big]\, w(x)\, dx
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{r_i \vee r_j}^{U} w(x)\,dx + \int_{r_i}^{r_i \vee -r_j} w(x)\,dx - \int_{r_i}^{U} w(x)\,dx \\
&\quad + \int_{r_j}^{-r_i \vee r_j} w(x)\,dx + \int_{L}^{-r_i \wedge -r_j} w(x)\,dx - \int_{r_i}^{U} w(x)\,dx \\
&\quad - \int_{r_j}^{U} w(x)\,dx - \int_{L}^{-r_j} w(x)\,dx + \int_{L}^{U} w(x)\,dx \\
&= \psi(U) - \psi(r_i \vee r_j) + \psi(r_i \vee -r_j) - \psi(r_i) - \psi(U) + \psi(r_i) + \psi(-r_i \vee r_j) \\
&\quad - \psi(r_j) + \psi(-r_i \wedge -r_j) - \psi(L) - \psi(-r_i) + \psi(L) - \psi(U) + \psi(r_j) \\
&\quad - \psi(-r_j) + \psi(L) + \psi(U) - \psi(L) \\
&= -\psi(r_i \vee r_j) + \psi(r_i \vee -r_j) + \psi(-r_i \vee r_j) + \psi(-r_i \wedge -r_j) + \psi(r_i) + \psi(r_j) \\
&= -2\psi(r_i \vee r_j) + \psi(-r_i \vee r_j) + \psi(r_i \vee -r_j) + \psi(r_i) + \psi(r_j).
\end{aligned}
$$

Notice $t_{ij} = t_{ji}$ and $t_{ii} = 2|\psi(r_i)|$. We can show by considering cases, that

$$
t_{ij} = \begin{cases} 2(|\psi(r_i)| \wedge |\psi(r_j)|) & \text{if } r_i r_j \ge 0 \\ -2(|\psi(r_i)| \wedge |\psi(r_j)|) & \text{if } r_i r_j < 0. \end{cases}
$$

For instance, if $0 < r_i < r_j$, then from the expression for t_{ij} we obtain

$$
\begin{aligned}
t_{ij} &= -2\psi(r_j) + \psi(r_j) + \psi(r_i) + \psi(r_i) + \psi(r_j) \\
&= 2\psi(r_i) = 2|\psi(r_i)|.
\end{aligned}
$$

The remaining configurations are all similar.

Now, by Lemma 3.4,

$$V_n^2(\lambda) = (1/n^2)\{ \sum_{i=1}^{n} t_{ii} + 2 \sum_{i<j} t_{ij} \},$$

and each t_{ij} is one of either $\pm 2|\psi(r_i)|$ or $\pm 2|\psi(r_j)|$, depending on the relative location of r_i and r_j. Thus the problem reduces to counting the number of times that each term $2|\psi(r_k)|$ appears with a positive or negative sign as the double sum over $(i<j)$ progresses. Let P_k be the number of times that $2|\psi(r_k)|$ is positive, and N_k the number of times that it is negative. Using the definitions of R_k and Q_k, the following two cases can be verified.

Case A. If r_k is negative, then

$$\begin{cases} P_k = R_k - 1 \\ N_k = (n - Q_k) - (n - R_k) \end{cases}.$$

Case B. If r_k is non-negative, then

$$\begin{cases} P_k = n - R_k \\ N_k = (n - Q_k) - (n - R_k) = R_k - Q_k \end{cases}.$$

Hence the coefficient for the term $2|\psi(r_k)|$ is

$$\begin{cases} R_k - 1 - (n - Q_k) + (R_k - 1) = 2R_k - 2 - n + Q_k & \text{if } r_k < 0 \\ n - R_k - (R_k - Q_K) = n - 2R_k + Q_k & \text{if } r_k \geq 0, \end{cases}$$

as stated.

■

3.3 Generalized M-Estimators

Relation (3.9) shows that the proposed estimator $\hat{\lambda}_{CVM}$ is a special case of a definite family of estimators. Serfling (1984) termed this the class of *Generalized M-Estimators*, while studying classes of a related type. One way of characterizing the general case is through the following description:

Let m be a fixed integer and $X_1, X_2, \ldots, X_n$ random variables on a probability space $(\mathcal{X}, \mathcal{A}, P)$, and Θ a parameter space which is a subset of $\mathbb{R}^p$. Consider a function $K : \mathcal{X}^m \times \Theta \longrightarrow \mathbb{R}$, such that $K(x_1, x_2, \ldots, x_m, \theta)$ is *symmetric* in its x arguments for each fixed $\theta \in \Theta$, that is $K(x_1, \ldots, x_m, \theta) = K(x_{i_1}, \ldots, x_{i_m}, \theta)$ for all permutations $(i_1, \ldots, i_m)$ of the integers $(1, \ldots, m)$.

We define a *Generalized M-Estimator* to be a random variable $\hat{\theta}_n$ that satisfies

$$(3.11) \qquad \inf_{\theta} (1/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} K(X_{i_1}, \ldots, X_{i_m}, \theta)$$
$$= (1/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} K(X_{i_1}, \ldots, X_{i_m}, \hat{\theta}_n).$$

In the event that the function K has a derivative with respect to θ, then after putting $K'(x_1, \ldots, x_m, \theta) = \frac{\partial}{\partial\theta} K(x_1, \ldots, x_m, \theta)$ we would find that the Generalized M-Estimator satisfies

$$(3.12) \qquad (1/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} K'(X_{i_1}, \ldots, X_{i_m}, \hat{\theta}_n) = 0.$$

If we take m = 1 we obtain Huber's M-Estimator as a special case, namely, a $\hat{\theta}_n$ which is defined by

$$\inf_{\theta} (1/n) \sum_{i=1}^{n} \rho(X_i, \theta) = (1/n) \sum_{i=1}^{n} \rho(X_i, \hat{\theta}_n)$$

for some function $\rho : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}$, or alternatively as satisfying

$$(1/n) \sum_{i=1}^{n} \Psi(X_i, \hat{\theta}_n) = 0$$

where $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^p$.

In the remaining part of this chapter we establish some conditions for consistency of a generalized M-estimate. The special case of estimating the transformation using the estimator described in section 3.1 will be examined closely in section 3.6. We first state a few preliminary definitions and key results for later reference.

3.4 U-Statistics

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with distribution F. In this section, let K be a function from $\mathbb{R}^m$ to $\mathbb{R}^q$, which we call a *Kernel of Order* m. We will assume that $K(x_1, \ldots, x_m)$ is symmetric in its m arguments, that is

$$K(x_{i_1}, \ldots, x_{i_m}) = K(x_1, \ldots, x_m)$$

for all permutations $(i_1, \ldots, i_m)$ of the integers $\{1, \ldots, m\}$, and that

$$E_F\{|K(X_1, \ldots, X_m)|\} = \int \ldots \int |K(x_1, \ldots, x_m)| \, dF(x_1) \ldots dF(x_m) < \infty.$$

For a given symmetric kernel K, the associated *U-Statistic* based on a sample of size $n > m$ is defined to be

$$U_n = \frac{1}{\binom{n}{m}} \sum_C K(X_{i_1}, \ldots, X_{i_m}) \tag{3.13}$$

where the sum $\sum_C$ is over all combinations of distinct elements $\{i_1, \ldots, i_m\}$ from the set $\{1, \ldots, n\}$.

A second statistic, also based on the kernel K and on the sample of size n, is termed a *V-Statistic* if it is defined as

$$V_n = \frac{1}{n^m} \sum_{i_1=1}^{n} \ldots \sum_{i_m=1}^{n} K(X_{i_1}, \ldots, X_{i_m}). \tag{3.14}$$

Some well-known results on U-Statistics for real-valued ($q = 1$) kernels will be quoted here without proof. For a more detailed discussion of U-Statistics, see for example, chapter 5 of Serfling (1980).

We define the functions

$$K_c(x_1, \ldots, x_c) = \begin{cases} E_F\{K(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m)\} & 1 \le c \le m-1 \\ K(x_1, \ldots, x_m) & c = m \end{cases}$$

as well as the numbers

$$\gamma = E_F\{K(X_1, \ldots, X_m)\} = \int \ldots \int K(x_1, \ldots, x_m)\, dF(x_1) \ldots dF(x_m) \tag{3.15}$$

and if $E_F\{K^2(X_1, \ldots, X_m)\} < \infty$,

$$\zeta_c = VAR_F\{K_c(X_1, \ldots, X_c)\} \qquad 1 \le c \le m.$$

<u>Lemma 3.6</u>: If $E_F\{K^2(X_1, \ldots, X_m)\} < \infty$, then

$$VAR_F\{U_n\} = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c}\binom{n-m}{m-c} \zeta_c$$

and also

(i) $(m^2/n)\zeta_1 \le VAR_F\{U_n\} \le (m/n)\zeta_m$

(ii) $(n + 1)VAR_F\{U_{n+1}\} \le n\ VAR_F\{U_n\}$

(iii) $VAR_F\{U_n\} = (m^2/n)\zeta_1 + O(n^{-2})$ as $n \longrightarrow \infty$.

∎

<u>Lemma 3.7</u>: Let $r \ge 2$, and suppose that $E_F\{|K(X_1, \ldots, X_m)|^r\} < \infty$. Then

$$E|\, U_n - \gamma|^r = O(n^{-r/2}), \text{ as } n \longrightarrow \infty.$$

∎

The next lemma shows the close connection between a U-Statistic and a V-Statistic when they are based on the same symmetric kernel K of order m.

Lemma 3.8: Let r be a positive integer and suppose that for all $1 \le i_1, \ldots, i_m \le m$

$$E_F\{| K(X_{i_1}, \ldots, X_{i_m})|^r\} < \infty.$$

Then the statistics U_n and V_n defined by (3.13) and (3.14) satisfy

$$E| U_n - V_n|^r = O(n^{-r}).$$

■

Lemma 3.9: If

$$E_F\{| K(X_{i_1}, \ldots, X_{i_m})|^2\} < \infty$$

for all $1 \le i_1, \ldots, i_m \le m$, then $n^{1/2}(U_n - V_n) \xrightarrow{P} 0$ and hence $n^{1/2}(U_n - \gamma)$ and $n^{1/2}(V_n - \gamma)$ possess the same asymptotic distribution.

■

Lemma 3.10: (Strong Law of Large Numbers for U-Statistics)

If $E_F\{|K(X_1, \ldots, X_m)|\} < \infty$, then $U_n \xrightarrow{a.s.} \gamma$.

■

Lemma 3.11: (Asymptotic Normality of U-Statistics)

If $E_F\{K^2(X_1, \ldots, X_m)\} < \infty$ and $\zeta_1 > 0$, then

$$n^{1/2}(U_n - \gamma) \xrightarrow{d} N(0, m^2\zeta_1).$$

■

Lemma 3.12: If $a > 0$, and $E_F\{K^2(X_1, \ldots, X_m)\} < \infty$, then

$$P(\ |V_n| > a\ ) \leq (m/na^2)\ VAR_F\{K(X_1, \ldots, X_m)\}$$

for sufficiently large n.

■

3.5 Consistency of Generalized M-Estimators

Let us now consider a collection of real-valued kernels of order m, $\{\ K(x_1, \ldots, x_m, \theta): \theta \in \Theta\}$, such that each K is symmetric in its x arguments for each fixed θ in a parameter space Θ. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with values in $\mathcal{X}$ and having the common distribution F. A Generalized M-Estimator was defined in section 3.3 as a random variable $\hat{\theta}_n$ that minimizes the expression

$$(3.16) \qquad H_n(\theta) = \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} K(X_{i_1}, \ldots, X_{i_m}, \theta)$$

over $\theta \in \Theta$.

Strong consistency of $\hat{\theta}_n$, with convergence to some constant θ_0, is the subject of this section. The following lemma can be viewed as an extension to Lemma 2.3, to establish a form of uniform convergence which shall be important. The assumptions make use of the notion of equicontinuity which was stated in section 2.1.

Lemma 3.13: Denote

$$(3.17) \qquad H(\theta) = E_F\ \{K(X_1, \ldots, X_m, \theta)\}$$

$$= \int \cdots \int K(x_1, x_2, \ldots, x_m, \theta)\, dF(x_1) \ldots dF(x_m).$$

If we assume

(i) $E_F \mid K(X_1, \ldots, X_m, \theta) \mid < \infty$ for all $\theta \in \Theta$,

(ii) $K(x_1, \ldots, x_m, \theta)$ is equicontinuous in θ for all $x_1, \ldots, x_m$ in their domain, and

(iii) The parameter space Θ is compact,

then

$$\sup_{\theta} \mid H_n(\theta) - H(\theta) \mid \xrightarrow{a.s.} 0.$$

<u>proof</u>: Let $\epsilon > 0$ be fixed. Since K is equicontinuous in θ for $(x_1, x_2, \ldots, x_m) \in \mathfrak{X}^m$ and the set Θ is compact, there exist $\theta_1, \theta_2, \ldots, \theta_Q \in \Theta$ and open subsets $U_1, \ldots, U_Q \subset \Theta$ such that

$$\bigcup_{j=1}^{Q} U_j \supset \Theta, \ \theta_j \in U_j,$$

and for all $\theta \in U_j$, and all $X_1, \ldots, X_m$,

$$\mid K(X_1, \ldots, X_m, \theta) - K(X_1, \ldots, X_m, \theta_j) \mid < \epsilon/3.$$

For any $\theta \in U_j$, $j=1, \ldots, Q$ we have

$$\begin{aligned}
\mid H_n(\theta) - H(\theta) \mid &\le \mid H_n(\theta) - H_n(\theta_j) + H_n(\theta_j) - H(\theta_j) \mid \\
&\le (1/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \mid K(X_{i_1}, \ldots, X_{i_m}, \theta) - K(X_{i_1}, \ldots, X_{i_m}, \theta_j) \mid \\
&\qquad + \mid H_n(\theta_j) - H(\theta_j) \mid \\
&\le \epsilon/3 + \mid H_n(\theta_j) - H(\theta_j) \mid .
\end{aligned}$$

Now,

$$\sup_{\theta} \mid H_n(\theta) - H(\theta) \mid \le \sup_{\theta \in \cup U_j} \mid H_n(\theta) - H(\theta) \mid$$

$$\leq \max_{1\leq j\leq Q} \sup_{\theta\in U_j} | H_n(\theta) - H(\theta) |.$$

Since

$$\sup_{\theta\in U_j} | H_n(\theta) - H(\theta) | \leq \sup_{\theta\in U_j} | H_n(\theta) - H(\theta_j) | + \sup_{\theta\in U_j} | H(\theta_j) - H(\theta) |$$

$$\leq \epsilon/3 + | H_n(\theta_j) - H(\theta_j) | + \epsilon/3,$$

we obtain

$$\sup_{\theta} | H_n(\theta) - H(\theta) | \leq \max_{1\leq j\leq Q} \{ (2/3)\epsilon + | H_n(\theta_j) - H(\theta_j) | \}$$

$$= (2/3)\epsilon + \max_{1\leq j\leq Q} | H_n(\theta_j) - H(\theta_j) |.$$

For any $\delta > 0$, by Lemma 3.10 of U-Statistic theory, there exists an integer N such that for $j=1, \ldots, Q$

$$P(\text{ some } n > N, \ | H_n(\theta_j) - H(\theta_j) | \geq \epsilon/3 \ ) < \delta/Q.$$

Denoting the event within parentheses by E_j, we note

$$\bigcap_{j=1}^{Q} \bar{E}_j \subset \{ \text{ all } n > N, \ \max_j | H_n(\theta_j) - H(\theta_j) | < \epsilon/3 \ \}$$

$$\subset \{ \text{ all } n > N, \ \sup_{\theta} | H_n(\theta) - H(\theta) | < \epsilon \ \}.$$

By Bonferroni's inequality,

$$P(\bigcap_{j=1}^{Q} \bar{E}_j) \geq 1 - \sum_{j=1}^{Q} P(E_j) \geq 1 - \delta$$

and hence

$$P(\text{ all } n > N, \ \sup_{\theta} | H_n(\theta) - H(\theta) | < \epsilon) \geq 1 - \delta,$$

completing the proof.

■

The following theorem establishes strong consistency of a Generalized M-Estimator $\hat{\theta}_n$. We note that condition (3.18) below is a characterization of the limit. Although (3.18) may appear as manageable, it may be the case that it is actually hard to check in a given situation, particularly if θ is multi-dimensional.

<u>Theorem 3.14</u>: Let $H(\theta)$ be defined by (3.17). Assume that the conditions (i)-(iii) of Lemma 3.13 hold, and that there exists $\theta_0 \in \Theta$ such that

$$(3.18) \qquad H(\theta) > H(\theta_0) \quad \text{for all } \theta \neq \theta_0,\ \theta \in \Theta.$$

Then $\hat{\theta}_n$ defined by minimizing (3.16) satisfies $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$.

<u>proof</u>: For each $\epsilon > 0$, let $U_\epsilon = \{\, \theta : |\, \theta - \theta_0 \,| < \epsilon \,\}$.

It is sufficient to prove

$$P(\text{there exists } n_0 : \hat{\theta}_n \in U_\epsilon \text{ , all } n > n_0) = 1.$$

Since $\Theta - U_\epsilon$ is a compact set,

$$\inf_{\Theta - U_\epsilon} H(\theta)$$

is attained and by assumption is strictly greater than $H(\theta_0)$, say greater than $H(\theta_0) + 3\delta$ for some $\delta > 0$. By Lemma 3.13, with probability one there exists an n_0 such that for all $n > n_0$

$$\sup_{\theta \in \Theta} |\, H_n(\theta) - H(\theta) | < \delta$$

and thus $H(\theta) - 2\delta < H(\theta) - \delta \leq H_n(\theta) \leq H(\theta) + \delta$, for all θ, and all $n > n_0$. This implies

$$\inf_{\Theta - U_\epsilon} H_n(\theta) \geq \inf_{\Theta - U_\epsilon} H(\theta) - 2\delta \text{ and}$$

$$\inf_{\theta\in U_\epsilon} H_n(\theta) \le \inf_{\theta\in U_\epsilon} H(\theta) + \delta = H(\theta_o) + \delta,$$

and thus for all $n > n_o$

$$\inf_{\theta\in\Theta-U_\epsilon} H_n(\theta) \ge \inf_{\theta\in\Theta-U_\epsilon} H(\theta) - 2\delta > H(\theta_o) + \delta \ge \inf_{\theta\in U_\epsilon} H_n(\theta).$$

Hence with probability one, $\hat{\theta}_n \in U_\epsilon$ for all $n > n_o$.

■

3.6 A Special Case: Minimum Distance Estimation of a Transformation

We return to the study of estimation of a transformation parameter by the method outlined in section 3.1. There we stated that the quantity

$$V_n^2(\lambda) = \int \{F_n(x,\lambda) + F_n(-x,\lambda) - 1\}^2 \, w(x) \, dx \tag{3.19}$$

is minimized over λ in order to obtain the estimate denoted by $\hat{\lambda}_{CVM}$. The notation $F_n(x,\lambda)$ stands for the empirical distribution function of the standardized residuals

$$r_i^*(\lambda) = \{ h(Y_i,\lambda) - h(f(X_i,\hat{\beta}(\lambda)),\lambda) \} / \hat{\sigma}(\lambda)$$

$$1 \le i \le n.$$

We showed in section 3.2 that an alternate expression for $V_n^2(\lambda)$ is

$$V_n^2(\lambda) = (1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} K(Z_i,Z_j,\lambda)$$

where $Z_i = (X_i,Y_i)$, $1 \le i \le n$, are independent and identically distributed observations, and furthermore, that

$$K(Z_i,Z_j,\lambda) = \int a_i(x,\lambda)\, a_j(x,\lambda)\, w(x)\, dx$$

where $a_i(x,\lambda) = 1(r_i^*(\lambda) \le x) + 1(r_i^*(\lambda) \le -x) - 1$. If the parameters β_o

and σ_0 are known, replacing $\hat{\beta}(\lambda)$ for β_0 and $\hat{\sigma}(\lambda)$ for σ_0 shows directly that $\hat{\lambda}_{CVM}$ is a generalized M-estimator with a kernel K of order m = 2. The topic in the current section is consistency of such an estimate of the transformation, under the structure of the transform-both-sides regression model.

To establish consistency using the result from Theorem 3.14, a relevant quantity to be examined closely is

$$V^2(\lambda) =: E\,[K(Z_1, Z_2, \lambda)]. \tag{3.20}$$

For consistency to be plausible a required condition is that there exists a value λ_0 such that $V^2(\lambda) > V^2(\lambda_0)$ for all $\lambda \neq \lambda_0$. We shall see in one of the results to follow, that this requisite is a consequence of assuming the existence of a parameter value such that the transform-both-sides model holds, but it will have to be checked for each particular family of transformations that is considered. With respect to other assumptions of Theorem 3.14, the fact that λ is restricted to lie in a compact set has already been assumed. The other crucial assumption in Theorem 3.14 is equicontinuity in λ of the function $K(Z_1, Z_2, \lambda)$ for Z_1 and Z_2 in their domain; in Lemma 3.15 we check that equicontinuity holds.

Lemma 3.15: Let $h(y, \lambda)$, the family of transformations in the TBS model (3.1), be such that it is continuously differentiable in λ for each fixed y. (This is A8 of Assumptions 3.1)

Then $K(Z_1, Z_2, \lambda)$ is equicontinuous in λ for $Z_1, Z_2 \in \mathcal{E} \times \mathfrak{X}$.

proof: We can show (see proof of Proposition 3.5), that $K(Z_1,Z_2,\lambda) = g(r_1^*(\lambda),\ r_2^*(\lambda))$, where the function g is defined by

$$g(u,v) = \int_{uVv}^{U} w(x)\,dx + \int_{u}^{uV-v} w(x)\,dx - \int_{u}^{U} w(x)\,dx$$
$$+ \int_{v}^{-uVv} w(x)\,dx + \int_{L}^{-u\Lambda-v} w(x)\,dx - \int_{u}^{U} w(x)\,dx$$
$$- \int_{u}^{U} w(x)\,dx - \int_{L}^{-v} w(x)\,dx + \int_{L}^{U} w(x)\,dx.$$

This g(u,v) is a continuous function of (u,v), and for (u,v) on a compact support, it is thus uniformly continuous. Then for given $\epsilon > 0$ there exists δ^* such that if $|u - u'| + |v - v'| < \delta^*$, then $|g(u,v) - g(u',v')| < \epsilon$ for all (u,u',v,v'). It will suffice to show that $r^*(\lambda)$ is equicontinuous in λ for $Z \in \mathcal{E}\times\mathcal{X}$, for then there is δ such that if $\|\lambda - \lambda'\| < \delta$, then

$$|r_1^*(\lambda) - r_1^*(\lambda')| < \delta^*/2 \text{ and } |r_2^*(\lambda) - r_2^*(\lambda')| < \delta^*/2,$$

and thus $\|\lambda - \lambda'\| < \delta$ would imply

$$|g(r_1^*(\lambda),r_2^*(\lambda)) - g(r_1^*(\lambda'),r_2^*(\lambda'))| < \epsilon,$$

which completes the proof of the lemma.

It thus remains to show that $r(\lambda)$ is equicontinuous in λ. By the form of the model and by definition of a residual,

$$r(\lambda) = h\{h^{-1}[h(f(X,\beta_o),\lambda_o)+\sigma_o\epsilon,\lambda_o],\lambda\} - h(f(X,\beta_o),\lambda).$$

Suppose λ is p-dimensional, $\lambda = (\lambda_1,\dots,\lambda_p)$. Since the derivatives are assumed continuous on bounded supports, there exists a constant C such that $|(\partial/\partial\lambda_j)\ r(\lambda)| < C$ for all $j=1,\dots,p$. By the Mean Value Theorem for functions from $\mathbb{R}^p$ to $\mathbb{R}$, for some $t \in [0,1]$,

$$r(\lambda) - r(\lambda') = \sum_{j=1}^{p} \frac{\partial}{\partial\lambda_j} r(\lambda^*)(\lambda_j - \lambda_j')$$

with $\lambda^* = t\lambda + (1-t)\lambda'$ and thus

$$| r(\lambda) - r(\lambda') | \le C \sum_{j=1}^{p} | \lambda_j - \lambda_j' |.$$

If $\| \lambda - \lambda' \| \le \epsilon/(Cp)$, then $| r(\lambda) - r(\lambda') | \le \epsilon$, showing that $r(\lambda)$ is equicontinuous in λ, as required.

∎

We next proceed to study special cases of parameterized families of transformations $\{ h(y,\lambda) : \lambda \in \Lambda \}$, and seek to verify assumption (3.18) in the case that β_0 and σ_0 are known. Recall that we have assumed that the transform-both-sides model (3.1) holds for some parameters on a compact subset, the supports of $\mathcal{X}$ and $\mathcal{E}$ are compact sets, and the observations (X_i, Y_i) $1 \le i \le n$ are independent.

We also recall from equation (3.20) that

$$V^2(\lambda) = E \int a_1(x,\lambda)\, a_2(x,\lambda)\, w(x)\, dx,$$

where

$$\begin{aligned} a_i(x,\lambda) &= 1(r_i^*(\lambda) \le x) + 1(r_i^*(\lambda) \le -x) - 1 \\ &= 1(r_i^*(\lambda) \le x) - 1(r_i^*(\lambda) > -x) \end{aligned}$$

and $r_i^*(\lambda)$ depends on the transformation family through (3.2) and (3.3). Given that $|r_i^*(\lambda)|$ is bounded for all (λ, X, ϵ), we can write

$$V^2(\lambda) = \int E\,\{ a_1(x,\lambda)\, a_2(x,\lambda)\, w(x) \}\, dx$$

$$= \int \{ P(r_1^*(\lambda) \le x) - P(r_2^*(\lambda) > -x) \}^2\, w(x)\, dx$$

since the $r_i(\lambda)$ are independent and identically distributed.

Define $F(x,\lambda) = P(r_1^*(\lambda) \le x) = P(r(\lambda) \le \sigma_o x)$ so that

$$V^2(\lambda) = \int \{ F(x,\lambda) + F(-x,\lambda) - 1 \}^2 w(x)\, dx. \tag{3.21}$$

Now let $f_o = f(X,\beta_o)$. We notice that if the TBS model with a family $\{ h(y,\lambda)\}$ holds, then for fixed X,

$$\begin{aligned} F(x,\lambda) &= P(r(\lambda) \le \sigma_o x) = P(h(Y,\lambda) - h(f_o,\lambda) \le \sigma_o x) \\ &= P(h(h^{-1}(h(f_o,\lambda_o)+\sigma_o\epsilon,\lambda_o),\lambda) - h(f_o,\lambda) \le \sigma_o x) \\ &= P(h^{-1}(h(f_o,\lambda_o)+\sigma_o\epsilon),\lambda_o) \le h^{-1}(\sigma_o x + h(f_o,\lambda),\lambda)) \\ &= P(\epsilon \le \{ h(h^{-1}(\sigma_o x + h(f_o,\lambda),\lambda),\lambda_o) - h(f_o,\lambda_o) \}/\sigma_o). \end{aligned}$$

Denote the right-hand side of the last inequality by $\varphi_x(\lambda)$. Since X and ϵ are independent, we can write

$$P(\epsilon \le \varphi_x(\lambda)) = E_X E_\epsilon 1(\epsilon \le \varphi_x(\lambda)) = E_X F_\epsilon(\varphi_x(\lambda))$$

where F_ϵ is the distribution function of ϵ.

We thus have

$$F(x,\lambda) = E_X F_\epsilon(\varphi_x(\lambda)) \tag{3.22}$$

with

$$\varphi_x(\lambda) = \{ h(h^{-1}(\sigma_o x + h(f_o,\lambda),\lambda),\lambda_o) - h(f_o,\lambda_o) \}/\sigma_o. \tag{3.23}$$

Notice of course, $F(x,\lambda_o) = F_\epsilon(x)$, and furthermore, if F_ϵ is symmetric then

$$F(x,\lambda) + F(-x,\lambda) - 1 = E_X\{ F_\epsilon(\varphi_x(\lambda)) - F_\epsilon(-\varphi_{-x}(\lambda)) \}. \tag{3.24}$$

We introduce one last piece of notation to use in proofs of lemmas that follow. Let

(3.25) $$\Psi(X,x,\lambda) = F_{\epsilon}(\varphi_x(\lambda)) - F_{\epsilon}(-\varphi_{-x}(\lambda))$$

so that (3.24), (3.25), and (3.21) imply

(3.26) $$V^2(\lambda) = \int \{ E_X\Psi(X,x,\lambda) \}^2 w(x)\, dx.$$

In the lemmas that ensue, the goal is to verify that there exists a value λ_0 such that $V^2(\lambda) > V^2(\lambda_0)$ for all $\lambda \neq \lambda_0$. We shall prove this assuming that the parameters β_0 and σ_0 are known; in this case we substitute σ_0 for $\hat{\sigma}(\lambda)$ and β_0 for $\hat{\beta}(\lambda)$ in (3.19) and (3.20). This is a first step towards the issue of consistency in the case in which these parameters are also estimated. The results will be summarized in a later section. Lemma 3.15 and Theorem 3.14 would then imply that the minimum distance estimator is consistent for the transformation parameter.

Lemma 3.16: Assume that the distribution $F_{\epsilon}(x)$ is symmetric and strictly increasing on the support $[-M,M]$ of ϵ, and that the weight $w(x)$ is strictly positive. If for each fixed $X \in \mathfrak{X}$, that is, for each fixed $f_0 = f(X,\beta_0)$, the quantity $\varphi_x(\lambda)$ defined by (3.23) is such that for $\lambda \neq \lambda_0$ and all $x \neq 0$, $\varphi_x(\lambda) \neq -\varphi_{-x}(\lambda)$, then $V^2(\lambda) > V^2(\lambda_0)$.

proof: First note that $\varphi_x(\lambda_0) = x$ for all f_0, and hence by (3.25) and (3.26), $V^2(\lambda_0) = 0$. If F_{ϵ} is strictly increasing and $\varphi_x(\lambda)$ satisfies the conditions of the lemma, then $\Psi(X,x,\lambda)$ in the notation defined by (3.25) is strictly non-zero. By (3.26), we conclude that $V^2(\lambda) > 0$ if $\lambda \neq \lambda_0$.

■

Lemma 3.17: Modified Power Transform, $\lambda_o \neq 0$.

Let $h(y,\lambda)$ be the modified power transform of Box and Cox given by

$$(3.27) \qquad h(y,\lambda) = \begin{cases} \{y^\lambda - 1\} / \lambda & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0. \end{cases}$$

Assume that the distribution $F_\epsilon(x)$ is symmetric and strictly increasing on the support $[-M,M]$ of ϵ, and that $\lambda_o \neq 0$.

Then, $V^2(\lambda) > V^2(\lambda_o)$ if $\lambda \neq \lambda_o$.

proof: We find that for this family of transformations,

$$h^{-1}(u,\lambda) = \begin{cases} (\lambda u + 1)^{1/\lambda} & \text{if } \lambda \neq 0 \\ \exp(u) & \text{if } \lambda = 0. \end{cases}$$

Consider two cases for λ.

Case A. $\lambda \neq 0$.

The quantity $\varphi_x(\lambda)$ in (3.23) is in this case

$$(3.28) \qquad \begin{aligned} \varphi_x(\lambda) &= \{ [\lambda(\sigma_o x + [f_o{}^\lambda - 1] / \lambda) + 1]^{\lambda_o/\lambda} - 1\} / \sigma_o\lambda_o \\ &\qquad - [f_o{}^{\lambda_o} - 1] / \sigma_o\lambda_o \\ &= ([f_o{}^\lambda + \lambda\sigma_o x]^{\lambda_o/\lambda} - 1) / \sigma_o\lambda_o - [f_o{}^{\lambda_o} - 1] / \sigma_o\lambda_o \\ &= ([f_o{}^\lambda + \lambda\sigma_o x]^{\lambda_o/\lambda} - f_o{}^{\lambda_o}) / \sigma_o\lambda_o. \end{aligned}$$

Fix $\lambda \neq \lambda_o$ and consider the function g defined by

$$g(u) = u^{\lambda_o/\lambda}.$$

Since

$$g'(u) = (\lambda_o/\lambda)\, u^{\lambda_o/\lambda - 1} \text{ and}$$
$$g''(u) = (\lambda_o/\lambda)(\lambda_o/\lambda - 1)\, u^{\lambda_o/\lambda - 2},$$

we observe that if $\lambda > \lambda_o$ or $\lambda < \lambda_o$, g is either strictly concave or strictly convex. Thus

$$[f_o{}^\lambda + \lambda\sigma_o x]^{\lambda_o/\lambda} - (f_o{}^\lambda)^{\lambda_o/\lambda}$$

is either strictly greater or strictly smaller than

$$(f_o^{\lambda})^{\lambda_o/\lambda} - [f_o^{\lambda} - \lambda\sigma_o x]^{\lambda_o/\lambda}$$

for all $x \neq 0$. Therefore $\varphi_x(\lambda) \neq -\varphi_{-x}(\lambda)$, and the conclusion follows from Lemma 3.16.

Case B. $\lambda = 0$.

In this case, (3.23) yields

$$\begin{aligned}
(3.29) \qquad \varphi_x(\lambda) &= \{[\exp(\sigma_o x + \log(f_o)]^{\lambda_o} - 1\} \,/\, \sigma_o\lambda_o \\
&\quad - [f_o^{\lambda_o} - 1] \,/\, \sigma_o\lambda_o \\
&= \{\exp(\sigma_o\lambda_o x + \log(f_o^{\lambda_o})) - f_o^{\lambda_o}\} \,/\, \sigma_o\lambda_o \\
&= \{\exp(\sigma_o\lambda_o x + \log(f_o^{\lambda_o})) - \exp(\log(f_o^{\lambda_o}))\} \,/\, \sigma_o\lambda_o .
\end{aligned}$$

Consider the function g defined by $g(u) = \exp(u)$. Since g is strictly convex, we conclude

$$\exp(\sigma_o\lambda_o x + \log(f_o^{\lambda_o})) - \exp(\log(f_o^{\lambda_o}))$$

is either strictly larger or strictly smaller than

$$\exp(\log(f_o^{\lambda_o})) - \exp(-\sigma_o\lambda_o x + \log(f_o^{\lambda_o}))$$

for all $x \neq 0$. Therefore $\varphi_x(\lambda) \neq -\varphi_{-x}(\lambda)$, and the conclusion also follows from Lemma 3.16.

■

Lemma 3.18: Log Transformation.

Let $h(y,\lambda)$ be the modified power transform of Box and Cox given by (3.27). Assume that the distribution $F_\epsilon(x)$ is symmetric and strictly increasing on the support $[-M,M]$ of ϵ, and that $\lambda_o = 0$, so that the true transformation is the log transform.

Then, $V^2(\lambda) > V^2(\lambda_o)$ if $\lambda \neq \lambda_o$.

proof: Fixing $\lambda \neq 0$, $h^{-1}(u,\lambda) = (\lambda u + 1)^{1/\lambda}$ and by (3.23) we obtain

$$\text{(3.30)} \quad \begin{aligned} \varphi_x(\lambda) &= \log[\lambda(\sigma_o x + (f_o^{\lambda} - 1)/\lambda) + 1)^{1/\lambda}] \,/\, \sigma_o - \log(f_o)/\sigma_o \\ &= \{ \log[(\lambda\sigma_o x + f_o^{\lambda})^{1/\lambda}] - \log(f_o) \} \,/\, \sigma_o \\ &= \{ \log(\lambda\sigma_o x + f_o^{\lambda}) - \log(f_o^{\lambda}) \} \,/\, \lambda\sigma_o . \end{aligned}$$

Since the function log(u) is concave, we conclude that

$$\log(\lambda\sigma_o x + f_o^{\lambda}) - \log(f_o^{\lambda}) \neq \log(f_o^{\lambda}) - \log(-\lambda\sigma_o x + f_o^{\lambda})$$

for all $x \neq 0$. Thus $\varphi_x(\lambda) \neq -\varphi_{-x}(\lambda)$, and the conclusion follows from Lemma 3.16.

■

Lemma 3.19: Estimating the shift for fixed power, $\lambda_o \neq 0$.

Let $h(y,\mu)$ be the modified shifted power transform of Box and Cox with fixed power λ_o, given by

$$\text{(3.31)} \quad h(y,\mu) = \begin{cases} \{(y - \mu)^{\lambda_o} - 1\}/\lambda_o & \text{if } \lambda_o \neq 0 \\ \log(y - \mu) & \text{if } \lambda_o = 0. \end{cases}$$

Assume that the distribution $F_\epsilon(x)$ is symmetric and strictly increasing on the support $[-M,M]$ of ϵ, and that $\lambda_o \neq 0$.

Then, $V^2(\mu) > V^2(\mu_o)$ if $\mu \neq \mu_o$.

proof: In this case, $h^{-1}(u,\mu) = \mu + (\lambda_o u + 1)^{1/\lambda_o}$. Equation (3.23) yields

$$\text{(3.32)} \quad \varphi_x(\mu) =$$

$$\begin{aligned} &(\{\mu + (\lambda_o[\sigma_o x + ((f_o - \mu)^{\lambda_o} - 1)/\lambda_o] + 1\,)^{1/\lambda_o} - \mu_o\}^{\lambda_o} - 1\,) \,/\, \lambda_o\sigma_o \\ &\qquad - ((f_o - \mu_o)^{\lambda_o} - 1) \,/\, \lambda_o\sigma_o \\ &= [\{ \mu + (\lambda_o\sigma_o x + (f_o - \mu)^{\lambda_o})^{1/\lambda_o} - \mu_o\}^{\lambda_o} - (f_o - \mu_o)^{\lambda_o}] \,/\, \lambda_o\sigma_o \\ &= (\lambda_o\sigma_o)^{-1} [\{ \mu - \mu_o + (\lambda_o\sigma_o x + (f_o - \mu)^{\lambda_o})^{1/\lambda_o} \}^{\lambda_o} \\ &\qquad - \{ \mu - \mu_o + ((f_o - \mu)^{\lambda_o})^{1/\lambda_o}\}^{\lambda_o}]. \end{aligned}$$

Now consider the function g defined by

$$g(u) = \{ \mu - \mu_o + u^{1/\lambda_o} \}^{\lambda_o}.$$

We find

$$g'(u) = \{ \mu - \mu_o + u^{1/\lambda_o} \}^{\lambda_o - 1} u^{1/\lambda_o - 1}$$

and

$$\begin{aligned} g''(u) &= u^{1/\lambda_o - 1}[(\lambda_o - 1)\{ \mu - \mu_o + u^{1/\lambda_o} \}^{\lambda_o - 2} \lambda_o^{-1} u^{1/\lambda_o - 1}] \\ &\quad + \{ \mu - \mu_o + u^{1/\lambda_o} \}^{\lambda_o - 1} (1/\lambda_o - 1) u^{1/\lambda_o - 2} \\ &= (1 - 1/\lambda_o) u^{2(1/\lambda_o - 1)} \{ \mu - \mu_o + u^{1/\lambda_o} \}^{\lambda_o - 2} \\ &\quad + (1/\lambda_o - 1) u^{1/\lambda_o - 2}\{ \mu - \mu_o + u^{1/\lambda_o} \}^{\lambda_o - 1} \\ &= (1 - 1/\lambda_o)\{ \mu - \mu_o + u^{1/\lambda_o} \}^{\lambda_o - 2} u^{1/\lambda_o - 2} \times \\ &\qquad [u^{1/\lambda_o} - \{ \mu - \mu_o + u^{1/\lambda_o} \}] \\ &= (1 - 1/\lambda_o)(\mu_o - \mu)\{ \mu - \mu_o + u^{1/\lambda_o} \}^{\lambda_o - 2} u^{1/\lambda_o - 2}. \end{aligned}$$

The sign of $g''(u)$ is constant for all $u > 0$, and changes as $\mu < \mu_o$ or $\mu > \mu_o$. Furthermore, $g''(u) \neq 0$ if $\mu \neq \mu_o$. Therefore, if $\mu \neq \mu_o$, the function g is either strictly concave or strictly convex in u. We then conclude that if $x \neq 0$,

$$\{\mu - \mu_o + (\lambda_o \sigma_o x + (f_o - \mu)^{\lambda_o})^{1/\lambda_o} \}^{\lambda_o} - \{ \mu - \mu_o + ((f_o - \mu)^{\lambda_o})^{1/\lambda_o}\}^{\lambda_o}$$

is either strictly larger or strictly smaller than

$$\{\mu - \mu_o + ((f_o - \mu)^{\lambda_o})^{1/\lambda_o}\}^{\lambda_o} - \{ \mu - \mu_o + (-\lambda_o \sigma_o x + (f_o - \mu)^{\lambda_o})^{1/\lambda_o}\}^{\lambda_o}$$

and hence $\varphi_x(\mu) \neq -\varphi_{-x}(\mu)$ if $\mu \neq \mu_o$. The conclusion now follows from Lemma 3.16.

■

Lemma 3.20: Shifted log transform.

Let $h(y,\mu)$ be the modified shifted power transform of Box and Cox with fixed power λ_o, given by (3.31), and suppose that $\lambda_o = 0$.

Assume that the distribution $F_\epsilon(x)$ is symmetric and strictly increasing on the support $[-M,M]$ of ϵ.

Then, $V^2(\mu) > V^2(\mu_o)$ if $\mu \neq \mu_o$.

proof: We check that in this case, $h^{-1}(u,\mu) = \mu + \exp(u)$. Equation (3.23) yields

(3.33) $\varphi_x(\mu) =$

$$\log\{ \mu + \exp[\sigma_o x + \log(f_o - \mu)] - \mu_o \} / \sigma_o - \log(f_o - \mu_o) / \sigma_o$$

$$= [\log\{ \mu - \mu_o + \exp[\sigma_o x + \log(f_o - \mu)] \} - \log\{f_o - \mu_o\}] / \sigma_o$$

$$= [\log\{ \mu - \mu_o + \exp[\sigma_o x + \log(f_o - \mu)] \}$$

$$- \log\{ \mu - \mu_o + \exp(\log(f_o - \mu))\}] / \sigma_o.$$

If we now consider the function g defined by

$$g(u) = \log\{ \mu - \mu_o + \exp(u) \},$$

we find

$$g'(u) = \{ \mu - \mu_o + \exp(u) \}^{-1} \exp(u)$$

and

$$g''(u) = \exp(u) [-\{ \mu - \mu_o + \exp(u) \}^{-2}\exp(u)]$$

$$+ \{ \mu - \mu_o + \exp(u) \}^{-1} \exp(u)$$

$$= \exp(u) \{ \mu - \mu_o + \exp(u) \}^{-2} \times$$

$$[-\exp(u) + \{ \mu - \mu_o + \exp(u) \}]$$

$$= \exp(u) (\mu - \mu_o) \{ \mu - \mu_o + \exp(u) \}^{-2}.$$

We note that the sign of $g''(u)$ is constant for all $u > 0$, changes sign as $\mu < \mu_o$ or $\mu > \mu_o$, and also $g''(u) \neq 0$ if $\mu \neq \mu_o$. Therefore, if $\mu \neq \mu_o$, the function g is either strictly concave or strictly convex in

u. We then conclude that if $x \neq 0$,

$$\log\{ \mu - \mu_0 + \exp[\sigma_0 x + \log(f_0 - \mu)] \} - \log\{ \mu - \mu_0 + \exp(\log(f_0 - \mu)) \}$$

is either strictly larger or strictly smaller than

$$\log\{\mu - \mu_0 + \exp(\log(f_0 - \mu)) \} - \log\{ \mu - \mu_0 + \exp[-\sigma_0 x + \log(f_0 - \mu)] \}$$

and hence $\varphi_x(\mu) \neq -\varphi_{-x}(\mu)$ if $\mu \neq \mu_0$. An application of Lemma 3.16 completes the proof.

■

We may summarize the results of this section in the following theorem.

Theorem 3.21: Assume the conditions A1-A7 of Assumptions 3.1 and that the distribution $F_\epsilon(x)$ is strictly increasing on the support $[-M,M]$ of ϵ. If β_0 and σ_0 are known, and the transformation h is either the modified power transformation (3.27) or the modified fixed-power transformation with shift (3.31), then the minimum distance estimator of the transformation parameter is strongly consistent.

proof: Follows from Theorem 3.14 and Lemmas 3.15 through 3.20.

■

3.7 Discussion

We showed that minimum distance estimation yields consistent estimates of the transformation parameter when β_0 and σ_0 are known. This was achieved under nonparametric assumptions on the error distribution, in the sense that we only assumed symmetry. If β_0 and σ_0 are unknown and are estimated by least-squares for fixed λ, consistency is a different matter. However, consistency in the (β_0,σ_0)-known case can be exploited to establish consistency in the more general case. We now sketch a procedure to obtain a consistent estimate of (λ,β,σ), which is similar to the method explained in Theorem 2.13 for the Ruppert-Aldershof transformation to symmetry. We let $Z_i = (X_i,Y_i)$, $i=1,\ldots,n$, denote the sequence of observations, and suppose that λ and β are t and r-dimensional, respectively.

1. Estimate β using the Least Absolute Deviation estimator $\hat{\beta}_{LAD}$ (Definition 2.10), that is, $\hat{\beta}_{LAD}$ minimizes

$$(3.34) \qquad \sum_{i=1}^{n} | Y_i - f(X_i,\beta) |$$

as a function of β. We know (section 2.3) that $\hat{\beta}_{LAD} \xrightarrow{a.s.} \beta_0$ if the observations follow the TBS model with errors that have a unique median at zero, regardless of the unknown value of λ_0.

2. Estimate λ by minimizing

$$(3.35) \qquad V_n^2(\lambda) = \int \{ F_n(x,\lambda) + F_n(-x,\lambda) - 1 \}^2 \, w(x)\, dx$$

where

$$F_n(x,\lambda) = (1/n) \sum_{i=1}^{n} 1(r(\lambda,\hat{\beta}_{LAD},Z_i) \le x)$$

and denote this estimate by $\hat{\lambda}_{LAD}$. By using arguments similar to the ones in Theorem 2.13, as well as Lemmas 3.16 through 3.20, we would obtain $\hat{\lambda}_{LAD} \xrightarrow{a.s.} \lambda_0$ whenever h is the modified power transformation or the modified fixed-power transformation with shift.

3. Define the estimate of σ by

$$\hat{\sigma}^2_{LAD}(\lambda) = (1/n) \sum_{i=1}^{n} r^2(\hat{\lambda}_{LAD}, \hat{\beta}_{LAD}, Z_i) \tag{3.36}$$

and put $\hat{\theta}_{LAD} = (\hat{\beta}_{LAD}, \hat{\lambda}_{LAD}, \hat{\sigma}_{LAD})$. Using $\hat{\theta}_{LAD}$ as a starting point, now let $\hat{\theta}^*_n$ be a one-step Newton-Raphson approximation (section 2.3) to the solution of the following system of equations:

$$(1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial}{\partial\lambda_k} \int a_i(x,\lambda)\, a_j(x,\lambda)\, w(x)\, dx = 0 \tag{3.37}$$

$$1 \le k \le t$$

$$(1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} \Big\{ r(\lambda,\beta,Z_i) \frac{\partial}{\partial\beta_k} r(\lambda,\beta,Z_i) + r(\lambda,\beta,Z_j) \frac{\partial}{\partial\beta_k} r(\lambda,\beta,Z_j) \Big\} = 0$$

$$1 \le k \le r$$

$$(1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} [(1/2)\{ r^2(\lambda,\beta,Z_i) + r^2(\lambda,\beta,Z_j) \} - \sigma^2] = 0$$

where

$$a_i(x,\lambda) = 1(r(\lambda,\beta,Z_i) \le \sigma x) + 1(r(\lambda,\beta,Z_i) \le -\sigma x) - 1.$$

By definition, the estimate $\hat{\theta}_{CVM}$ solves system (3.37), but we are once again confronted with the possibility that multiple solutions exist. The algorithm just described serves as a means of assuring that a consistent solution is chosen. Since it is true that $\hat{\theta}_{LAD} \xrightarrow{a.s.} \theta_o$ at an appropriate rate, we could then prove $\hat{\theta}^*_n \xrightarrow{a.s.} \theta_o$, using similar techniques as were applied in the proof of Theorem 2.13. Theorem 3.21 is the foundation for this last result. Additionally, we conclude that $n^{1/2}\hat{\theta}_{CVM}$ and $n^{1/2}\hat{\theta}^*_n$ have the same asymptotic distribution which will be derived in Chapter IV; therefore, $\hat{\theta}_{CVM}$ and $\hat{\theta}^*_n$ are asymptotically equivalent for estimation purposes.

In Lemmas 3.19 and 3.20, we considered the transformation which applies a shift before taking a power. This transformation is a restricted case of a more general family of transformations known as the *family of modified power transformations with shift*, which is defined by

$$(3.38) \qquad h(y,\mu,\lambda) = \begin{cases} [(y-\mu)^{\lambda} - 1]/\lambda & \lambda \neq 0 \\ \log(y-\mu) & \lambda = 0 \end{cases}.$$

It is interesting to consider this two-parameter family of transformations, because it may generate observations Y_i that are not necessarily positive as we had to assume when we considered the Box-Cox transform. By an appropriate choice of μ, all $Y_i - \mu$ can be made positive. Adding a positive constant to all the Y_i's would be desirable, for instance, if one of them is zero and we wished to include the possibility of applying the log transform. Carroll and Ruppert (1988) study an example (Example 6.2) using the modified power transformation (3.27) where some of the Y measurements were zero and inference on λ is performed after adding an arbitrary fixed constant to all observations. They found that estimates $\hat{\lambda}$ of the transformation parameter for fixed shifts are very sensitive to the choice of μ. This observation illustrates that the capability of using data to estimate (μ,λ) is essential. Atkinson (1985) discusses shift in modified power transformation in detail. In particular, he shows that the maximum likelihood estimate of (μ,λ) under normal errors fails to be consistent. The problem lies in the fact that the likelihood function is unbounded. More precisely, the likelihood tends to infinity when $-\mu$ tends to the minimum of the Y_i's.

If the power parameter in (3.38) is fixed at λ_o, the problem reduces to estimation of the shift. In practical terms, this situation is interesting because it is quite common that a researcher wishes to apply a specified transformation, for example the log transformation if the data seem to be positively skewed and having variance increasing with the mean. However, it may be the case that some observations are zero or even negative and adding a constant to observations is required before taking logs. Since this constant is an unknown parameter, it would be an advantage to have the capability of estimating its value. Lemmas 3.29 and 3.20 show that minimum distance estimation is consistent in this fixed-λ_o case, whereas methods such as maximum likelihood still exhibit the difficulties mentioned above.

The minimum distance estimator of a transformation also has the potential to consistently estimate both of the parameters of a power transformation with shift, simultaneously. Of course, the computation of the minimum distance estimator would involve solving (3.6) in two-dimensional (μ,λ) space, and this introduces the need for numerical optimization methods in higher dimensions that do not rely on differentiability. However, it turns out that numerical procedures are not the main problem here. Rather, the complications are statistical in nature: Jointly estimating (μ,λ) in (3.38) is not a task that can be accomplished efficiently by symmetry considerations alone. More shall be said about this problem in Chapter IV, after we derive the asymptotic covariance of the minimum distance estimate.

CHAPTER IV

ASYMPTOTIC NORMALITY OF THE MINIMUM DISTANCE ESTIMATOR

4.0 Introduction

The goal in this chapter is to establish asymptotic normality of the minimum distance estimator of transformation parameters which was described in Chapter III. We saw that the proposed estimator is a special case of generalized M-estimation, minimizing the expression

$$(4.1) \qquad (1/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} K(X_{i_1}, \ldots, X_{i_m}, \theta)$$

over the parameter space Θ, thus giving rise to the generalized M-estimator $\hat{\theta}_n$. We wish to investigate conditions under which a sequence of estimates $\hat{\theta}_n$ that minimizes (4.1), and that satisfies convergence in probability to some fixed $\theta_0 \in \Theta$, will also satisfy

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0,V)$$

for some (positive definite) matrix V.

Application of the result to the special case of estimating the transformation parameters in the transform-both-sides regression setting will directly yield asymptotic normality of the minimum distance estimator of section 3.1. Of course, what we achieve with this conclusion is the capability of obtaining large sample confidence sets and approximate tests of hypotheses with respect to the parameter $\theta = (\lambda, \beta, \sigma)$.

Testing hypotheses about transformation parameters is of special interest because tests can be performed to assess the validity of a postulated model. For instance, if the family of transformations is $h(y,\lambda) = (y^{\lambda} - 1)/\lambda$, then testing for the null hypothesis H_o: $\lambda = 1$ amounts to testing for additivity of the error in the regression model. More general hypotheses, say linear combinations of the elements of θ, are obtained by any one of standard asymptotic tests. Examples of these asymptotic methods are the Wald-type confidence sets and tests, and the Rao score tests.

Assuming that consistency holds, we proceed in section 4.1 with the theory of asymptotic normality for generalized M-estimators in the general setting, and then return to the particular case of estimating a transformation parameter in section 4.2. In section 4.3 we study a special case of a regression model which was selected to serve as an example for further examination; we carry out actual computations for the matrix V assuming the single sample model $f(X,\beta_o) = \beta_o$ for all X, that is,

$$h(Y_i,\lambda_o) = h(\beta_o,\lambda_o) + \sigma_o\epsilon_i. \tag{4.2}$$

This concrete case will be enlightening because of two main reasons. First, some comparisons with other estimation procedures can be accomplished, for instance with maximum likelihood, the Taylor estimates, or the Ruppert-Aldershof estimator. Secondly, we will be able to assess the effect of the weighting function w(x) used in the computation of the minimum distance estimate. It is important to investigate how the asymptotic variance is dependent on the particular choice of w(x) for different error distributions. Understanding the effect of

$w(x)$, we then could give some guidelines for the choice of $w(x)$ in order to obtain an efficient estimator $\hat{\theta}_n$.

4.1 Asymptotic Normality of Generalized M-Estimators

We consider the setting of generalized M-estimation, which was introduced in section 3.3:

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables on a space $(\mathcal{X}, \mathcal{A}, P)$ and let $\Theta \subset \mathbb{R}^p$ be a parameter space. We assume there is a function $\Psi : \mathcal{X}^m \times \Theta \longrightarrow \mathbb{R}^p$ such that $\Psi(x_1, \ldots, x_m, \theta)$ is symmetric in its x arguments for each fixed $\theta \in \Theta$, and that a sequence $\{\hat{\theta}_n\}$ of random variables satisfies

$$(4.3) \qquad (n^{1/2}/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \Psi(X_{i_1}, \ldots, X_{i_m}, \hat{\theta}_n) \xrightarrow{P} 0.$$

It is possible that each $\hat{\theta}_n$ is obtained by minimizing the expression

$$(1/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} K(X_{i_1}, \ldots, X_{i_m}, \theta)$$

over θ for some function $K : \mathcal{X}^m \times \Theta \longrightarrow \mathbb{R}$, so that $\hat{\theta}_n$ satisfies (4.3) with $\Psi = \frac{\partial}{\partial \theta} K$.

We present the proof of asymptotic normality after several preliminary results. The first of these extends asymptotic normality of U-statistics (Lemma 3.11) to kernels that have values in $\mathbb{R}^k$, $k \geq 1$.

<u>Lemma 4.1</u>: Let $K : \mathcal{X}^m \longrightarrow \mathbb{R}^k$ be a symmetric kernel and define:

$$U_n = \frac{1}{\binom{n}{m}} \sum_{C} K(X_{i_1}, \ldots, X_{i_m}),$$

$$\gamma = E[K(X_1, \ldots, X_2)],$$

$$h_1(x) = E[K(x, X_2, \ldots, X_m)], \text{ and}$$

$$\zeta_1 = \mathrm{COV}[h_1(X)].$$

Then, if $E|\ K(X_1, \ldots, X_m)|^2 < \infty$ and ζ_1 is positive definite,

$$n^{1/2}(U_n - \gamma) \xrightarrow{d} N_k(0,\ m^2\zeta_1).$$

<u>proof</u>: Using the Cramer-Wald device, it is sufficient to prove $n^{1/2}(a^TU_n - a^T\gamma) \xrightarrow{d} N_1(0,\ m^2a^T\zeta_1a)$ for every $a \in \mathbb{R}^k$. Therefore fix $a \in \mathbb{R}^k$ and define

$$\begin{aligned} K^*(X_1, \ldots, X_m) &= a^TK(X_1, \ldots, X_m), \\ \gamma^* &= E[K^*(X_1, \ldots, X_m)], \\ U_n^* &= a^TU_n, \\ h_1^*(x) &= E[K^*(x, X_2, \ldots, X_m)], \\ \text{and} \quad \zeta_1^* &= \mathrm{VAR}[h_1^*(X_1)]. \end{aligned}$$

Then notice that

$$h_1^*(x) = a^Th_1(x),$$

$$\zeta_1^* = \mathrm{VAR}[a^Th_1(X_1)] = a^T\mathrm{COV}[h_1(X_1)]a = a^T\zeta_1a,$$

and $\gamma^* = a^T\gamma$.

Now, $E|\ h^*(X_1, \ldots, X_m)| < \infty$, and since ζ_1 is positive definite, then $\zeta_1^* > 0$, so by Lemma 3.11, we have

$$n^{1/2}(U_n^* - \gamma^*) \xrightarrow{d} N(0, m^2\zeta_1^*)$$

that is,

$$n^{1/2}(a^T U_n - a^T \gamma) \xrightarrow{d} N(0,\ m^2 a^T \zeta_1 a).$$

■

The following two results extend the work of Huber (1967) on the asymptotic normality of M-estimators. We put $x = (x_1, \ldots, x_m)$ and define

$$u(x_1, \ldots, x_m, \theta, d) = u(x, \theta, d) = \sup_{|\tau-\theta| \leq d} |\ \Psi(x, \tau) - \Psi(x, \theta)|$$

and

$$\lambda(\theta) = E\Psi(X, \theta) = \int \ldots \int \Psi(x_1, \ldots, x_m, \theta) dP(x_1) \ldots dP(x_m).$$

We notice that $u(x_1, \ldots, x_m, \theta, d)$ is symmetric in $x_1, \ldots, x_m$ for each fixed (θ, d), if Ψ is symmetric. We also make the following assumptions.

<u>Assumptions 4.2</u>:

A1. For each fixed $\theta \in \Theta \subset \mathbb{R}^p$, $\Psi(x, \theta)$ is $\mathcal{A}$-measurable

A2. There is a $\theta_o \in \Theta$ such that $\lambda(\theta_o) = 0$

A3. There are strictly positive numbers a, b, c, d_o such that

(i) $|\ \lambda(\theta)| \geq a|\ \theta - \theta_o|$ for $|\ \theta - \theta_o| \leq d_o$

(ii) $E[u(X, \theta, d)] \leq b \cdot d$ for $|\ \theta - \theta_o| + d \leq d_o$

(iii) $E[u(X, \theta, d)]^2 \leq c \cdot d$ for $|\ \theta - \theta_o| + d \leq d_o$

A4. $E|\ \Psi(X, \theta_o)|^2 < \infty$

■

Lemma 4.3: Let

$$Z_n(\tau,\theta) = \frac{\left| \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} K(X_{i_1},\ldots,X_{i_m},\tau) - K(X_{i_1},\ldots,X_{i_m},\theta) - \lambda(\tau) + \lambda(\theta) \right|}{n^m/n^{1/2} + n^m|\lambda(\tau)|}.$$

Assumptions A1, A2, and A3 imply

$$\sup_{|\tau-\theta_o|\leq d_o} Z_n(\tau,\theta_o) \xrightarrow{P} 0, \text{ as } n \longrightarrow \infty.$$

proof: Huber's rigorous proof of Lemma 3 in his 1967 paper pushes through. Only one of the major changes required will be mentioned here. Basically, a sum $\sum_{i=1}^{n} \Psi(X_i,\theta)$ is replaced by the sum

$$(1/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \Psi(X_{i_1},\ldots, X_{i_m},\theta),$$

and wherever a reference to Chebyshev's inequality is encountered, Lemma 3.12 is used instead.

■

Lemma 4.4: Assume that Assumptions 4.2 hold, and that $\hat{\theta}_n$ satisfies

$$(n^{1/2}/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \Psi(X_{i_1},\ldots, X_{i_m},\hat{\theta}_n) \xrightarrow{P} 0 \tag{4.4}$$

and that

$$P(|\hat{\theta}_n - \theta_o| \leq d_o) \longrightarrow 1 \text{ as } n \longrightarrow \infty. \tag{4.5}$$

Then

$$(n^{1/2}/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \Psi(X_{i_1},\ldots, X_{i_m},\theta_o) + n^{1/2}\lambda(\hat{\theta}_n) \xrightarrow{P} 0.$$

proof: The reader is again referred to Huber's 1967 paper and the proof of his Theorem 3 therein. The current proof is supported by Lemma 4.3, and similar changes are required to make the former proof go through.

■

The following theorem is the main result of this chapter:

Theorem 4.5: Suppose Assumptions 4.2 hold, and that $\{\hat{\theta}_n\}$ is a sequence of random variables that satisfies (4.4) and (4.5) for some $\theta_o \in \Theta$. Also assume that the derivative $\frac{\partial}{\partial\theta}\lambda(\theta)$ exists and is nonsingular when evaluated at θ_o. Let

$$h_1(x) = E[\Psi(x, X_2, \ldots, X_m, \theta_o)],$$

$$A = \frac{\partial}{\partial\theta}\lambda(\theta)\bigg|_{\theta=\theta_o}, \text{ and}$$

$$B = COV[h_1(X)] = E[h_1(x)h_1^T(x)].$$

Then,

$$n^{½}(\hat{\theta}_n - \theta_o) \xrightarrow{d} N(0, m^2A^{-1}BA^{-T}).$$

proof: Using the assumption that $\lambda(\theta_o) = 0$, the definition of a derivative yields

$$n^{½}\lambda(\hat{\theta}_n) = n^{½}A(\hat{\theta}_n - \theta_o) + o_p(|\hat{\theta}_n - \theta_o|)$$

and thus

$$n^{1/2} A (\hat{\theta}_n - \theta_0) = n^{1/2} \lambda(\hat{\theta}_n) + o_p(|\hat{\theta}_n - \theta_0|)$$

$$= [\, n^{1/2} \lambda(\hat{\theta}_n) + (n^{1/2}/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \Psi(X_{i_1}, \ldots, X_{i_m}, \theta_0) \,]$$

$$- (n^{1/2}/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \Psi(X_{i_1}, \ldots, X_{i_m}, \theta_0) + o_p(|\hat{\theta}_n - \theta_0|).$$

By Lemma 4.4, the term in square brackets converges to zero in probability, and by Lemmas 4.1 and 3.9,

$$(n^{1/2}/n^m) \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \Psi(X_{i_1}, \ldots, X_{i_m}, \theta_0) \xrightarrow{d} N(0, m^2 B).$$

Thus, by Slutsky's theorem, we obtain

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0,\ A^{-1} m^2 B A^{-T}).$$

■

4.2 A Special Case: Minimum Distance Estimation of a Transformation

We turn to the minimum distance estimator described in Definition 3.3. We restate the definition here, and then show explicitly how the minimum distance estimator is viewed as a generalized M-estimator. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ under the assumption that the transform-both-sides regression model holds:

$$h(Y,\lambda) = h(f(X,\beta),\lambda) + \sigma\epsilon.$$

Notation 4.6:

N1. $$Z=(X,Y), \quad Z_i=(X_i,Y_i)$$

N2. $$r(\lambda,\beta,Z) = h(Y,\lambda) - h(f(X,\beta),\lambda)$$

N3. $\hat{\beta}(\lambda)$ is the least-squares estimator for fixed λ, that is, $\hat{\beta}(\lambda)$ minimizes $\sum_{i=1}^{n} \{r(\lambda,\beta,Z_i)\}^2$ as a function of β.

N4. $$\hat{\sigma}^2(\lambda) = (1/n) \sum_{i=1}^{n} r^2(\lambda,\hat{\beta}(\lambda),Z_i)$$

N5. $$r_i^*(\lambda) = r(\lambda,\hat{\beta}(\lambda),Z_i) \,/\, \hat{\sigma}(\lambda).$$

N6. $$F_n(x,\lambda) = (1/n) \sum_{i=1}^{n} 1(\, r_i^*(\lambda) \le x \,)$$

N7. $$Q_n(x,\lambda) = F_n(x,\lambda) + F_n(-x,\lambda) - 1$$

N8. $$V_n^2(\lambda) = \int Q_n^2(x,\lambda)\; w(x)\; dx$$

■

<u>Definition 4.7</u>: The *Minimum Distance Estimator* $\hat{\lambda}_{CVM}$ was defined by minimizing the quantity $V_n^2(\lambda)$ over λ:

(4.6) $$\inf_{\lambda\in\Lambda} V_n^2(\lambda) = V_n^2(\hat{\lambda}_{CVM}).$$

This also defined the following estimators of β and σ:

$$\hat{\beta}_{CVM} = \hat{\beta}(\hat{\lambda}_{CVM})$$

$$\hat{\sigma}^2_{CVM} = \hat{\sigma}(\hat{\lambda}_{CVM})$$

■

Letting $\theta = (\lambda,\beta,\sigma)$ be the parameter, we wish to show that the minimum distance estimator $\hat{\theta}_{CVM}$ thus defined is truly a generalized M-estimator in the sense of (4.3). We assume that λ is t-dimensional and β is r-dimensional, and we record the notation

$$a_i(x,\lambda) = 1(r(\lambda,\beta,Z_i) \le \sigma x) + 1(r(\lambda,\beta,Z_i) \le -\sigma x) - 1$$

$$1 \le i \le n.$$

Now, in order to observe how $\hat{\theta}_{CVM}$ is a Generalized M-Estimator even when (β,σ) is also estimated using least-squares, we note that $\hat{\theta}_{CVM}$ is a solution to the following set of equations:

$$(4.7) \qquad (1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial}{\partial \lambda_k} \int a_i(x,\lambda)\, a_j(x,\lambda)\, w(x)\, dx = 0$$

$$1 \le k \le t$$

$$(1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ r(\lambda,\beta,Z_i) \frac{\partial}{\partial \beta_k} r(\lambda,\beta,Z_i) + r(\lambda,\beta,Z_j) \frac{\partial}{\partial \beta_k} r(\lambda,\beta,Z_j) \right\} = 0$$

$$1 \le k \le r$$

$$(1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} [(1/2)\{ r^2(\lambda,\beta,Z_i) + r^2(\lambda,\beta,Z_j) \} - \sigma^2] = 0.$$

We do note that all the summands are symmetric functions of Z_i and Z_j. The system of equations (4.7) defines the function $\Psi : \mathfrak{X}^2 \times \Theta \longrightarrow \mathbb{R}^p$ such that the estimator $\hat{\theta}_{CVM}$ satisfies

$$(1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} \Psi(Z_i, Z_j, \hat{\theta}_{CVM}) = 0.$$

If consistency of $\hat{\theta}_{CVM}$ is assumed, then in order to achieve asymptotic normality, the conditions of Theorem 4.5 must be verified. The stated conclusion would then be

$$n^{1/2}(\hat{\theta}_{CVM} - \theta_o) \xrightarrow{d} N(0,\ 4A^{-1}BA^{-T})$$

for specified matrices A and B, where θ_o is such that $E\Psi(Z_1, Z_2, \theta_o) = 0$.

The situation that we are mainly interested in is the transform-both-sides regression setting. We check that Assumptions 4.2 are verified for this situation, in the following lemma.

Lemma 4.8: If $Z_i = (X_i, Y_i)$ for $i=1, \ldots, n$ are independent and identically distributed observations following the TBS model, the supports of X and ϵ are compact, the errors ϵ are symmetric, the random variables X and ϵ are independent, and if Ψ is defined by equations (4.7), then Assumptions 4.2 are satisfied.

proof: Measurability of Ψ follows from continuity. We next wish to show that there exists a θ_o such that $E[\Psi(Z_1, Z_2, \theta_o)] = 0$. If there is a parameter value θ_o for which the TBS model holds with symmetric errors, then $r(\lambda_o, \beta_o, Z) = \sigma_o \epsilon$. Also, the model implies that Y can be expressed in the form $Y = g(X, \epsilon)$ for some function g. Thus $Z = (X, Y)$ may be written $Z = (X, g(X, \epsilon))$, where the random variables X and ϵ have been assumed to be independent of each other. Since

$$r(\lambda, \beta, Z) = h(Y, \lambda) - h(f(X, \beta), \lambda),$$

the derivative $(\partial/\partial\beta_k) r(\lambda, \beta, Z)$ is independent of Y, and thus it does not depend on ϵ. Therefore

$$\begin{aligned}(4.8) \qquad E[r(\lambda_o, \beta_o, Z) \frac{\partial}{\partial\beta_k} r(\lambda_o, \beta_o, Z)] &= E[\sigma_o \epsilon \frac{\partial}{\partial\beta_k} r(\lambda_o, \beta_o, Z)] \\ &= E_\epsilon E_X[\sigma_o \epsilon \frac{\partial}{\partial\beta_k} r(\lambda_o, \beta_o, Z)] = \sigma_o E_\epsilon[\epsilon E_X\{ \frac{\partial}{\partial\beta_k} r(\lambda_o, \beta_o, Z)\}] \\ &= \sigma_o E_X[\frac{\partial}{\partial\beta_k} r(\lambda_o, \beta_o, Z)] E_\epsilon[\epsilon] = 0\end{aligned}$$

and since $E\epsilon^2 = 1$,

$$(4.9) \qquad E[r^2(\lambda,\beta,Z) - \sigma_o^2] = E[\sigma_o^2\epsilon^2 - \sigma_o^2] = 0.$$

Now let $F(x,\theta) = P(\ r(\lambda,\beta,Z) \leq \sigma x\ )$. The function

$$\int a_1(x,\lambda)\ a_2(x,\lambda)\ w(x)\ dx$$

is Lipchitz continuous as a function of λ, so the following interchange of expectation and differentiation is allowed:

$$\begin{aligned} E[\ \frac{\partial}{\partial\lambda_k} \int a_1(x,\lambda)\ a_2(x,\lambda)\ w(x)\ dx\ ] &= \frac{\partial}{\partial\lambda_k} E[\int a_1(x,\lambda)\ a_2(x,\lambda)\ w(x)\ dx\ ] \\ &= \frac{\partial}{\partial\lambda_k} \int E[a_1(x,\lambda)\ a_2(x,\lambda)]\ w(x)\ dx \\ &= \int \frac{\partial}{\partial\lambda_k} \{\ F(x,\theta) - F(-x,\theta) + 1\ \}^2\ w(x)\ dx \\ &= \int 2\{F(x,\theta) - F(-x,\theta) + 1\} \frac{\partial}{\partial\lambda_k} \{F(x,\theta) - F(-x,\theta) + 1\}\ w(x)\ dx. \end{aligned}$$

Since $F(x,\theta_o) = F_\epsilon(x)$ and F_ϵ is a symmetric distribution, then the above quantity is zero when evaluated at $\theta = \theta_o$ for all $k=1,\ldots,t$. Thus by (4.8) and (4.9), assumption A2 holds.

Assumptions A3 and A4 follow from the compact supports that were assumed for X and ϵ; since $\Psi(Z_1,Z_2,\theta)$ is bounded, then $u(x,\theta,d)$ is bounded.

■

Theorem 4.5 and Lemma 4.8 yield the following result:

Theorem 4.9: (Asymptotic normality of the minimum distance estimator)

Under the assumptions of Lemma 4.8, if $\hat{\theta}_{CVM} \xrightarrow{P} \theta_o$, then

$$n^{1/2}(\hat{\theta}_{CVM} - \theta_o) \xrightarrow{d} N(0,\ 4A^{-1}BA^{-T})$$

where

$$A = \frac{\partial}{\partial \theta} E[\Psi(Z_1, Z_2, \theta)] \Big|_{\theta=\theta_o} \quad \text{and}$$

$$B = \mathrm{COV}_{Z_1}(E_{Z_2}[\Psi(Z_1, Z_2, \theta_o)])$$

$$= \int \{ \int \Psi(z_1, z_2, \theta_o)\, dP(z_2) \}^2\, dP(z_1).$$

■

The next theorem shows that the nuisance parameter σ can be disregarded since it does not have an effect on the asymptotic covariance of $(\hat{\lambda}, \hat{\beta})$ when compared to the covariance that results from assuming that σ was known and did not have to be simultaneously estimated. This enables us to concentrate only on the covariance of $(\hat{\lambda}, \hat{\beta})$ when performing actual computations. For this reason, we will hardly refer to minimum distance estimates of σ in the remaining part of this work. The absence of σ estimates will be particularly noticeable in Chapter V.

<u>Lemma 4.10</u>: (Ruppert and Aldershof (1989))

Let A and B be $q \times q$ matrices such that A is partitioned as follows

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where A_{11} and B_{11} are $q_1 \times q_1$, and A_{22} and B_{22} are $q_2 \times q_2$. Then the $q_1 \times q_1$ upper-left corner of $A^{-1}BA^{-T}$ is $A_{11}^{-1}B_{11}A_{11}^{-T}$.

proof: Matrix algebra using $A^{-1} = \begin{bmatrix} A_{11}^{-1} & 0 \\ -A_{22}^{-1}A_{21}A_{11}^{-1} & A_{22}^{-1} \end{bmatrix}$.

■

Theorem 4.11: Let $(\hat{\lambda}^*_{CMV}, \hat{\beta}^*_{CVM})$ be the minimum distance estimate of the parameter (λ, β) assuming σ_o is known and did not need to be estimated, and let $(\hat{\lambda}_{CMV}, \hat{\beta}_{CVM}, \hat{\sigma}_{CVM})$ be the minimum distance estimate of (λ, β, σ) when σ is also estimated. Denote the corresponding asymptotic covariances by Σ^* and Σ, respectively. Then, the $(t+r)\times(t+r)$ upper-left corner of Σ is equal to Σ^*.

proof: Using Notation 4.6, by Theorem 4.9 and Lemma 4.10, it is sufficient to prove that

$$(4.10) \qquad A^k_{23}(\theta) =: \frac{\partial}{\partial\sigma} E[(1/2)\{ r(\lambda,\beta,Z_1) \frac{\partial}{\partial\beta_k} r(\lambda,\beta,Z_1) + r(\lambda,\beta,Z_2) \frac{\partial}{\partial\beta_k} r(\lambda,\beta,Z_2) \}] \qquad 1 \leq k \leq r$$

and

$$(4.11) \qquad A^k_{13}(\theta) =: \frac{\partial}{\partial\sigma} E[\frac{\partial}{\partial\lambda_k} \int a_1(x,\lambda)\, a_2(x,\lambda)\, w(x)\, dx] \qquad 1 \leq k \leq t$$

are all zero when evaluated at $\theta = \theta_o$.

Clearly $A^k_{23}(\theta) = 0$ for all θ and all k, since the function that is differentiated with respect to σ in equation (4.10) does not depend on σ.

For the terms $A^k_{13}(\theta)$, let us note, putting

$$F(x,\theta) = P(r(\lambda,\beta,Z) \leq \sigma x)$$

and

$$M(x,\theta) = Ea_1(x,\theta) = F(x,\theta) + F(-x,\theta) - 1,$$

that we have assumed enough conditions so that

$$(4.12) \qquad A_{13}^k(\theta) = \frac{\partial}{\partial\sigma}\int \frac{\partial}{\partial\lambda_k} E[a_1(x,\lambda)\ a_2(x,\lambda)\ w(x)\ ]\ dx$$

$$= \frac{\partial}{\partial\sigma}\int \frac{\partial}{\partial\lambda_k} M^2(x,\theta)\ w(x)\ dx = \frac{\partial}{\partial\sigma}\int 2\ M(x,\theta)\ \frac{\partial}{\partial\lambda_k} M(x,\theta)\ w(x)\ dx$$

$$= 2\int [\frac{\partial}{\partial\sigma} M(x,\theta)\ \frac{\partial}{\partial\lambda_k} M(x,\theta) + M(x,\theta)\ \frac{\partial^2}{\partial\sigma\partial\lambda_k} M(x,\theta)]\ w(x)\ dx.$$

Next we note by letting $f = f(X,\beta)$ in the TBS model with family of transformations $h(y,\lambda)$ and error distribution F_ϵ (and density f_ϵ), that

$$\begin{aligned} F(x,\theta) &= P(r(\lambda,\beta,Z) \le \sigma x\ ) = P(\ h(Y,\lambda) - h(f,\lambda) \le \sigma x\ ) \\ &= P(\ h(h^{-1}(h(f_o,\lambda_o)+\sigma_o\epsilon,\lambda_o),\lambda) - h(f,\lambda) \le \sigma x\ ) \\ &= P(\ h^{-1}(h(f_o,\lambda_o)+\sigma_o\epsilon),\lambda_o) \le h^{-1}(\sigma x + h(f,\lambda),\lambda)\ ) \\ &= P(\ \ \epsilon \le \{\ h(h^{-1}(\sigma x + h(f,\lambda),\lambda),\lambda_o) - h(f_o,\lambda_o)\ \}/\sigma_o\ ) \\ &= E_X F_\epsilon(\ \{\ h(h^{-1}(\sigma x + h(f,\lambda),\lambda),\lambda_o) - h(f_o,\lambda_o)\ \}/\sigma_o\ ) \end{aligned}$$

and hence

$$\begin{aligned} \frac{\partial}{\partial\sigma} F(x,\theta) = E_X\{(1/\sigma_o)f_\epsilon(\{\ h(h^{-1}(\sigma x + h(f,\lambda),\lambda),\lambda_o) - h(f_o,\lambda_o)\ \}/\sigma_o)\times \\ \dot{h}(h^{-1}(\sigma x + h(f,\lambda),\lambda),\lambda_o)\ \dot{h}^{-1}(\sigma x + h(f,\lambda),\lambda)\ x. \end{aligned}$$

The second factor above simplifies considerably using the fact that when the indicated derivatives exist for any function $f(x)$,

$$(f^{-1})'(x) = [f'(f^{-1}(x))]^{-1}.$$

From this we obtain

$$\frac{\partial}{\partial\sigma} F(x,\theta)\Big|_{\theta=\theta_o} = E_X[(1/\sigma_o)f_\epsilon(x)x],$$

showing that if f_ϵ is a symmetric density,

$$(4.13) \qquad \frac{\partial}{\partial\sigma} M(x,\theta_o) = \frac{\partial}{\partial\sigma} F(x,\theta_o) + \frac{\partial}{\partial\sigma} F(-x,\theta_o) = 0$$

and

$$(4.14) \qquad M(x,\theta_o) = 0.$$

Equations (4.12), (4.13), and (4.14) show that $A_{13}^k(\theta_o) = 0$ for all k, and this completes the proof.

■

4.3 The Single-sample Model with a Modified Power Transformation with Shift

We define the *modified power transformation with shift* by

$$(4.15) \qquad h(y,\mu,\lambda) = \begin{cases} [(y-\mu)^\lambda - 1]/\lambda & \lambda \neq 0 \\ \log(y-\mu) & \lambda = 0 \end{cases}$$

for $-\infty < \mu, \lambda < \infty$. In section 3.7 we discussed why this family of transformations is suitable when there is the need for accounting for non-positive observations of the response. We also cited that conventional estimation methods such as maximum likelihood are unable to properly estimate the shift parameter, whereas minimum distance produces consistent estimates.

Next, we present the results of computations for the matrices A and B of Theorem 4.9 under a simple one-sample model with the shifted power transformation. If μ is restricted to be zero, we obtain the Box Cox transformation and the one-sample model considered also by Taylor (1985), and Hinkley (1975, 1977). This model will be the basis for a simulation study in Chapter V, where we vary error densities, estimators, and weights used in minimum distance estimation.

Proposition 4.12: Consider the one-sample model

$$h(Y_i,\mu,\lambda) = h(\beta,\mu,\lambda) + \sigma\epsilon \tag{4.16}$$

where h is given by (4.15), ϵ is assumed to have a continuous density $f_\epsilon(\cdot)$ which is symmetric on [-M, M], and $Var(\epsilon) = 1$.

We adopt the notation $\theta = (\mu,\lambda,\beta)$,

$$A = \begin{bmatrix} A_{\mu\mu} & A_{\mu\lambda} & A_{\mu\beta} \\ A_{\lambda\mu} & A_{\lambda\lambda} & A_{\lambda\beta} \\ A_{\beta\mu} & A_{\beta\lambda} & A_{\beta\beta} \end{bmatrix},$$

and a similar notation for B. We assume (4.16) holds for some value $\theta_o = (\mu_o,\lambda_o,\beta_o)$. Also, we define the following:

$$F_\mu(x) = f_\epsilon(x)\ \sigma_o^{-1}\{\ [(\beta_o - \mu_o)^{\lambda_o} + \lambda_o\sigma_o x]^{1-1/\lambda_o} - (\beta_o - \mu_o)^{\lambda_o - 1}\ \},$$

$$F_\lambda(x) = f_\epsilon(x)\ \sigma_o^{-1}\ [(\beta_o - \mu_o)^{\lambda_o} + \lambda_o\sigma_o y]$$
$$\times \Big\{ \lambda_o^{-1}(\sigma_o x + (\beta_o - \mu_o)^{\lambda_o} \log(\beta_o - \mu_o)) \,/\, ((\beta_o - \mu_o)^{\lambda_o} + \lambda_o\sigma_o x)$$
$$- \lambda_o^{-2}\log[(\beta_o - \mu_o)^{\lambda_o} + \lambda_o\sigma_o x] \Big\},$$

$$F_\beta(x) = f_\epsilon(x)\ \sigma_o^{-1}\ (\beta_o - \mu_o)^{\lambda_o - 1},$$

$$P_\mu(x) = F_\mu(x) + F_\mu(-x),$$

$$P_\lambda(x) = F_\lambda(x) + F_\lambda(-x),\ \text{and}$$

$$P_\beta(x) = F_\beta(x) + F_\beta(-x).$$

We then have

$$n^{1/2}(\hat{\theta}_{CVM} - \theta_o) \xrightarrow{d} N(0,\ 4A^{-1}BA^{-T})$$

where:

$$A_{\mu\mu} = 4 \int_0^M P_\mu^2(x)\ w(x)\ dx,$$

$$A_{\mu\lambda} = 4 \int_0^M P_\lambda(x)\ P_\mu(x)\ w(x)\ dx,$$

$$A_{\mu\beta} = 4 \int_0^M P_\beta(x)\ P_\mu(x)\ w(x)\ dx,$$

$$A_{\lambda\mu} = 4 \int_0^M P_\mu(x)\ P_\lambda(x)\ w(x)\ dx,$$

$$A_{\lambda\lambda} = 4 \int_0^M P_\lambda^2(x)\ w(x)\ dx,$$

$$A_{\lambda\beta} = 4 \int_0^M P_\beta(x)\ P_\lambda(x)\ w(x)\ dx,$$

$$A_{\beta\mu} = 2(\beta_o - \mu_o)^{\lambda_o - 1}\left\{ \int_{-M}^M [(\beta_o - \mu_o)^{\lambda_o} + \lambda_o\sigma_o y]^{1-1/\lambda_o}\ f_\epsilon(y)\ dy - (\beta_o - \mu_o)^{\lambda_o - 1} \right\},$$

$$A_{\beta\lambda} = -2(\beta_o - \mu_o)^{\lambda_o}\ \lambda_o{}^{-1}$$

$$\times \left\{ \int_{-M}^M [(\beta_o - \mu_o)^{\lambda_o} + \lambda_o\sigma_o y]\ \log[(\beta_o - \mu_o)^{\lambda_o} + \lambda_o\sigma_o y]^{1/\lambda_o}\ f_\epsilon(y)\ dy - (\beta_o - \mu_o)^{\lambda_o}\ \log(\beta_o - \mu_o) \right\},$$

$$A_{\beta\beta} = 2(\beta_o - \mu_o)^{2(\lambda_o - 1)},$$

$$B_{\mu\mu} = 8 \int_0^M \{ \int_0^y P_\mu(x)\ w(x)\ dx \}^2\ f_\epsilon(y)\ dy,$$

$$B_{\mu\lambda} = B_{\lambda\mu} = 8 \int_0^M \{\int_0^y P_\lambda(x)\ w(x)\ dx\}\ \{\int_0^y P_\mu(x)\ w(x)\ dx\}\ f_\epsilon(y)\ dy,$$

$$B_{\mu\beta} = B_{\beta\mu} = 4\sigma_o(\beta_o - \mu_o)^{\lambda_o - 1} \int_0^M y\ \{\int_0^y P_\mu(x)\ w(x)\ dx\}\ f_\epsilon(y)\ dy,$$

$$B_{\lambda\lambda} = 8 \int_0^M \{ \int_0^y P_\lambda(x)\ w(x)\ dx \}^2 f_\epsilon(y)\ dy,$$

$$B_{\lambda\beta} = B_{\beta\lambda} = 4\sigma_o(\beta_o - \mu_o)^{\lambda_o - 1} \int_0^M y \{ \int_0^y P_\lambda(x)\ w(x)\ dx \} f_\epsilon(y)\ dy,$$

and

$$B_{\beta\beta} = \sigma_o^2\ (\beta_o - \mu_o)^{2(\lambda_o - 1)}.$$

<u>proof</u>: Omitted. It follows from lengthy algebra and calculus, starting from the expressions obtained in Theorem 4.9.

■

We may also use these expressions for obtaining the asymptotic covariance matrix of $n^{1/2}(\hat{\lambda}, \hat{\beta})$ if the transformation is the one-parameter Box-Cox modified power transformation without the shifting. In this case, we simply set $\mu_o = 0$ and let

$$A = \begin{bmatrix} A_{\lambda\lambda} & A_{\lambda\beta} \\ A_{\beta\lambda} & A_{\beta\beta} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{\lambda\lambda} & B_{\lambda\beta} \\ B_{\beta\lambda} & B_{\beta\beta} \end{bmatrix}.$$

Likewise, when estimating the shift μ for fixed known power λ_o, we obtain the asymptotic covariance matrix of $n^{1/2}(\hat{\mu}, \hat{\beta})$ by setting

$$A = \begin{bmatrix} A_{\mu\mu} & A_{\mu\beta} \\ A_{\beta\mu} & A_{\beta\beta} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{\mu\mu} & B_{\mu\beta} \\ B_{\beta\mu} & B_{\beta\beta} \end{bmatrix}.$$

Since no mention has been made about estimating the matrix $4A^{-1}BA^{-T}$, we note in short that the expressions in Proposition 4.12 provide one feasible way of estimating A and B, and hence an estimate of the asymptotic covariance of $n^{1/2}\hat{\theta}_{CVM}$ may be found by $4\hat{A}^{-1}\hat{B}\hat{A}^{-T}$. It is possible to replace $f_\epsilon(y)$ for a symmetric density estimate $\hat{f}_\epsilon(y)$ of the residuals $\{ r_i^*(\hat{\lambda}_{CVM}) \}$ and substitute θ_0 for $\hat{\theta}_{CVM}$. If $\hat{f}_\epsilon(y) \xrightarrow{a.s.} f_\epsilon(y)$ uniformly in y, we obtain that $\hat{A}$ and $\hat{B}$ are consistent estimators of A and B. Other methods for estimating A and B could and should also be explored in the future.

It was interesting to take into consideration the shifted power transformations in the first place, because of the reasons given in section 3.7. In addition, it is interesting to explore the estimability of the (μ,λ) parameter by a consistent method, as when using minimum distance. In actuality, estimating μ and λ simultaneously is a very hard problem. One example should suffice to illustrate the difficulties that we encounter when estimating (μ,λ) jointly. With this intention, we consider model (4.16) under the two-parameter transformation (4.15). By definition, the estimate of the transformation parameter (μ,λ) is obtained by minimizing the quantity $V_n^2(\mu,\lambda)$ (which was defined in N8 of Notation 4.6) over two-dimensional (μ,λ) space. A mere glance at the immense order of magnitudes that result on the diagonal elements of the matrix $4A^{-1}BA^{-T}$, would show that estimating (μ,λ) is a difficult problem. However, we consider a graphic explanation of this fact instead, which perhaps better clarifies the nature of the problem. If n is large, for each fixed (μ,λ), $V_n^2(\mu,\lambda)$ is an approximation to its expected value (the limit of $V_n^2(\mu,\lambda)$); thus, a large n gives the general picture. The usefulness of $V_n^2(\mu,\lambda)$ for estimating (μ,λ) is related to the location of the minimum of $V_n^2(\mu,\lambda)$ with respect to the true parameter values, and how discriminating the minimum really is with respect to surrounding values of (μ,λ). With this notion in mind, we examine contour plots of the quantity

$$(4.17) \qquad T_n(\mu,\lambda) =: \log \{ (n^2/2)\, V_n^2(\mu,\lambda) \}$$

as a function of (μ,λ). Here, the log function is used solely to make graphics more well-behaved and to improve resolution. One example was selected to be studied closely when the distribution of errors is standard normal, by setting the parameters at $n = 250$, $\beta_0 = 16$, $\mu_0 = 0$, $\sigma_0 = 0.25$; while λ_0 was allowed to take values -0.25 (Figure 4.1), 0.25 (Figure 4.2), and 1.25 (Figure 4.3).

As we examine the contour plots, we discover peculiar "topographic" features in the vicinity of the true (μ,λ) pair, depending on the true value of λ: Long narrow gorges, wide flat valleys, or virtually flat plains. As the regions plotted in the figures show a wide range of values for both λ and μ, in any case they tend to show that $V_n^2(\mu,\lambda)$ is unable to pick out departures from the true values, unless the sample size was drastically increased. Under this pictorial representation, consistency means that eventually, the plots conform so that they all have a unique minimum at the true value; a well-determined estimate would occur when the true value lies at the bottom of a pronounced and rounded "hole".

All of this exposition translates into the fact that (μ,λ) cannot be discerned unless the sample size is very large (perhaps unattainable); it also means that the (μ,λ) parameterization (4.15) of the shifted power transformations is ineffective. This strongly suggests that reparameterizations of the shifted power need to be investigated. We also note that the one-sample model which we considered does not have any heteroscedasticity since there is no variation in the median. Symmetry considerations alone seem to be unable to pinpoint (μ,λ) in this case. In the more general model, the inclusion of a device which also measures heteroscedasticity might possibly improve the situation.

4.4 Figures

Figures for Chapter IV follow.

Figure 4.1: Contour plot of $T_n(\mu,\lambda) = \log\{(n^2/2)\, V_n^2(\mu,\lambda)\}$, when $n = 250$, $\beta_o = 16$, $\mu_o = 0$, $\lambda_o = -0.25$, and $\sigma_o = 0.25$ in model (4.16) under the modified power transformation with shift (4.15). The error density is standard normal.

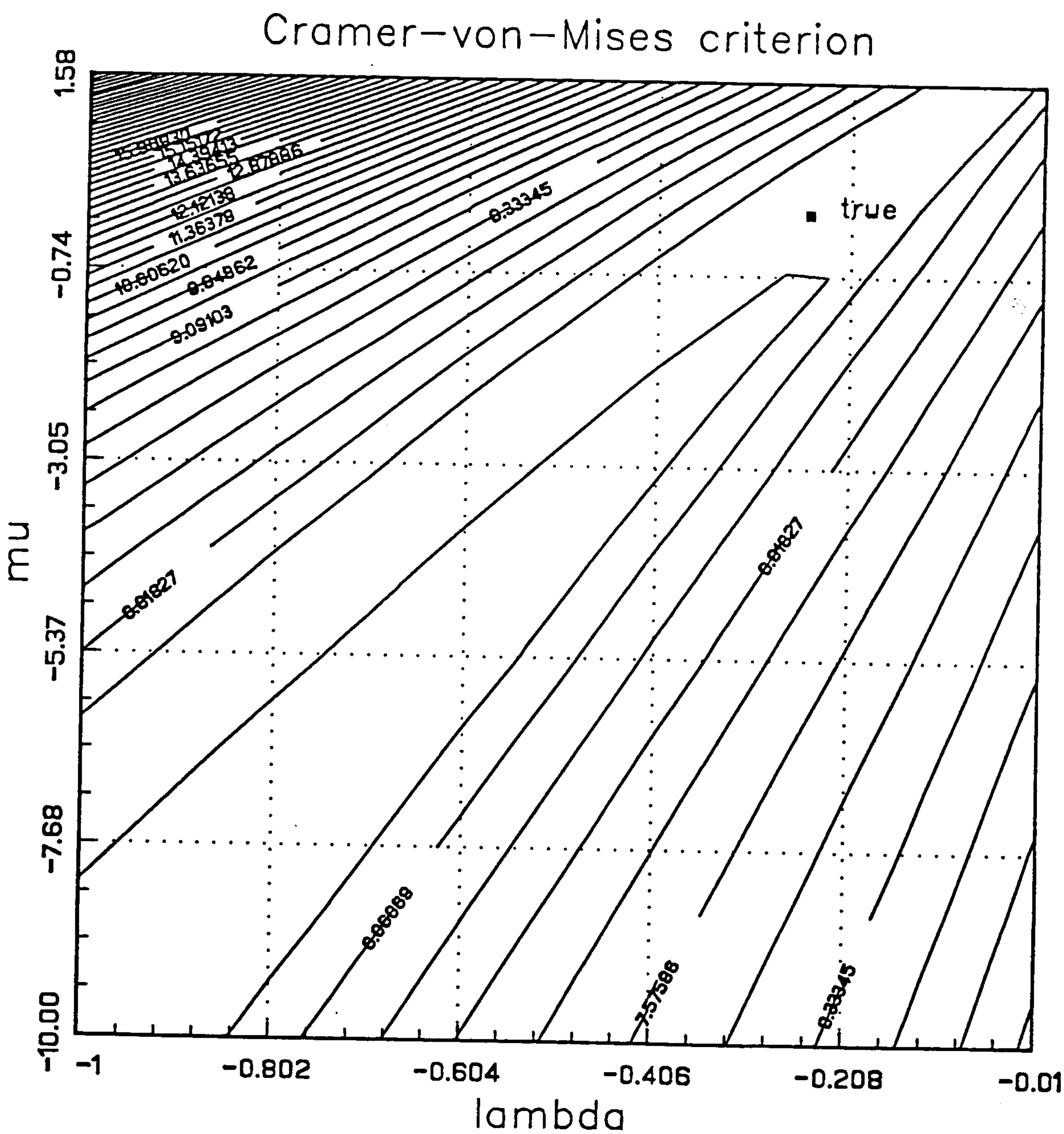


Figure 4.2: Contour plot of $T_n(\mu,\lambda) = \log\{(n^2/2)\,V_n^2(\mu,\lambda)\}$, when $n = 250$, $\beta_0 = 16$, $\mu_0 = 0$, $\lambda_0 = 0.25$, and $\sigma_0 = 0.25$ in model (4.16) under the modified power transformation with shift (4.15). The error density is standard normal.

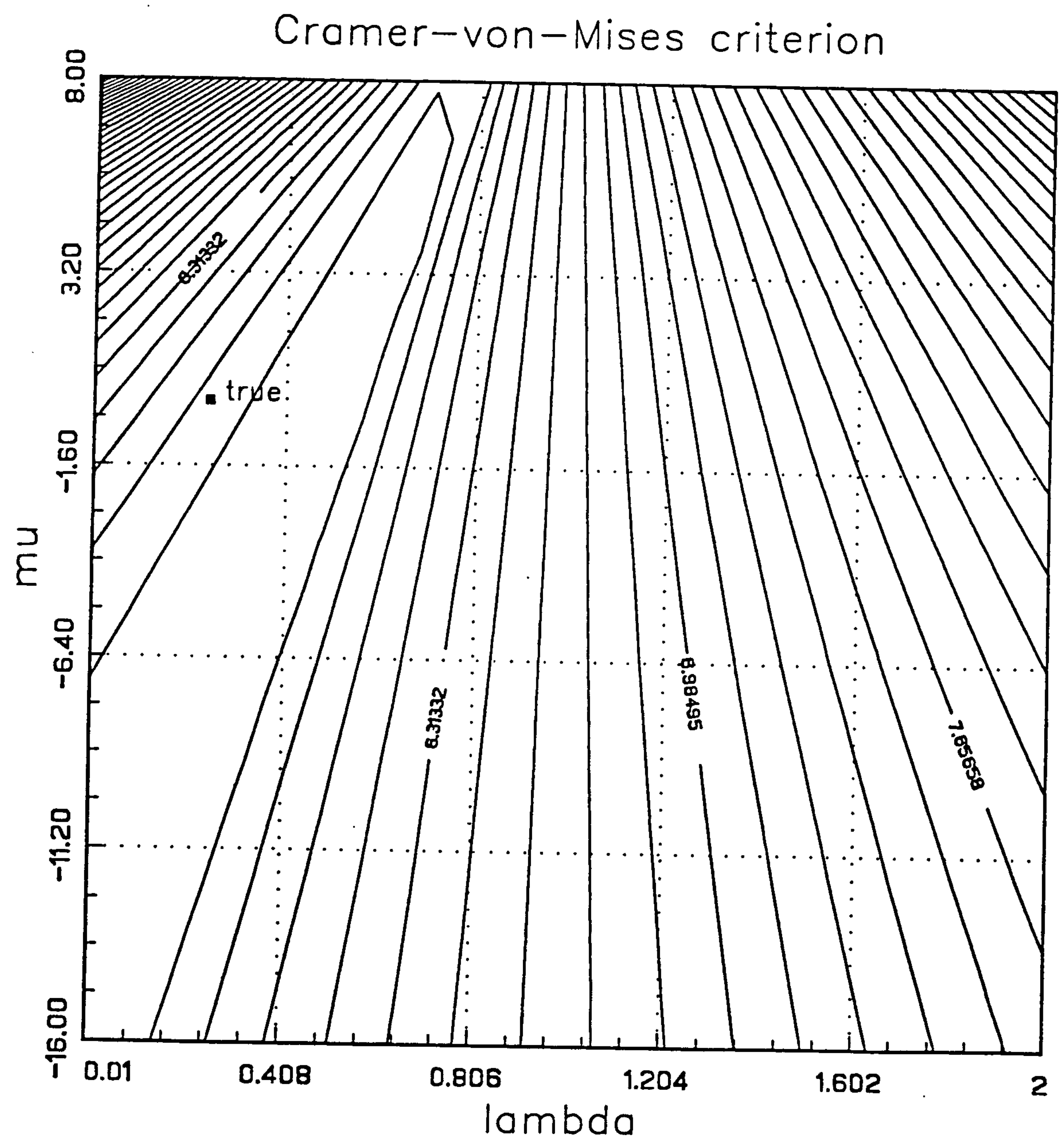
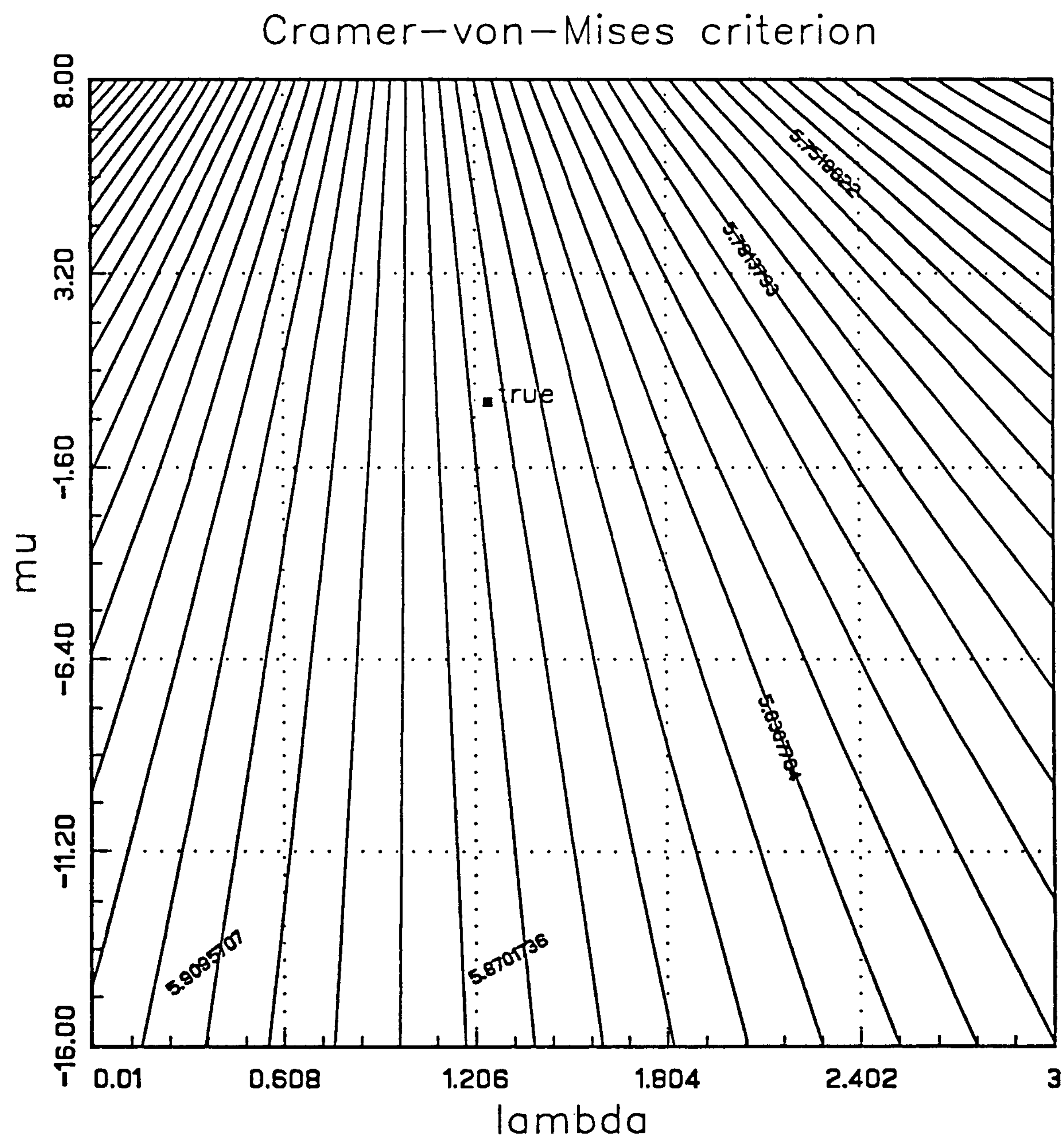


Figure 4.3: Contour plot of $T_n(\mu,\lambda) = \log \{ (n^2/2)\ V_n^2(\mu,\lambda) \}$, when $n = 250$, $\beta_o = 16$, $\mu_o = 0$, $\lambda_o = 1.25$, and $\sigma_o = 0.25$ in model (4.16) under the modified power transformation with shift (4.15). The error density is standard normal.



CHAPTER V

WEIGHTS IN MINIMUM DISTANCE ESTIMATION, AND A MONTE CARLO STUDY

5.0 Introduction

In the definition of the minimum distance estimator introduced in Chapter III, we considered the empirical distribution function (e.d.f.) of standardized residuals, and sought to minimize a criterion that measures symmetry. Explicitly, denoting by $F_n(x,\lambda)$ the e.d.f. of standardized residuals, the quantity to be minimized is

$$V_n^2(\lambda) = \int \{ F_n(x,\lambda) + F_n(-x,\lambda) - 1 \}^2 \, w(x) \, dx \tag{5.1}$$

where the weight function w(x) was assumed symmetric and non-negative. The weight was arbitrary, but considered fixed throughout Chapters III and IV. When the weight was expressible as the derivative of an odd, nondecreasing function, then the computation of (5.1) for arbitrary λ is simplified (Proposition 3.4).

In the present chapter, we wish to explore possible forms of the weight function so that a better estimator of the transformation parameter is obtained. This matter is the subject of section 5.1, where we consider an intuitive argument that points in the direction of the weight $\exp(x^2/2)$ if the error distribution is standard normal and the transformation is the modified power. The remaining part of this chapter summarizes the results of a Monte Carlo study performed on the minimum distance estimator. The purposes of this study are to evaluate the effect of different weights, to attain asymptotic theory, and to explore the performance of the estimate for small sample sizes.

5.1 A Heuristic Argument for an Optimal Weight

In Theorem 4.9 we established that the covariance of the limiting distribution of the minimum distance estimator is given by

$$4A^{-1}BA^{-T}$$

where

$$A = \frac{\partial}{\partial\theta} E[\ \Psi(Z_1, Z_2, \theta)\ ]_{\theta=\theta_o},$$

$$B = \int \{ \int \Psi(z_1, z_2, \theta_o)\ dP(z_2)\ \}^2\ dP(z_1).$$

and Ψ is defined by equations (4.7).

While exploring the effects of different weights on the estimates of λ, intuitively we expect at least one property that the weight w should possess in order to improve the estimates. To see this notion, we consider the quantity

$$Q_n(x,\lambda) = F_n(x,\lambda) + F_n(-x,\lambda) - 1. \tag{5.2}$$

For any set value of λ, we see from properties of distribution functions, that when the absolute value of x increases, the value of $Q_n(x,\lambda)$ tends to zero. This behavior suggests that we should upweight for large values of $|x|$ in order to increase the sensitivity of Q_n to departures of λ from its true value. In this way, a weight with heavier values at the tails is expected to yield more efficient estimates of λ.

Although the most general problem is finding the weight which is optimal for a given error distribution, it would be interesting as a preliminary step to discover the form of an optimal weight in the case that errors are normal. The use of such weight would insure that we obtain an efficient estimator whenever the errors happened to be normal, but we would need to evaluate the behavior of the identified weight when departures from normality occur. If the weight is fixed we obtain consistent estimators of λ under any symmetric (not necessarily normal) error density, but it is plausible that one weight will not perform uniformly well for all error densities.

We will concentrate on the case in which the parameter λ is one-dimensional, and the transformation is the modified power family of Box and Cox which we denoted by $\{ y^{(\lambda)}\}$. The model to be considered is then

$$(5.3) \qquad Y^{(\lambda)} = f^{(\lambda)}(X,\beta) + \sigma\epsilon .$$

We assume ϵ possesses a symmetric distribution $F_\epsilon(\cdot)$ and a density $f_\epsilon(\cdot)$. For the remaining part of this section, we also assume that the parameters β and σ are known, and denote

$$\mu = \mu(X) = f(X,\beta).$$

The model gives rise to the family of distribution functions for Y, indexed by λ, given by

$$(5.4) \qquad F_Y(y,\lambda) = P(Y \le y)$$

$$= P((Y^{(\lambda)}- \mu^{(\lambda)})/\sigma \le (y^{(\lambda)} - \mu^{(\lambda)})/\sigma) = F_\epsilon((y^{(\lambda)} - \mu^{(\lambda)})/\sigma)$$

and corresponding densities

(5.5) $$f_\lambda(y) = \sigma^{-1} f_\epsilon((y^{(\lambda)} - \mu^{(\lambda)})/\sigma)\; y^{\lambda-1}.$$

We assume i.i.d. observations $Y_1, \ldots, Y_n$ following model (5.3) and we let

(5.6) $$F_n(y,\lambda) = \text{e.d.f.}\{\ (Y_i^{(\lambda)} - \mu^{(\lambda)})/\sigma\ \}_{i=1}^{n}$$

and

(5.7) $$F_n(y) = \text{e.d.f.}\{\ Y_i\ \}_{i=1}^{n}.$$

We note that

(5.8) $$F_n(z,\lambda) = F_n[\ (\mu^\lambda + \lambda\sigma z)^{1/\lambda}\ ].$$

By definition, the minimum distance estimator $\hat{\lambda}_{CVM}$ minimizes

(5.9) $$V_n^2(\lambda) = \int \{\ F_n(y,\lambda) + F_n(-y,\lambda) - 1\ \}^2\ w(y)\ dy$$

for a fixed function w(y). However, we shall now allow w(y) to depend on λ in the minimization, that is, we let $w(y) = w(x,\lambda)$.

If, on the other hand, we were to define $\hat{\lambda}$ by minimizing

(5.10) $$W_n^2(\lambda) = \int \{\ F_n(y) - F_Y(y,\lambda)\ \}^2\ w^*(y,\lambda)\ dy,$$

for some given weight w^*, we would obtain an estimate of the type which has been examined by Parr and DeWet (1981). By considering an influence function argument, these authors show that in order to obtain an efficient estimate $\hat{\lambda}$, the form of the weight should be

(5.11) $$w^*(y,\lambda) = -\frac{\partial^2}{\partial\lambda\partial y}\log f_\lambda(y)\ \Big/\ \frac{\partial}{\partial\lambda} F_Y(y,\lambda).$$

By the change of variable $z = (y^{(\lambda)} - \mu^{(\lambda)})/\sigma = (y^\lambda - \mu^\lambda)/\sigma\lambda$ so that $y = (\mu^{\lambda} + \lambda\sigma z)^{1/\lambda}$ and $dy = (\mu^{\lambda} + \lambda\sigma z)^{1/\lambda-1}\sigma\, dz$ in (5.10), we can use (5.4) and (5.8) to obtain

$$(5.12)\qquad W_n^2(\lambda) = \int \{\ F_n(y) - F_\epsilon((y^{(\lambda)}-\mu^{(\lambda)})/\sigma)\ \}^2\ w^*(y,\lambda)\ dy$$

$$= \int \{\ F_n(z,\lambda) - F_\epsilon(z)\ \}^2\ w^*[(\mu^{\lambda}+\lambda\sigma z)^{1/\lambda},\lambda]\ (\mu^{\lambda}+\lambda\sigma z)^{1/\lambda-1}\sigma\ dz.$$

Next, we consider approximating $F_\epsilon(z)$ in (5.12). Since the error distribution was assumed symmetric, it seems natural to consider a symmetrization of the empirical distribution of the residuals, given by

$$(5.13)\qquad \hat{F}_\epsilon(z) = [F_n(z,\lambda) + (1 - F_n(-z,\lambda))\ ]\ /\ 2.$$

If we also define

$$(5.14)\qquad w(z,\lambda) = w^*[(\mu^{\lambda}+\lambda\sigma z)^{1/\lambda},\lambda]\ (\mu^{\lambda}+\lambda z)^{1/\lambda-1}\sigma,$$

then we will obtain that

$$(5.15)\qquad W_n^2(\lambda) \cong (1/4) \int \{\ F_n(z,\lambda) + F_n(-z,\lambda) - 1\ \}^2\ w(z,\lambda)\ dz.$$

By comparing with (5.9), we can see

$$(5.16)\qquad W_n^2(\lambda) \cong (1/4)\ V_n^2(\lambda).$$

Thus, the two estimators $\hat{\lambda}$ and $\hat{\lambda}_{CVM}$ are equivalent in the sense that one minimizes its defining criterion if and only if it also minimizes the opposing criterion. That is, the minimum distance estimator is approximately equal to an estimator of the Parr-DeWet type, and therefore an (approximately) optimal $w(z,\lambda)$ weight for the minimum distance estimator would be given by

(5.17) $$w(z,\lambda) = w^{*}[(\mu^{\lambda}+\lambda\sigma z)^{1/\lambda},\lambda]\ (\mu^{\lambda}+\lambda\sigma z)^{1/\lambda-1}\sigma,$$

where $w^{*}(z,\lambda)$ is chosen to be (5.11).

We acknowledge the fact that (5.17) was obtained by an important approximation, when F_{ϵ} was replaced by an estimate $\hat{F}_{\epsilon}$ in (5.12). So at best, this optimal weight would only be crude. There is yet a second reason why (5.17) should be taken with precaution, because it was obtained assuming that the value of the β parameter was known. When β is also estimated, the optimal weight which we seek may be different; the problem is understandably more complex in this case.

There is no reason why (5.17) should give rise to a symmetric weight, and furthermore, it may be possible that (5.17) varies greatly with λ so that the weight is not a function of y only, in (5.9). Next, we compute the function $w(z,\lambda)$ in the case that errors ϵ are standard normal. Consider (5.4) and (5.5) which hold for a general error distribution $F_{\epsilon}(\cdot)$ and error density $f_{\epsilon}(\cdot)$. If we assume

$$f_{\epsilon}(u) = (2\pi)^{-1/2} e^{-u^2/2}$$

then we obtain

$$\frac{\partial}{\partial y} \log f_{\epsilon}(\ (y^{(\lambda)} - \mu^{(\lambda)})/\sigma\ ) = -\sigma^{-2}(y^{(\lambda)} - \mu^{(\lambda)})\ y^{\lambda-1}$$

and

$$\frac{\partial^2}{\partial\lambda\partial y} \log f_{\epsilon}(\ (y^{(\lambda)} - \mu^{(\lambda)})/\sigma\ ) =$$
$$-\sigma^{-2}[\ (y^{(\lambda)} - \mu^{(\lambda)})\ y^{\lambda-1}\log(y) + y^{\lambda-1}\frac{\partial}{\partial\lambda}\ (y^{(\lambda)} - \mu^{(\lambda)})\ ].$$

Substituting, obtain using (5.11),

(5.18) $w^{*}(y,\lambda) =$

$$- \frac{\partial^2}{\partial\lambda\partial y} \{ \log f_\epsilon((y^{(\lambda)}-\mu^{(\lambda)})/\sigma) + (\lambda-1) \log(y) \}$$

$$\frac{\qquad}{f_\epsilon((y^{(\lambda)}-\mu^{(\lambda)})/\sigma) \sigma^{-1}\frac{\partial}{\partial\lambda} (y^{(\lambda)}-\mu^{(\lambda)})}$$

$$= \frac{\sigma^{-2}(y^{(\lambda)}-\mu^{(\lambda)}) y^{\lambda-1} \log(y) + \sigma^{-2}y^{\lambda-1} \frac{\partial}{\partial\lambda} (y^{(\lambda)}-\mu^{(\lambda)}) - 1/y}{f_\epsilon((y^{(\lambda)}-\mu^{(\lambda)})/\sigma) \sigma^{-1}\frac{\partial}{\partial\lambda} (y^{(\lambda)}-\mu^{(\lambda)})}.$$

Next we find an expression for $\frac{\partial}{\partial\lambda} (y^{(\lambda)}- \mu^{(\lambda)})$.

$$\frac{\partial}{\partial\lambda} (y^{(\lambda)}- \mu^{(\lambda)}) = \frac{\partial}{\partial\lambda} ((y^\lambda- \mu^\lambda)/\lambda)$$

$$= \{ \lambda(y^\lambda\log(y) - \mu^\lambda\log(\mu)) - (y^\lambda- \mu^\lambda) \} / \lambda^2$$

which evaluated at $y_o = (\mu^\lambda + \lambda\sigma z)^{1/\lambda}$ becomes

$$(5.19) \qquad D(z,\lambda) =: \frac{\partial}{\partial\lambda} (y^{(\lambda)}- \mu^{(\lambda)})\Big|_{y_o} =$$

$$\{ \lambda((\mu^\lambda + \lambda\sigma z)(1/\lambda) \log(\mu^\lambda + \lambda\sigma z) - \mu^\lambda\log(\mu)) - \lambda\sigma z \} / \lambda^2.$$

Using (5.18) and (5.19),

$$w^*[(\mu^\lambda + \lambda\sigma z)^{1/\lambda},\lambda] =$$

$$(2\pi)^{1/2} e^{z^2/2} \Big\{ \sigma^{-1}[z(\mu^\lambda + \lambda\sigma z)^{(\lambda-1)/\lambda}\log(\mu^\lambda + \lambda\sigma z)^{1/\lambda}] / D(z,\lambda)$$
$$- \sigma(\mu^\lambda + \lambda\sigma z)^{-1/\lambda} / D(z,\lambda)$$
$$+ \sigma^{-1}(\mu^\lambda + \lambda\sigma z)^{(\lambda-1)/\lambda} \Big\}.$$

Finally, using (5.17), we obtain

$$w(z,\lambda) = w^*[(\mu^\lambda + \lambda\sigma z)^{1/\lambda},\lambda] (\mu^\lambda + \lambda\sigma z)^{1/\lambda-1}\sigma = (2\pi)^{1/2} e^{z^2} \times$$
$$\{ [z \log(\mu^\lambda + \lambda\sigma z)^{1/\lambda}] / D(z,\lambda) - \sigma^2[(\mu^\lambda + \lambda\sigma z)^{-1}] / D(z,\lambda) + 1 \}.$$

If we denote

$$(5.20) \qquad T_\lambda(z) =$$

$$(2\pi)^{1/2}\{ [z \log(\mu^\lambda + \lambda\sigma z)^{1/\lambda}] / D(z,\lambda) - \sigma^2[(\mu^\lambda + \lambda\sigma z)^{-1}] / D(z,\lambda) + 1\},$$

the result we have obtained is that

$$w(z,\lambda) = e^{z^2/2}\, T_\lambda(z). \tag{5.21}$$

In this way, it is the reciprocal of the standard normal density which produces the term $\exp(x^2/2)$ that appears as a leading term for the optimal weight. The function $\exp(x^2/2)$ is appealing as a weight because of its smoothness and because it upweights at the tails in a way that was intuitively foreseen for an optimal weight. In order to justify using $w(x) = \exp(x^2/2)$ in (5.1), it is necessary to compare $w(z,\lambda)$ and $\exp(z^2/2)$, or equivalently, to check that $T_\lambda(z)$ is approximately constant for all (λ,z).

Various plots of $T_\lambda(z)$ for different values of λ have shown that its general behavior is not dependent on λ or σ. The weight $w(z,\lambda)$ obtained in this way has a peculiar feature in that that $w(0^-,\lambda) = \infty$ and $w(0^+,\lambda) = -\infty$. This fact seems troublesome at first, since $w(z,\lambda)$ resists proper interpretation around $z = 0$. However, we next see that this point has only a marginal effect. We consider the example $(\lambda,\mu,\sigma) = (.25, 12.25, 1)$, which shows the typical trend of the function $w(z,\lambda)$. Figure 5.1 shows a plot of the functions $\log\{w(z,\lambda)/6\}$ on $0.3 < |z| \leq 4$, and $z^2/2$ on $0 \leq |z| \leq 4$. We observe that, excluding the singularity at $z = 0$, both curves resemble each other over most of the range of z, and that $\log\{w(z,\lambda)/6\}$ is (unintentionally) almost a symmetric function. Now, if we consider the symmetrized version of $w(z,\lambda)$ given by $w_s(z,\lambda) = [w(z,\lambda) + w(-z,\lambda)]/2$, we find that the problem with the singularity at $z = 0$ is no longer present and the resemblance with $z^2/2$ is markedly stronger. Figure 5.2 is a simple plot which shows graphs of the functions $\log\{w_s(z,\lambda)/6\}$ and $z^2/2$ over the range

$0 \leq |z| \leq 4$; we notice that these two curves are virtually indistinguishable from each other. As a result of these findings, we will subsequently consider the weight $w(x) = \exp(x^2/2)$ as a likely candidate for minimum distance estimation under normal errors, and next proceed to entertain this $w(x)$ in a Monte Carlo study.

5.2 Monte Carlo Study

We intend to study minimum distance estimation in the single-sample TBS transformation model

$$h(Y_i, \lambda) = h(\beta, \lambda) + \sigma\epsilon_i \tag{5.22}$$

where the errors ϵ_i are symmetric with density f_ϵ. If h is the modified shifted power transformation, this is precisely the model we considered in Proposition 4.12. There, we computed the asymptotic covariance of the estimate that was derived from generalized M-estimation theory.

The densities f_ϵ which were counted were a standard normal truncated at ± 4, a standard normal truncated at ± 2, a uniform density, a triangular density, Student's t-distribution with 3 degrees of freedom truncated at ± 6, and a standard normal contaminated with probability .15 by a normal with mean zero and variance 5. In all cases, the densities were standardized so that they had mean zero and variance one.

The number of Monte Carlo replications was fixed at 1000 in each case; a selected case was defined by a fixed setting of the parameter $\theta = (\lambda, \beta, \sigma)$, the density f_ϵ, the sample size n, and the chosen estimating method. A replication then consisted in generating a sample following model (5.22) and then computing the estimator.

When dealing with the minimum distance estimate, the calculation requires minimization of a non-differentiable function in λ. The solution requires some numerical elaboration: The *Nelder-Mead simplex algorithm* (Nelder and Mead (1965)) was the numerical algorithm employed for this matter. This iterative procedure is based only on evaluations of the function itself and not of its derivatives, and requires a starting point which in this study was found by search over a grid of λ-values. Apart from the fact that this algorithm renders convergence that is painstakingly slow (especially in higher dimensions), it was otherwise effective for the problem at hand.

For comparative purposes, we shall consider three estimators of the parameter λ:

(a) $\hat{\lambda}_{CVM}$, the minimum distance estimate using weight function $w(x)$, of Chapter III.

(b) $\hat{\lambda}_{B\&C}$, the Maximum Likelihood estimate of Box and Cox (1964), which is the MLE computed under the assumption that f_ϵ is standard normal.

(c) $\hat{\lambda}_{R\&A}$, the Ruppert-Aldershof skewness estimator, of Chapter II.

This simulation study is organized into three main segments:

<u>Part A.</u> Here we consider minimum distance estimation under the simple power transform

$$h(y,\lambda) = y^{\lambda} \tag{5.23}$$

and a large sample size ($n = 120$) in order to compare with the asymptotics of Theorem 4.9, and to observe the action of a few selected weight functions $w(x)$. The parameter settings were chosen to be

$(\lambda,\beta,\sigma) = (.5,100,2)$ (Table 5.1), $(\lambda,\beta,\sigma) = (.5,100,1)$ (Table 5.2), and $(\lambda,\beta,\sigma) = (.5,100,.2)$ (Table 5.3).

Part B. Here we consider the modified power transformation of Box and Cox

$$(5.24) \qquad h(y,\lambda) = \begin{cases} (y^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$$

using a large sample size (n = 120). In order to perceive the effects of weight functions, several choices of the weight function $w(x)$ are examined when using $\hat{\lambda}_{CVM}$. We are particularly interested in observing the performance of the weight $w(x) = \exp(x^2/2)$ obtained in section 5.1, but other weights considered for experimentation were the unit weight $w(x)\equiv 1$, and the power weights $w(x) = x^k$ for k=2,4, and 6. The other two estimates were also included in this part of the study. Since we are interested in assessing the performance of the estimate $\hat{\lambda}_{CVM}$ under normal errors, the case when f_{ϵ} is normal will be of special interest. We will be able to gauge the loss of efficiency when we compare against $\hat{\lambda}_{B\&C}$, which is asymptotically optimal in the normal case. Of course, when the errors are not normal then $\hat{\lambda}_{B\&C}$ is not necessarily a consistent estimate of λ_0. However, other studies have shown that $\hat{\lambda}_{B\&C}$ produces meaningful estimates even in non-consistent cases (Hernandez and Johnson (1980)). Parameter settings for Part B of the Monte Carlo study were $(\lambda,\beta,\sigma) = (-.24, .037432,2)$ (Table 5.4) and $(\lambda,\beta,\sigma) = (-.24, .104383,1)$ (Table 5.5).

Part C. As in Part B, in this third part we also consider the transformation (5.24), but we concentrate only on the minimum distance estimate that is obtained when using the weight $w(x) = \exp(x^2/2)$. Some comparisons between $\hat{\lambda}_{CVM}$, $\hat{\lambda}_{B\&C}$, and $\hat{\lambda}_{R\&A}$ under different error densities are provided, when the sample sizes are not quite as large as in parts A and B. For sample size n = 40, the selected parameter settings were $(\lambda,\beta,\sigma) = (.5, 12.25, .25)$ (Table 5.6), $(\lambda,\beta,\sigma) = (.5, 36, 1)$ (Table 5.7), $(\lambda,\beta,\sigma) = (1.5, 4.16497, 1)$ (Table 5.8), and $(\lambda,\beta,\sigma) = (.25, 25.6289, .25)$ (Table 5.9).

In tables, whenever an estimate $\hat{\theta} = (\hat{\lambda},\hat{\beta})$ satisfies

$$n^{\frac{1}{2}}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma)$$

we resort to the notation

$$\Sigma = \begin{bmatrix} \Sigma(\lambda,\lambda) & \Sigma(\lambda,\beta) \\ \Sigma(\lambda,\beta) & \Sigma(\beta,\beta) \end{bmatrix}.$$

5.3 Discussion

The one-sample model which we considered in the Monte Carlo study provides a first step towards the more general regression model, and several general conclusions with respect to minimum distance estimation can be drawn. An important issue is to decide on the weight function which should be employed for minimum distance estimation. For estimating a power parameter, we derived the weight $w(x) = \exp(x^2/2)$ as optimal for normal errors, and this weight was one of the main focal points in the simulation study.

In the tables that correspond to Part A, we generally observe agreement with respect to asymptotic theory. Severe computational difficulties were encountered in some of the asymptotics, particularly in the case of normal errors and minimum distance estimation using the unit weight. These problems are presumably due to the two-level numerical integrations that are present in Proposition 4.12; any numerical errors in the inner integrals are also carried over to the outer ones and compounded. As with any numerical calculations, we should be cautious about any lurking numerical instability; in this particular problem, the numerical difficulties were unquestionably present in some of the cases considered. The problems are very likely due to shortcomings in the intrinsic numerical integration routines that were employed. For this reason, while a few of the entries in Tables 5.1, 5.2 and 5.3 show discrepancies between asymptotic and simulation results that are in excess of 10%, we will not be too rigorous. It is of greater importance here, to notice relative behavior across different weights.

While looking exclusively at the variance of $\hat{\lambda}_{CVM}$, even at an early stage we begin observing from Tables 5.1, 5.2, and 5.3 that the weight $\exp(x^2/2)$ performs better than the other weights considered. This exponential weight is preferred at all error densities except the uniform, where the squared weight seems to do just as well. The unit weight performs badly in all cases and clearly should be disregarded for any estimation purposes; it appears imperative to upweight for values away from zero in order to induce estimates that are worthy.

In tables of Part B, we again observe that among different weight choices for $\hat{\lambda}_{CVM}$, the weight $\exp(x^2/2)$ is without doubt the one that gives the best estimates of λ. The few exceptions where it is not the optimal choice are where uniform or normal truncated at ± 2 densities are involved; these densities are the less likely to be interesting from a point of view of a real-life application. It is significant to observe that the $\exp(x^2/2)$ weight is performing well despite the fact that β has not been assumed to be known and is also being estimated in the Monte Carlo. We recall that in section 5.1, the weight $\exp(x^2/2)$

was obtained by an informal argument in which we assumed that β was known, and that the errors were standard normal. Since the previous findings only establish that the $\exp(x^2/2)$ weights are good within the family of minimum distance estimators with varying weights, there is the need for comparing optimal minimum distance estimation to other procedures.

Strictly speaking, since we are restricting the densities to be truncated, we did not consider a case with normal errors. A close approximation to normality is given by the standard normal density truncated at ± 4. In the normal case it is known that $\hat{\lambda}_{B\&C}$ is asymptotically optimal, but this estimator is not even consistent for other error distributions. This last fact is evident in some of the large bias terms in Tables 5.6-5.9, but we note that even for several non-normal cases, the bias for $\hat{\lambda}_{B\&C}$ was negligible. This agrees with the findings of Taylor (1980), and Hernandez and Johnson (1980), who showed that the Box-Cox estimate has an interpretation even in inconsistent cases. Even if minimum distance estimation is not expected to yield fully efficient estimators, it is certainly desirable to observe how it compares to $\hat{\lambda}_{B\&C}$ in the normal case, as a basis for judgment. This loss of efficiency is foreseen for minimum distance estimation because we only made the nonparametric assumption of symmetry with respect to the errors. In Tables 5.4 and 5.5, we see that when estimating the modified power transformation, the loss of efficiency for $\hat{\lambda}_{CVM}$ in the normal case is roughly of the order of 18% and 11% respectively. Comparison with the optimal in this case is merely a standard; the full potential of minimum distance estimation is called for in situations where other procedures fail to be consistent, for example $\hat{\lambda}_{B\&C}$ under non-normality, $\hat{\lambda}_{B\&C}$ when estimating a shift, or $\hat{\lambda}_{R\&A}$ under multi-dimensional λ.

An accessory remark that results from Part B is that the weight $w(x) = x^2$ appears to be a reasonable contender when we grant that computation of the estimate is much simplified when using x^2 instead of $\exp(x^2/2)$. For uniform errors, squared weights are evidently better; for other cases they are dominated by $\exp(x^2/2)$ weight, but not severely.

We detect no difference in overall behavior in Part C when the sample size is n = 40. That is, the estimates compare amongst themselves in much the same way as they did for n = 120. However, in this part, we considered a greater collection of error densities and therefore we can better assess the behavior of estimates when departures from normality occur. Here, the weight for minimum distance was fixed at $\exp(x^2/2)$, which was a weight designed to yield minimum variance for normal errors. While it would have been favorable to find that this weight is particularly efficient also in non-normal cases, there does not seem to be evidence that this is the case. It is self-evident, for example noting the bias terms in Table 5.8, how departures from normality influence $\hat{\lambda}_{B\&C}$; in order to note what the effects are on $\hat{\lambda}_{CVM}$, we must compare with $\hat{\lambda}_{R\&A}$, which is also consistent. In most instances reported in Tables 5.6 though 5.9, $\hat{\lambda}_{CVM}$ has a larger variance than $\hat{\lambda}_{R\&A}$ does. In a few cases they are behaving practically in the same way, while $\hat{\lambda}_{CVM}$ is superior particularly in the rather radical examples that involve uniform errors. This discovery exposes the need for investigating the form of an optimal weight for heavy-tailed error distributions (e.g. the t-distribution). The minimum distance estimator is based on full symmetry of the errors, but it seems that the wrong choice of a weight function causes $\hat{\lambda}_{CVM}$ to be comparable to $\hat{\lambda}_{R\&A}$ which is based only on skewness.

A natural question arises: What happens if we apply the heuristic argument for finding an optimal weight to a heavy-tailed error density? The answer involved only some straightforward calculations, but the result was hard to interpret. For a t-distribution, the argument yields a weight function which was not as clear-cut as the ones illustrated in Figures 5.1 and 5.2. The problem is that the result could not be expressed as a simple function of z; this property was desirable to simplify computation if simulations were to be performed. A possible solution is to approximate the weight by some rational function, say a symmetric polynomial, and then use the approximation as the weight. Since the original weight might initially be in error because of the approximations that were involved in its derivation, this issue was not elaborated upon. Thus an optimal weight for heavy-tailed distributions remains an open problem. It is not clear at this point what characteristics the optimal weight should have to accommodate heavy tails.

Conceivably, the next effort in further research on minimum distance estimation should be focused on estimating a shifted power transformation. As we remarked in Chapter IV, maximum likelihood is unable to cope with a shift parameter. Two related problems are estimating the shift for fixed power, and estimating the shift and the power simultaneously. Questions about choice of weights and/or possible alternate parameterizations remain to be answered.

Figure 5.1: Plot of $\log(w(z,\lambda)/6)$ for $0.3 < |z| \le 4$ (solid line) and $z^2/2$ for $0 \le |z| \le 4$ (broken line), when $(\lambda,\mu,\sigma)=(.25,12.25,1)$.

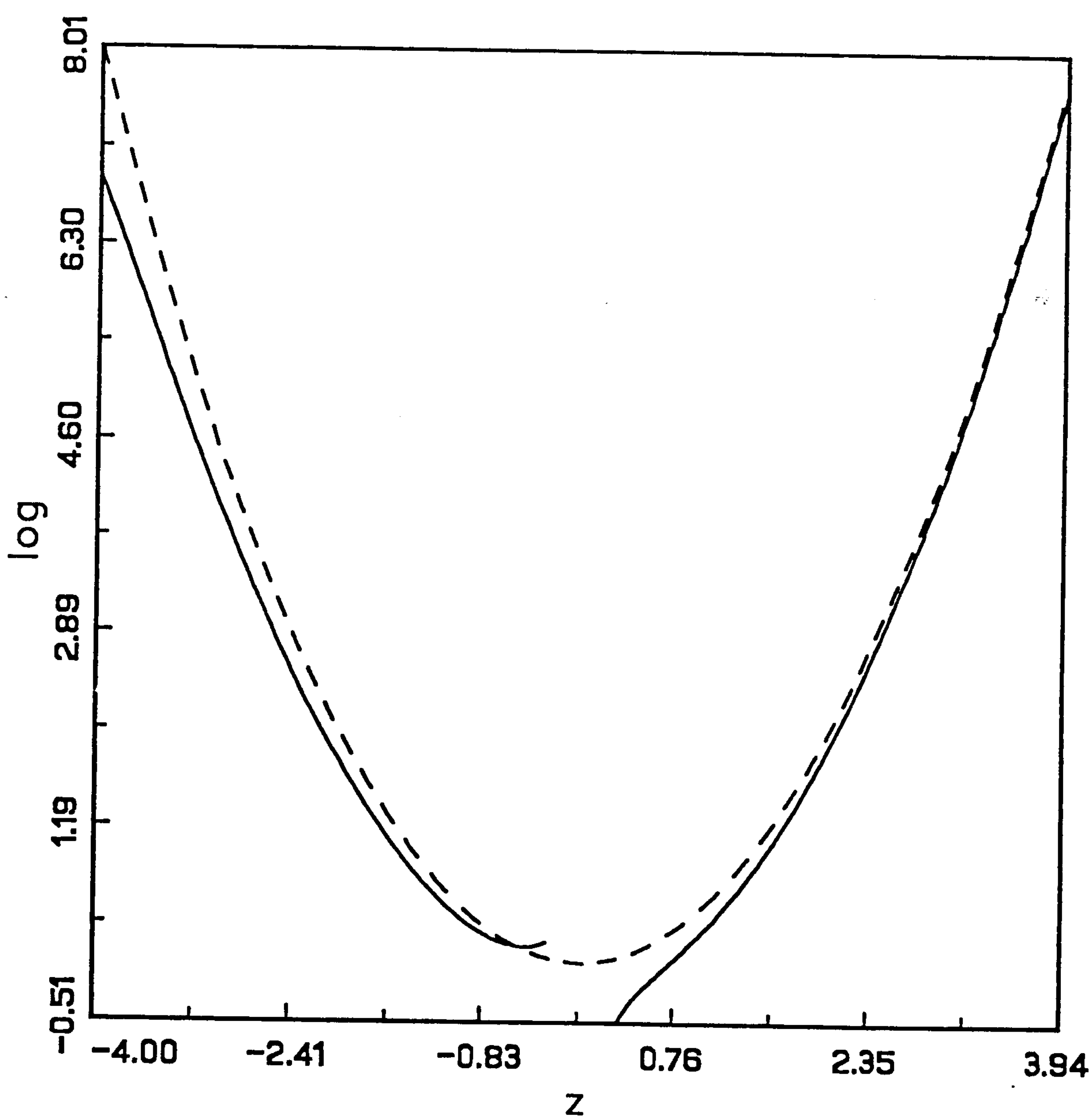


Figure 5.2: Plot of $\log(w_s(z,\lambda)/6)$ (solid line) and $z^2/2$ (broken line) for $0 \leq |z| \leq 4$ when $(\lambda,\mu,\sigma)=(.25,12.25,1)$.

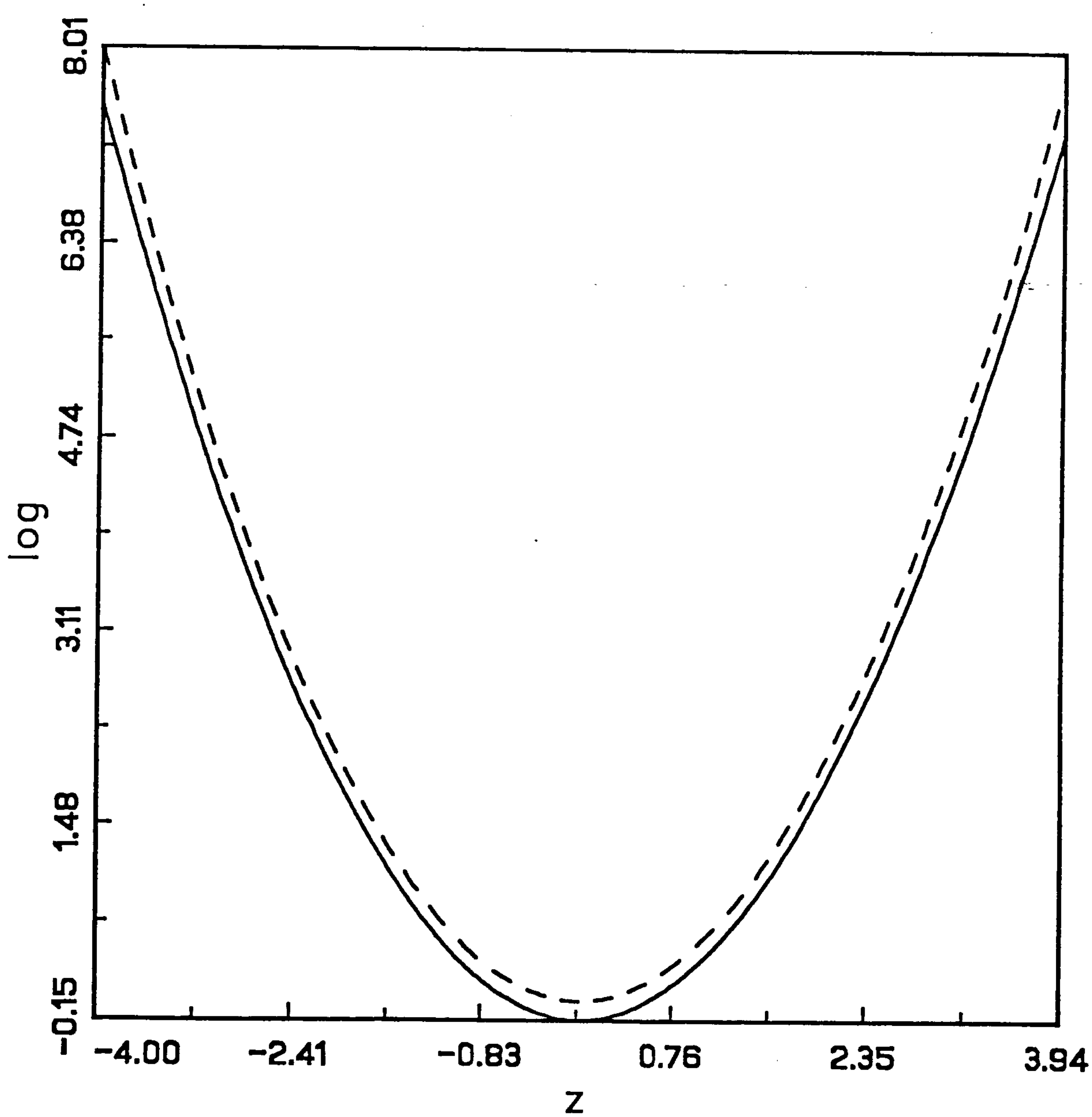


Table 5.1: Minimum distance estimation of $\theta=(\lambda,\beta)$ in model (5.22) with simple power transformation (5.23). Monte Carlo results (MC) using n=120, and asymptotic values (asy) of covariance matrix Σ when $\lambda_0=.5$, $\beta_0=100$, $\sigma_0=2$. (*** denotes an entry that had a numerical difficulty)

		ERROR DENSITY normal ±4	normal ±2	uniform	triang.
unit wt					
$\Sigma(\lambda,\lambda)$	MC	6.10	6.05	8.73	5.71
	asy	***	5.67	8.59	5.82
$\Sigma(\beta,\beta)$	MC	1890.73	2221.36	3376.73	2322.00
	asy	***	2381.51	3655.36	2308.15
$\Sigma(\lambda,\beta)$	MC	46.96	69.42	157.59	71.02
	asy	***	71.11	167.87	67.21
x^2 wt					
$\Sigma(\lambda,\lambda)$	MC	5.43	3.90	7.93	4.06
	asy	4.91	3.39	7.64	3.71
$\Sigma(\beta,\beta)$	MC	1938.23	2422.68	3630.97	2391.81
	asy	1930.30	2306.52	3690.33	2287.93
$\Sigma(\lambda,\beta)$	MC	41.13	60.33	160.00	61.26
	asy	40.30	57.23	159.99	57.38
$\exp(x^2/2)$ wt					
$\Sigma(\lambda,\lambda)$	MC	4.73	3.57	8.72	3.64
	asy	3.69	2.87	7.60	3.22
$\Sigma(\beta,\beta)$	MC	1894.55	2322.42	3775.01	2276.24
	asy	1854.55	2301.67	3687.49	2272.54
$\Sigma(\lambda,\beta)$	MC	38.66	58.09	171.39	54.87
	asy	30.67	54.83	159.65	54.43

Table 5.2: Minimum distance estimation of $\theta=(\lambda,\beta)$ in model (5.22) with simple power transformation (5.23). Monte Carlo results (MC) using n=120, and asymptotic values (asy) of covariance matrix Σ when $\lambda_0=.5$, $\beta_0=100$, $\sigma_0=1$. (*** denotes an entry that had a numerical difficulty)

		ERROR DENSITY			
		normal ± 4	normal ± 2	uniform	triang.
unit wt					
$\Sigma(\lambda,\lambda)$	MC	26.24	24.62	41.12	23.33
	asy	***	23.64	35.38	24.20
$\Sigma(\beta,\beta)$	MC	430.01	595.34	964.84	561.21
	asy	***	596.78	941.88	577.76
$\Sigma(\lambda,\beta)$	MC	37.38	74.14	182.38	66.79
	asy	***	72.74	170.55	68.56
x^2 wt					
$\Sigma(\lambda,\lambda)$	MC	19.73	15.92	31.71	16.97
	asy	20.64	14.18	31.51	15.47
$\Sigma(\beta,\beta)$	MC	478.55	542.46	853.15	573.72
	asy	483.86	578.34	926.13	573.39
$\Sigma(\lambda,\beta)$	MC	35.06	56.67	155.54	58.81
	asy	41.61	58.65	162.74	58.71
$\exp(x^2/2)$ wt					
$\Sigma(\lambda,\lambda)$	MC	18.36	14.79	33.39	14.25
	asy	15.87	12.05	31.34	13.46
$\Sigma(\beta,\beta)$	MC	444.47	573.84	926.99	552.16
	asy	466.02	577.33	925.44	569.75
$\Sigma(\lambda,\beta)$	MC	32.56	60.65	165.36	51.95
	asy	32.37	56.26	162.39	55.78

Table 5.3: Minimum distance estimation of $\theta=(\lambda,\beta)$ in model (5.22) with simple power transformation (5.23). Monte Carlo results (MC) using n=120, and asymptotic values (asy) of covariance matrix Σ when $\lambda_0=.5$, $\beta_0=100$, $\sigma_0=0.2$. (*** denotes an entry that had a numerical difficulty)

		ERROR DENSITY			
		normal ±4	normal ±2	uniform	triang.
unit wt					
$\Sigma(\lambda,\lambda)$	MC	663.39	520.08	889.38	550.35
	asy	***	598.68	892.53	612.21
$\Sigma(\beta,\beta)$	MC	19.04	20.24	34.87	23.60
	asy	***	23.88	37.71	23.11
$\Sigma(\lambda,\beta)$	MC	46.48	58.29	160.26	59.96
	asy	***	73.24	171.39	68.98
x^2 wt					
$\Sigma(\lambda,\lambda)$	MC	568.55	459.36	787.67	472.27
	asy	523.62	359.43	795.14	391.86
$\Sigma(\beta,\beta)$	MC	18.22	24.40	34.42	23.26
	asy	19.36	23.15	37.08	22.95
$\Sigma(\lambda,\beta)$	MC	40.75	72.84	155.03	62.60
	asy	42.00	59.08	163.60	59.12
$\exp(x^2/2)$ wt					
$\Sigma(\lambda,\lambda)$	MC	495.22	400.69	807.18	359.70
	asy	404.87	305.70	790.85	341.08
$\Sigma(\beta,\beta)$	MC	18.64	22.64	38.20	21.13
	asy	18.66	23.11	37.06	22.80
$\Sigma(\lambda,\beta)$	MC	38.16	62.45	165.17	52.18
	asy	32.86	56.70	163.25	56.19

Table 5.4: Monte Carlo results for covariance matrix Σ of $n^{1/2}(\hat{\theta} - \theta_0)$ for different estimators of $\theta=(\lambda,\beta)$, based on a samples of size n=120 when $\lambda_0=-.24$, $\beta_0=.037432$, $\sigma_0=2$ in the single-sample model (5.22) with the modified power transformation (5.24).

	ERROR DENSITY				
	normal ±4	normal ±2	uniform	triang.	Student-t
MD unit					
$\Sigma(\lambda,\lambda)$	1.164	0.937	2.074	1.031	1.530
$\Sigma(\beta,\beta)$	0.0014	0.0017	0.0029	0.0017	0.0010
$\Sigma(\lambda,\beta)$	0.018	0.023	0.072	0.024	0.003
MD x^2					
$\Sigma(\lambda,\lambda)$	0.889	0.806	1.457	0.763	1.354
$\Sigma(\beta,\beta)$	0.0015	0.0018	0.0026	0.0018	0.0010
$\Sigma(\lambda,\beta)$	0.014	0.026	0.058	0.024	0.004
MD x^4					
$\Sigma(\lambda,\lambda)$	0.915	0.573	1.499	0.640	1.222
$\Sigma(\beta,\beta)$	0.0014	0.0016	0.0027	0.0015	0.0011
$\Sigma(\lambda,\beta)$	0.013	0.020	0.060	0.0208	0.006
MD x^6					
$\Sigma(\lambda,\lambda)$	0.997	0.552	1.689	0.622	1.309
$\Sigma(\beta,\beta)$	0.0015	0.0017	0.0032	0.0016	0.0010
$\Sigma(\lambda,\beta)$	0.014	0.021	0.071	0.021	0.007
MD $\exp(\frac{1}{2}x^2)$					
$\Sigma(\lambda,\lambda)$	0.890	0.679	1.669	0.673	1.205
$\Sigma(\beta,\beta)$	0.0013	0.0017	0.0028	0.0015	0.0011
$\Sigma(\lambda,\beta)$	0.013	0.022	0.064	0.019	0.007
B&C MLE		←note: inconsistent for these errors→			
$\Sigma(\lambda,\lambda)$	0.754	0.500	0.920	0.544	1.320
$\Sigma(\beta,\beta)$	0.0014	0.0016	0.0025	0.0015	0.0010
$\Sigma(\lambda,\beta)$	0.011	0.017	0.044	0.017	0.003
R&A Skew.					
$\Sigma(\lambda,\lambda)$	0.806	0.722	1.772	0.686	1.117
$\Sigma(\beta,\beta)$	0.0014	0.0017	0.0028	0.0016	0.0010
$\Sigma(\lambda,\beta)$	0.012	0.023	0.065	0.020	0.003

Table 5.5: Monte Carlo results for covariance matrix Σ of $n^{1/2}(\hat{\theta} - \theta_0)$ for different estimators of $\theta=(\lambda,\beta)$, based on a samples of size n=120 when $\lambda_0=-.24$, $\beta_0=.104383$, $\sigma_0=1$ in the single-sample model (5.22) with the modified power transformation (5.24).

	ERROR DENSITY				
	normal ±4	normal ±2	uniform	triang.	Student-t
MD unit					
$\Sigma(\lambda,\lambda)$	2.499	2.880	4.653	2.417	4.099
$\Sigma(\beta,\beta)$	0.004	0.005	0.009	0.005	0.003
$\Sigma(\lambda,\beta)$	0.037	0.075	0.189	0.056	0.005
MD x^2					
$\Sigma(\lambda,\lambda)$	2.331	1.850	4.138	1.865	3.377
$\Sigma(\beta,\beta)$	0.005	0.005	0.009	0.005	0.003
$\Sigma(\lambda,\beta)$	0.037	0.061	0.188	0.060	0.006
MD x^4					
$\Sigma(\lambda,\lambda)$	2.306	1.655	4.144	1.718	3.242
$\Sigma(\beta,\beta)$	0.004	0.006	0.009	0.005	0.003
$\Sigma(\lambda,\beta)$	0.041	0.058	0.160	0.063	0.017
MD x^6					
$\Sigma(\lambda,\lambda)$	2.633	1.470	3.519	1.614	3.336
$\Sigma(\beta,\beta)$	0.004	0.005	0.009	0.005	0.003
$\Sigma(\lambda,\beta)$	0.047	0.064	0.166	0.059	0.020
MD $\exp(\frac{1}{2}x^2)$					
$\Sigma(\lambda,\lambda)$	2.214	1.700	4.055	1.920	3.467
$\Sigma(\beta,\beta)$	0.004	0.005	0.009	0.005	0.003
$\Sigma(\lambda,\beta)$	0.036	0.059	0.181	0.065	0.023
B&C MLE		←note: inconsistent for these errors→			
$\Sigma(\lambda,\lambda)$	1.990	1.460	2.433	1.361	3.298
$\Sigma(\beta,\beta)$	0.004	0.005	0.008	0.005	0.003
$\Sigma(\lambda,\beta)$	0.035	0.057	0.122	0.053	0.015
R&A Skew.					
$\Sigma(\lambda,\lambda)$	2.075	1.733	4.476	1.669	3.005
$\Sigma(\beta,\beta)$	0.004	0.005	0.009	0.005	0.003
$\Sigma(\lambda,\beta)$	0.038	0.060	0.186	0.059	0.004

Table 5.6: Monte Carlo results for selected estimates of λ when λ_0=.5, β_0=12.25, σ_0=.25, and n=40 in the single-sample model (5.22) under the modified power transformation (5.24).

errors ↓		ESTIMATOR $\hat{\lambda}_{CVM}$	$\hat{\lambda}_{B\&C}$	$\hat{\lambda}_{R\&A}$
Normal ±4	bias	.0137	.0165	-.0300
	var	4.936	3.908	4.545
Normal ±2	bias	.0347	-.1790	-.0446
	var	4.066	2.942	3.539
Student-t	bias	-.0311	.1319	-.0836
	var	7.116	6.789	7.257
Cont.Normal	bias	-.0461	-.0714	-.0562
	var	7.734	4.807	7.218
Uniform	bias	.0770	-.1315	.1078
	var	8.348	4.279	7.779

Table 5.7: Monte Carlo results for selected estimates of λ when $\lambda_0=.5$, $\beta_0=36$, $\sigma_0=1$, and $n=40$ in the single-sample model (5.22) under the modified power transformation (5.24).

errors ↓		ESTIMATOR $\hat{\lambda}_{CVM}$	$\hat{\lambda}_{B\&C}$	$\hat{\lambda}_{R\&A}$
Normal ± 4	bias	-.0103	-.0580	-.0577
	var	0.926	0.709	0.874
Normal ± 2	bias	-.0122	-.0713	.0020
	var	0.701	0.505	0.634
Student-t	bias	.0068	.0081	.0392
	var	1.4224	1.243	1.391
Cont.Normal	bias	-.0260	-.0044	-.0047
	var	1.616	1.157	1.292
Uniform	bias	.0557	-.1516	.0724
	var	1.422	0.771	1.479

Table 5.8: Monte Carlo results for selected estimates of λ when $\lambda_0=1.5$, $\beta_0=4.164977$, $\sigma_0=1$, and n=40 in the single-sample model (5.22) under the modified power transformation (5.24).

errors ↓		ESTIMATOR $\hat{\lambda}_{CVM}$	$\hat{\lambda}_{B\&C}$	$\hat{\lambda}_{R\&A}$
Normal ± 4	bias	-.0003	-.0980	-.0007
	var	1.709	1.342	1.642
Normal ± 2	bias	.0564	-.1678	.0075
	var	1.378	0.957	1.372
Student-t	bias	-.0397	.0420	-.0361
	var	2.768	2.353	2.726
Cont.Normal	bias	-.0740	-.0584	-.0188
	var	2.874	2.114	2.503
Uniform	bias	.0836	-.4053	.1494
	var	2.715	1.455	2.905

Table 5.9: Monte Carlo results for selected estimates of λ when $\lambda_0=.25$, $\beta_0=25.629$, $\sigma_0=.25$, and n=40 in the single-sample model (5.22) under the modified power transformation (5.24).

errors ↓		ESTIMATOR $\hat{\lambda}_{CVM}$	$\hat{\lambda}_{B\&C}$	$\hat{\lambda}_{R\&A}$
Normal ± 4	bias	.0232	-.0199	.0178
	var	2.033	1.607	1.829
Normal ± 2	bias	.0244	-.0418	.0397
	var	1.591	1.097	1.580
Student-t	bias	.0113	.0088	-.0259
	var	3.188	2.991	3.055
Cont.Normal	bias	-.0053	.0319	.0626
	var	3.161	1.936	2.850
Uniform	bias	.0090	-.1046	.0293
	var	3.273	1.656	3.041

BIBLIOGRAPHY

Andrews, D. F. (1971) "A Note on the Selection of Data Transformations", *Biometrika* 58, 249-254

Atkinson, A. C. (1973) "Testing Transformations to Normality", *Journal of the Royal Statistical Society*, *B* 35, 473-479

Atkinson, A. C. (1985) *Plots, Transformations, and Regression*, Clarendon Press, Oxford.

Bartlett, M. S. (1947) "The Use of Transformations", *Biometrics* 3, 39-52

Bassett, G. and Koenker, R. (1978) "Asymptotic Theory of Least Absolute Error Regression", *Journal of the American Statistical Association* 73, 618-622

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

Bickel, P. J. (1975) "One-step Huber Estimates in the Linear Model", *Journal of the American Statistical Association*, 70, 428-434

Bickel, P. J. and Doksum, K. A. (1981) "An Analysis of Transformations Revisited", *Journal of the American Statistical Association* 76, 296-311

Boos, D. D. (1981) "Minimum Distance Estimators for Location and Goodness of Fit", *Journal of the American Statistical Association*, 76, 663-670

Box, G. E. P. and Cox, D. R. (1964) "An Analysis of Transformations", *Journal of the Royal Statistical Society*, *B* 26, 211-252

Box, G. E. P. and Cox, D. R. (1982) "An Analysis of Transformations Revisited", *Journal of the American Statistical Association* 77, 209-210

Box, G. E. P. and Hill, W. J. (1974), "Correcting Inhomogeneity of Variance with Power Transformation Weighting", *Technometrics* 16, 385-389

Box, G. E. P. and Tidwell, P. W. (1962) "Transformation of the Independent Variables", *Technometrics* 4, 531-550

Breiman, L. and Friedman, J. H. (1985) "Estimating Optimal Transformations for Multiple Regression and Correlation", *Journal of the American Statistical Association* 80, 580-619

Butler, C. C. (1969) "A Test for Symmetry Using the Sample Distribution Function", *Annals of Mathematical Statistics*, 40, 2209-2210

Carroll, R. J.(1980) "A Robust Method for Testing Transformations to Achieve Approximate Normality", *Journal of the American Statistical Association* 74, 674-679

Carroll, R. J. and Ruppert, D. (1981) "On Prediction and the Power Transformation Family", *Biometrika* 68, 609-615

Carroll, R. J. and Ruppert, D. (1984) "Power Transformations when Fitting Theoretical Models to Data", *Journal of the American Statistical Association* 79, 321-328

Carroll, R. J. and Ruppert, D. (1985) "Transformations in Regression: A Robust Analysis", *Technometrics* 27, 1-12

Carroll, R. J. and Ruppert, D. (1987) "Diagnostics and Robust Estimation When Transforming the Regression Model and the Response", *Technometrics* 29, 287-299

Carroll, R. J. and Ruppert, D. (1988) *Transformations and Weighting in Regression*, Chapman and Hall, New York and London.

Dahlquist, G., Bjorck, A. and Anderson, N. (1974) *Numerical Methods*, Prentice-Hall, Englewood Cliffs, New Jersey.

Draper, N. R. and Cox, D. R. (1969) "On Distributions and Their Transformation to Normality", *Journal of the Royal Statistical Society*, *B* 31, 472-476

Duan, N. (1983) "Smearing Estimate: A Nonparametric Retransformation Method", *Journal of the American Statistical Association* 78, 605-610

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986) *Robust Statistics: The Approach based on Influence Functions*, Wiley, New York.

Hastie, T. and Tibshirani, R. (1986) "Generalized Additive Models", *Statistical Science* 1, 297-318

Hernandez, F. and Johnson, R. A. (1980) "The Large-Sample Behaviour 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

Hinkley, D. V. (1977) "On Quick Choice of Power Transformation", *Applied Statistics* 26, 67-68

Hinkley, D. V. and Runger, G. (1984) "Analysis of Transformed Data", *Journal of the American Statistical Association* 79, 302-309

Huber, P. J. (1967) "Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions", *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1, 221-233

Jennrich, R. I. (1969) "Asymptotic Properties of Nonlinear Least Squares Estimators", *Annals of Mathematical Statistics* 40, 633-643

Kettl, E. (1987) "Some Applications of the Transform-Both-Sides Regression Model (dissertation)", Mimeo Series #1725, Department of Statistics, University of North Carolina at Chapel Hill.

Koziol, J. A. (1980) "On a Cramér-von Mises-Type Statistic for Testing Symmetry", *Journal of the American Statistical Association*, 75, 161-167

Malinvaud, E. (1970) "The Consistency of Nonlinear Regression", *Annals of Mathematical Statistics* 41, 956-969

Nelder, J.A. and Mead, R. (1965) "A Simplex Method for Function Minimization", *The Computer Journal* 7, 308

Oberhofer, W. (1982) "The Consistency of Nonlinear Regression Minimizing the L_1-norm", *The Annals of Statistics* 10, 316-319

Orlov, A. I. (1972) "On Testing the Symmetry of Distributions", *Theory of Probability and Its Applications*, 17, 357-361

Parr, W. C., and De Wet, T. (1981) "On Minimum Cramér-von Mises-Norm Parameter Estimation", *Communications is Statistics, Part A - Theory and Methods* 10(12), 1149-1166

Rothman, E. D., and Woodroofe, M. (1972) "A Cramér-von Mises Type Statistic for Testing Symmetry", *Annals of Mathematical Statistics*, 43, 2035-2038

Rubin, H. (1956) "Uniform Convergence of Random Functions with Applications to Statistics", *The Annals of Mathematical Statistics* 27, 200-203

Ruppert, D. and Aldershof, B. (1989) "Transformations to Symmetry and Homoscedasticity", *Journal of the American Statistical Association* 84, 437-446

Ruppert, D. and Carroll, R. J. (1985) "Data Transformations in Regression Analysis with Applications to Stock-Recruitment Relationships", In *Resource Management*, M. Mangel editor, *Lecture Notes in Biomathematics* 61, Springer Verlag, New York.

Ruppert, D., Cressie, N., and Carroll, R. J. (1989) "A Transformation/ Weighting Model for Estimating Michaelis-Menten Parameters", *Biometrics* 45, 637-656

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Serfling, R. J. (1984) "Generalized L-, M-, and R-Statistics", *The Annals of Statistics* 12, 76-86

Snee, R. D. (1986) "An Alternative Approach to Fitting Models when Re-expression of the Response is Useful", *Journal of Quality Technology* 18, 211-225

Taylor, J. M. G. (1985) "Power Transformations to Symmetry", *Biometrika* 72, 145-152

Taylor, J. M. G. (1986) "The Retransformed Mean After a Fitted Power Transformation", *Journal of the American Statistical Association* 81, 114-118

Van Zwet, W. R. (1964) "Convex Transformations of Random Variables", *Mathematisch Centrum*, Amsterdam.

Wu, C. F. (1981) "Asymptotic Theory of Nonlinear Least Squares Estimation", *The Annals of Statistics* 9, 501-513