SOME APPLICATIONS OF THE
TRANSFORM-BOTH-SIDES REGRESSION MODEL

by

Ernestine E. Kettl

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

1987

Approved by

[Signatures of Advisor and Readers]
ERNESTINE E. KETTL. Some Applications of the Transform-Both-Sides Regression Model. (Under the direction of RAYMOND J. CARROLL.)

Transform-both-sides regression is a useful method when an assumed relationship exists between the response and the predictors but the way in which error enters the model is not well understood. Transform-both-sides applies the same transformation to both the response and the regression function thereby preserving the relationship between the response and the predictors yet allows variability to enter the model in different ways. In this dissertation we investigate three applications of the transform-both-sides regression model.

First the transform-both-sides model is extended to include power-of-the-mean weighting. Power transformation and weighting are used together to remove heteroscedasticity and induce symmetry in the error distribution. The regression parameters are estimated using a generalized least squares method where the transformation parameters are fixed. An approximate confidence region is obtained by allowing the transformation parameters to range over a grid of values and then the regression parameters are estimated for each fixed pair of the transformation parameters. We compute the asymptotic distribution of the estimated parameters and show that as \( N \to \infty \) and then \( \sigma \to 0 \) there is no effect due to fixing the transformation parameters. This approach is then applied to three sets of data.

Next we consider the Errors-in-Variables problem in the transform-both-sides model. We compute the asymptotic distribution for the estimated parameters when the measurement error is ignored and form corrected estimators to remove the resulting bias.
Finally we study the nonparametric ACE procedure in the transform-both-sides setting. We show that the parameterized transform-both-sides ACE criterion yields inconsistent estimates for the transformation parameter when it is equal to zero. We find that ACE does not account for heteroscedasticity nor does it always identify the correct transformation in a collection of examples generated by a transform-both-sides model. In some numerical examples we find ACE to be useful in discriminating between competing models and able to identify known influential points.
ACKNOWLEDGEMENTS

I want to express my gratitude to my advisor, Raymond Carroll, for the amount of time and effort he put forth to direct this work. I appreciate his unlimited patience and support, and thank him for his guidance and encouragement throughout the preparation of this thesis.

I acknowledge the members of my committee, David Ruppert, Don Richards, Bob Rodriguez, and M. R. Leadbetter, for their diligence in reading the manuscript in such a timely fashion and for the careful attention given to detail. In addition I thank Bob Rodriguez for making available the programmed ACE procedure, and the SAS Institute for allowing me the use of their computing facilities.

I also thank the members of the staff who have been most helpful throughout the past five years, especially June Maxwell for somehow always managing to keep things in perspective.

A special note of appreciation to my roommate, classmate and friend, Marie Davidian. More than anyone else she knows exactly what obtaining this degree has meant. She has been there to pick up the pieces when things fell apart and to celebrate when things finally did go right. Thanks for being there when it counted.

I also want to thank my fellow graduate students for their friendship and support. Without you, David, Yin, Charu, George, ..., this would not have been possible.

Finally I want to thank my family and friends for their encouragement throughout my education. It is truly appreciated.

This work was supported by the US Air Force Office of Scientific Research under Grant AFOSR F49620-82-C-0009.
# TABLE OF CONTENTS

## CHAPTER I: INTRODUCTION

1.0 Introduction .................................................. 1

1.1 Transformations ............................................... 2

1.2 Heteroscedasticity ........................................... 8

1.3 Errors-in-Variables ......................................... 11

1.4 Nonparametric Estimation of the Transformation ............ 14

1.5 Summary of Results .......................................... 14

## CHAPTER II: TRANSFORM-BOTH-SIDES AND POWER-OF-THE-MEAN WEIGHTING

2.0 Introduction .................................................. 17

2.1 Theoretical Analysis ......................................... 18

2.2 Examples ..................................................... 30

2.2.1 Capacitance Data ......................................... 31

2.2.2 The Skeena River Sockeye Salmon Data .................. 35

2.2.3 Population A Sockeye Salmon Data ....................... 39

2.2.4 Discussion ................................................ 41

2.3 Appendix ...................................................... 42

2.3.1 Heuristic Sketch of the Carroll-Ruppert Efficiency Result ................. 42

2.3.2 A Heuristic Derivation of (2.3) ........................ 43

2.3.3 Approximate Confidence Regions for ($\lambda$, $\theta$) ........ 45

2.4 Tables and Figures .......................................... 48
CHAPTER III: ERRORS-IN-VARIABLES AND THE TRANSFORM-BOTH-SIDES MODEL
3.0 Introduction ................................................. 66
3.1 Asymptotic Theory ............................................. 66
3.2 Detailed Calculations ......................................... 69
3.3 Appendix ...................................................... 76

CHAPTER IV: THE ACE ALGORITHM AND TRANSFORM-BOTH-SIDES REGRESSION
4.0 Introduction .................................................... 78
4.1 The ACE Criterion in the Transform-Both-Sides Setting
4.1.1 Theoretical Analysis ....................................... 81
4.1.2 Examples ..................................................... 92
4.2 Using ACE to Determine Model Suitability .................. 97
4.2.1 The Atlantic Menhaden Fishery Data ..................... 98
4.2.2 Chemical Reaction Data .................................... 100
4.3 ACE as a Diagnostic Tool .................................... 102
4.4 Discussion ..................................................... 106
4.5 Appendix ....................................................... 107
4.6 Tables and Figures ............................................ 107

BIBLIOGRAPHY ..................................................... 132
CHAPTER I
INTRODUCTION

1.0 Introduction

In regression one seeks to find a relationship between the response, \( y_1 \), and the independent variables, \( x_i = (x_{i1}, \ldots, x_{ip}) \). Frequently there is a previously known and well understood functional relationship between the response variable \( y_1 \) and the explanatory variables \( x_i \). This relationship might be derived from some physical process or biological system that is known in the absence of error to generate data according to \( y_1 = f(x_i, \beta) \), where \( f \) is some known function and \( \beta \) is a vector of unknown parameters. Error is introduced into the system through measurement and modeling error as well as through random variability.

While it is not always clear how this error enters the model, it is common to assume that the error is additive, i.e.,

\[
y_1 = f(x_i, \beta) + \sigma \epsilon_i
\]

where the \( \epsilon_i \) are independent and identically distributed standard normal random variables. When the data indicate that this model is inadequate one might then assume that the errors are multiplicative and lognormal so that

\[
\log(y_1) = \log[f(x_i, \beta)] + \sigma \epsilon_i.
\]
Models (1.1) and (1.2) are based on the same theoretical transfor-
mation model, where we write

\[ h(y_i) = h[f(x_i, \beta)] + \sigma \varepsilon_i \]

for some monotonic transformation \( h(.) \). This transform-both-sides
approach preserves the assumed relationship between \( y_i \) and \( x_i \), yet
allows variability to enter the model in different ways.

Although transforming both sides of the regression equation is
not a new technique, Carroll and Ruppert (1984) are the first to
advocate its routine use for reasons other than linearizing the model,
i.e., symmetrizing the error distribution and stabilizing the
variance. We consider some applications of this approach in the
following chapters. In the first two sections of this chapter we
describe in detail the transform-both-sides model and review the
transformation and heteroscedasticity literature. In Section 1.3 the
relevant Errors-in-Variables literature for the measurement error
problem is discussed. In Section 1.4 we briefly discuss the ACE
algorithm, a nonparametric procedure for estimating transformations.
Section 1.5 contains a brief summary of the results obtained.

1.1 Transformations

Transformations have long been used to stabilize variance and to
transform responses from a known nonnormal distribution to approximate
normality, thereby allowing the use of normal theory methods. Trans-
formations have also been used to linearize regression models when the
responses are related to the independent variables by a known
nonlinear function of the unknown parameters. Standard least squares methods can then be used to analyze the model if the resulting errors are independent, homogeneous and normally distributed. For review of transformations see Draper and Smith (1981, pp. 220–241) and Cook and Weisberg (1982, pp. 58–86).

In certain problems, standard regression procedures sometimes yield residuals that are skewed, heteroscedastic or otherwise nonnormal. Either some violation of the assumptions of independence, homogeneity, or normality has occurred, or there may exist some other deficiency in the postulated model. Box and Cox (1964) propose for a positive response variable $y$, a family of transformations indexed by the parameter $\lambda$.

\[
y^{(\lambda)} = \begin{cases} 
\frac{y^{\lambda-1}}{\lambda} & \text{if } \lambda \neq 0 \\
\log(y) & \text{if } \lambda = 0
\end{cases}
\]

They assume that there exists some transformation for which $y^{(\lambda)}$ will be normally distributed with constant variance and additive errors. Since (1.4) assumes that $y$ is positive, only transformations to approximate normality can be achieved, except for $\lambda = 0$. Furthermore, a single transformation may not exist that will simultaneously achieve all the stated objectives and the scale chosen may be a compromise selection.

The Box-Cox transformation family has received much attention in the literature. Maximum likelihood estimation and normal theory methods have been investigated by Draper and Cox (1969). Andrews
(1971), Atkinson (1973), Hinkley (1975), Hernandez and Johnson (1980),
Robust estimation and diagnostic methods for this model are reviewed
by Carroll and Ruppert (1985).

Much discussion has been generated over the question of how one
should proceed once an appropriate transformation has been selected.
In practice one usually estimates $\lambda$ and then proceeds as if the
estimated $\hat{\lambda}$ is the correct scale. In the "transform the response
only" model where $y_1^{(\lambda)} = f(x_1, \beta) + \sigma \varepsilon_1$, Bickel and Doksum (1981)
estimate ($\beta$, $\sigma$, $\lambda$) jointly and show that the estimate of $\beta$
when $\lambda$ is unknown can be much more variable than when $\lambda$ is known. A meaningful
interpretation of results in the joint estimation scheme is difficult
since it is made in reference to an unknown scale. Although the error
of the estimated parameters may be large, Carroll and Ruppert (1981)
show that prediction error in the original scale is not much larger
than when $\lambda$ is known. In many important cases, because the
differences between knowing and estimating $\lambda$ are small and since
linear model parameters are meaningful only in reference to a
particular scale, Box and Cox (1982) and Hinkley and Runger (1984)
recommend conditioning on $\hat{\lambda}$ as if it were known and reporting
subsequent results on the estimated scale. If one is willing to
accept the estimated scale as the true scale, such an approach leads
to results that are more easily interpreted since they are based on a
known scale.

Often in practice, after estimating $\lambda$, one then chooses the
nearest value of $\lambda$ that makes sense within the framework of the
problem. The selected transformation may have more scientific
relevance and lead to simpler meaningful interpretations of the results than analyses done on the \( \hat{\lambda} \)-scale. Carroll (1982b) addresses this problem and shows that restricted estimation of \( \lambda \) can lead to inferences different from the maximum likelihood estimation of \( \lambda \), although in many practical problems the restricted estimate works well.

In the transform-both-sides problem, the same transformation is applied to the response as well as the regression function. Carroll and Ruppert (1984) investigate maximum likelihood estimation in the "transform-both-sides" problem using the Box-Cox family of transformations, i.e.,

\[
y_1^{(\lambda)} = f^{(\lambda)}(x_1, \beta) + \sigma \varepsilon_1.
\]

This is model (1.3) where \( h(.) \) is the Box-Cox transformation.

Besides preserving the assumed relationship between \( x_1 \) and \( y_1 \), transform-both-sides is appealing in that unlike the Box-Cox "transform the response only" model, \( f(x_1, \beta) \) still has meaning after transformation. According to model (1.5), \( y_1^{(\lambda)} \) is normally distributed about \( f^{(\lambda)}(x_1, \beta) \), therefore \( f^{(\lambda)}(x_1, \beta) \) gives the conditional mean and median of \( y_1^{(\lambda)}|x_1 \). Unless \( \lambda = 1 \), \( f(x_1, \beta) \) is not the conditional mean for \( y_1|x_1 \). Regardless of the value of \( \lambda \), \( f(x_1, \beta) \) is the conditional median for the untransformed response \( y_1|x_1 \). See Ruppert and Carroll (1985) for a discussion of the conditional distribution of \( y_1|x_1 \) in the transform-both-sides setting.

Transformations have also been applied to both sides of the equation where the response and each of the explanatory variables are transformed by not necessarily the same transformation. See for
example Box and Tidwell (1962) and Boylan, et al., (1982). This differs from the transform-both-sides of Carroll and Ruppert (1984), where the same transformation is applied to both the response and the known regression function.

There are settings in which transformations can be used to remove heteroscedasticity and to induce symmetry. Using Bartlett’s (1947) method for stabilizing variance, by Taylor approximation, we get

\[
\text{Var}(y_i^{(\lambda)}) \approx \left[ \frac{dy_i^{(\lambda)}}{dy_i} \right] \text{Var}(y_i) \\
= [f(x_i, \beta)]^{2(\lambda-1)} \text{Var}(y_i).
\]

In order for \( \text{Var}(y_i^{(\lambda)}) \) to be approximately constant, the \( \text{Var}(y_i) \) is proportional to \([f(x_i, \beta)]^{2(1-\lambda)} \). Therefore if the variance of \( y_i \) is proportional to its expected value then a value for \( \lambda \) can be found so that the transformed response will have approximately homogeneous errors.

Since the Jacobian of the transformation, \( y^{\lambda-1} \), forces the transformed density to alter its shape, skewness can also be reduced by transformation. Suppose that the density of \( y \) is positively skewed. To induce symmetry, the large values of \( y \) that come from the extended right tail must be brought together and the smaller values in the shorter left tail must be spread out. Since \( y^{\lambda-1} \) is increasing in \( \lambda \), and is concave for \( \lambda \leq 1 \), the desired adjustment in shape will occur. Smaller values of \( \lambda \) will increase the amount of compression in the right tail. For \( \lambda \geq 1 \), \( y^{\lambda-1} \) is convex. Similarly, left skewed distributions can be transformed to approximate symmetry.
As demonstrated, power transformations can reduce both skewness and heteroscedasticity in a model. However, the value of $\lambda$ that transforms to approximate symmetry may not be the same value that stabilizes the variance. Therefore if a model shows substantial skewness and heteroscedasticity it may be an unrealistic expectation to require that a single transformation simultaneously achieve both symmetry and homoscedastic errors.

It is noted by Zarembka (1974) that the estimate of the transformation parameter of the response is biased in the direction of stabilizing the variance. In some econometric problems where the interest in $\lambda$ is to determine the optimal functional form, it is important for the researcher to know what contributed to the estimate $\hat{\lambda}$, i.e., heteroscedastic errors, nonnormality, or structural nonlinearity. Some work has been done in an attempt to model out the heteroscedasticity. See Caudry and Dagenais (1979) and Blaylock and Smallwood (1985) for examples in Box-Tidwell type transformation models. Ruppert and Carroll (1985) extend their transform-both-sides model by including a separate parameter to model the heteroscedasticity. Their expanded model,

$$y_{i}^{(\lambda)} = f(\lambda)(x_{i}, \beta) + \sigma x_{i}^{\theta} \epsilon_{i},$$

is more flexible allowing $\theta$ to control the heteroscedasticity and $\lambda$ to remove the skewness. Egy and Lahiri (1979) construct a similar model where they also model the heteroscedasticity by weighting with a power of the predictor variable, $x_{1}$. Their model, in our notation,

$$y_{i}^{(\lambda)} = a_{0} + a_{1} f^{(\lambda)}(x_{i}, \beta) + \sigma x_{i}^{\theta} \epsilon_{i},$$
differs slightly from the Carroll-Ruppert transform-both-sides model with the inclusion of the intercept \(a_0\) and the slope \(a_1\). Their intent is to create a model such that the transformation parameter \(\lambda\) represents the nonlinearity of the structure and not the heteroscedasticity or nonnormality.

1.2 Heteroscedasticity

Heteroscedastic regression models, where

\[
y_i = f(x_i, \beta) + \sigma_1 \epsilon_i \quad i = 1, 2, ..., n,
\]

\(f\) is a known function, \(\beta\) is a vector of unknown parameters, and \(\sigma_1\) represents the heteroscedasticity, have been used successfully to account for nonconstant variance. The variance \(\sigma_1\) can be of unknown form or modeled as a parametric function of the explanatory variables. One popular form of \(\sigma_1\) is the "power-of-the-mean" model where \(\sigma_1 = \sigma f^\theta(x_i, \beta)\) for \(f(x_i, \beta) > 0\) and \(\theta\) and \(\sigma\) are unknown scalars. This model is widely used in applications where it is reasonable to expect a positive mean response. See for example, Box and Hill (1974), Pritchard, Downie and Bacon (1977), and Bates, Wolf and Watts (1985).

When theoretical considerations have already determined the regression function, transform-both-sides is a way to preserve the functional relationship and still allow for transformation. The "transform-both-sides" model of Carroll and Ruppert (1984) is related to the "power-of-the-mean" model when \(\sigma\) is small. Carroll and Ruppert (1984) show by a Taylor expansion that

\[
y_i^{(\lambda)} = f^{(\lambda)}(x_i, \beta) + \sigma \epsilon_i.
\]
may be written as
\[ y_i = f(x_i, \beta) + \sigma [f(x_i, \beta)]^{1-\lambda} \varepsilon_i + O(\sigma^2). \]

Thus for small \( \sigma \), where \( \theta \equiv 1 - \lambda \), transform-both-sides and power-of-the-mean models should yield approximately the same results. See Carroll and Ruppert (1984), Snee (1985), and Bates, Wolf and Watts (1985) for examples where transform-both-sides is used to analyze data.

Ruppert and Carroll (1985) expand their transform-both-sides model by modeling heteroscedasticity separately. In addition to transforming both the response and the regression function by the Box-Cox transformation, they model the variance as a power of the explanatory variable,
\[ y_i^{(\lambda)} = f^{(\lambda)}(x_i, \beta) + \sigma x_i^\theta \varepsilon_i, \]
thus allowing for greater flexibility in the model.

While maximum likelihood is the standard procedure used to estimate parameters in transformation models, the usual method to estimate parameters in heteroscedastic regression is some form of generalized least squares (GLS). Consider the power-of-the-mean model,
\[ y_i = f(x_i, \beta) + \sigma f^\theta(x_i, \beta) \varepsilon_i, \]
where \( f \) is a known function, \( \beta \) is a vector of unknown parameters, \( \theta \) is an unknown scalar and \( \varepsilon_i \) are independent standard normal errors. In a generalized least squares procedure we start with some "good" preliminary estimate for \( \beta \), say \( \hat{\beta}_p \). For example, as a preliminary estimate for \( \beta \), we might use the unweighted least squares estimate.
If $\theta$ is known, we form the estimated weights, $f^{2\theta}(x_i, \hat{\beta}_p)$, and then apply weighted least squares to estimate $\beta$. If $\theta$ is unknown, it too must be estimated. Box and Hill (1974), Pritchard, Downie and Bacon (1977), Jobson and Fuller (1980), and Carroll and Ruppert (1982a) study this model. Jobson and Fuller (1980) and Carroll and Ruppert (1982a) prove that if $(\hat{\beta}, \hat{\theta})$ are consistent estimates of $(\beta, \theta)$, then all generalized least squares estimates of $\beta$ will have the same asymptotic distribution as weighted least squares with known weights. Carroll and Ruppert (1982a) propose a pseudo-likelihood approach for obtaining a consistent estimate of $\theta$.

Although asymptotically optimal when the model and the distributional assumptions are correct, maximum likelihood estimators are sensitive to outliers and are not robust against departures from the assumptions. McCullagh (1983) and Carroll and Ruppert (1982a) caution against the routine use of normal theory maximum likelihood estimation, since the maximum likelihood estimates can be much less efficient than generalized least squares estimates for nonnormal distributions. Carroll and Ruppert (1982b) show that misspecifications in the form of the variance function lead to a bias in $\hat{\beta}_{MLE}$ while hardly affecting the asymptotic distribution of $\hat{\beta}_{GLS}$.

As a numerical indicator of heteroscedasticity Carroll and Ruppert (1987) compute the Spearman's rank correlation coefficient $(\rho)$ on the absolute studentized residuals and the predictor or the predicted values. This correlation is computed by ranking each variable and computing the usual correlation based on the ranks. Since Spearman's correlation is not changed by monotone transformations of either variable, we can use squared residuals or their
logarithms as well as the logarithms of the predictors or predicted values. The Spearman correlation significance levels can be used as a rough indicator of heteroscedasticity. However one should not rely too strongly on these significance levels since no formal theory has yet been developed.

1.3 Errors-in-Variables

In the usual regression situation we assume the responses $y_i$ are random variables but that the $x_i$'s are fixed and known. However in many practical situations the $x_i$'s are subject to measurement error; for example, physical quantities such as temperature, pressure, and time cannot be measured precisely. Typically one ignores this variability and treats the observed $x_i$'s as known. Such an analysis will lead to estimates that are biased, and the predicted values may differ from those based on the correct model.

There is a vast literature on the subject of measurement error. The problem was considered even before the turn of the century by Adcock (1878) and Kummell (1879). Much of the subsequent work has concentrated on the linear model. For discussion and review see Madansky (1959) and Kendall and Stuart (1979, Ch. 29). The asymptotic normality for the maximum likelihood estimates of the regression parameters is established in the univariate case by Fuller (1980) and in the multivariate case by Gleser (1981). Recent interest has been in the area of nonlinear models, see Wolter and Fuller (1982a and 1982b), Stefanski (1983) and Carroll, et al. (1984). Amemiya and Fuller (1985) consider estimation in an implicit nonlinear functional relationship. Stefanski (1985) presents a general formulation for the
Errors-In-Variables (EIV) problem encompassing both linear and nonlinear models and functional and structural relationships.

We want to consider the nonlinear transform-both-sides model where the observations \((y_1, X_1)\) are such that the predictors \(x_1\) are measured with error. Suppose we have the following model when \(x_1\) contains no error.

\[
y_1^{(\lambda)} = f^{(\lambda)}(x_1, \beta) + \sigma_1 \varepsilon_1
\]

where \(f\) is a known function, \(\beta\) is an unknown vector of parameters, and \(z^{(\lambda)}\) is the Box-Cox transformation. Further assume that the \(\varepsilon_1\) are independent, symmetric with mean 0 and variance 1. Assume also that we can model the measurement error as \(X_1 = x_1 + \sigma_2 v_1\), where the \(v_1\)'s are independent, symmetric, with mean 0 and variance 1, and \(\varepsilon_1\) is independent of \(v_1\), for all \(i\). Model (1.8) postulated for \((y_1, X_1)\) does not necessarily hold for the observed data \((y_1, X_1)\).

To demonstrate this, let us consider simple linear regression. Assume \(y_1\) and \(X_1\) are independent given \(x_1\), and that \(x_1 \sim N(\xi, \sigma_3^2)\). Assume also that \(y_1 | x_1 \sim N(\alpha + \beta x_1, \sigma_1^2)\) where \(y_1 = \alpha + \beta x_1 + \sigma_1 \varepsilon_1\), and \(X_1 | x_1 \sim N(x_1, \sigma_2^2)\) where \(X_1 = x_1 + \sigma_2 v_1\). Working with the appropriate joint and marginal densities we find that \(X_1 \sim N(\xi, \sigma_2^2 + \sigma_3^2)\) and \(y_1 | X_1 \sim N(\alpha + \beta x_1, \sigma^2)\), where \(\alpha = \alpha + \beta \xi \sigma_2^2 / (\sigma_2^2 + \sigma_3^2)\), \(\beta = \beta \sigma_3^2 / (\sigma_2^2 + \sigma_3^2)\), and \(\sigma^2 = \sigma_1^2 + \beta^2 \sigma_2^2 \sigma_3^2 / (\sigma_2^2 + \sigma_3^2)\). Thus in this simple linear regression we retain the linear structure, but the slope is less than the original slope \(\beta\), the intercept \(\alpha\) differs from \(\alpha\) by an additive term, and the variance \(\sigma^2\) is increased by a term depending on the original slope and the error variances of \(x_1\) and \(X_1 | x_1\).

Conditioning on the observed \(X_1\) changes the model considerably.
even in this simple case. It is not at all clear what will happen in a more complicated nonlinear model. If we look at the transform-both-sides model, where \( y_1^{(\lambda)} = f^{(\lambda)}(x_1, \beta) + \sigma_1 e_1 \) corresponds to \( y_1^{(\lambda)} | x_1 \sim N(f^{(\lambda)}(x_1, \beta), \sigma_1^2) \), and if we assume that \( X_1 | x_1 \sim N(x_1, \sigma_2^2) \) and \( x_1 \) has density \( g \), then the density of \( y_1^{(\lambda)} | X_1 \), say \( h \), is

\[
1 \over \sqrt{2\pi\sigma_1} \int \exp \left\{ -1 \left[ \frac{(X - x)^2}{\sigma_2} + \frac{t - f^{(\lambda)}(x, \beta)^2}{\sigma_1^2} \right] \right\} g(x) \, dx
\]

Depending on \( g \), we may or may not be able to get a closed form for \( h \), and the resulting conditional model may bear little resemblance to the original model. If \( h(t) \) has a closed form then standard likelihood procedures can be used to estimate the parameters. We hope that

\[
\lim_{\sigma_2 \to 0} \left[ \text{E}(y^{(\lambda)} | x) - f^{(\lambda)}(x, \beta) \right] = 0
\]

In most cases (1.9) will be difficult if not impossible to work with and since (1.10) is the desired result, we can use (1.8) to obtain estimates for \( \beta \), by pretending there is no error in the observed \( x_1 \)'s. If we then substitute \( X_1 \) for \( x_1 \) into the estimator, the resulting estimates are biased. Employing methods similar to those of Wolter and Fuller (1982) and Amemiya and Fuller (1985) we derive the asymptotic distribution of the estimators. Stefanski (1983) uses a similar asymptotic theory in a logistic regression model. See also Stefanski and Carroll (1985).
1.4 Nonparametric estimation of the transformation

Some research has been done in estimating transformations through the use of nonparametric regression, notably that of Breiman and Friedman (1985a, b), Young (1981, 1985), and Young, de Leeuw, and Takane (1976). The work of Young (1981, 1985) and Young, et al., (1976) centers around an algorithm called MORALS ("Multiple Optimal Regression by Alternating Least Squares"). Originally developed for use with qualitative data in a psychometric setting, the MORALS procedure deals with discrete data, whereas the more recent ACE ("Alternating Conditional Expectation") algorithm of Breiman and Friedman (1985) is more flexible. The ACE procedure employs nonparametric smoothing and yields numerical transformations such that the subsequent plots are easier to interpret than the MORALS output. While the general transformations of ACE and MORALS are not necessarily monotone, they can be determined quite readily with ACE although some grouping of data into discrete intervals must occur before MORALS can be applied. See Rodriguez (1985) for a comparison of the ACE and MORALS procedures. Breiman and Friedman (1985a) develop a large body of theoretical results for the ACE procedure. Nonparametric monotone transformations have previously been studied by Kruskal (1964, 1965).

1.5 Summary of results

The discussion of Sections 1.1 and 1.2 reviews the role of the transform-both-sides approach in nonlinear regression. We investigate some applications of the transform-both-sides model where specifically
we consider heteroscedasticity, measurement error and a nonparametric regression procedure for estimating the transformation.

In Chapter 2 we use transform-both-sides but model the variance separately as a power-of-the-mean. We consider the model

\[
y_1^{(\lambda)} = f^{(\lambda)}(x_1, \beta) + \sigma f^{(\lambda)}(x_1, \beta) \varepsilon_{1}.
\]

where the \( \varepsilon_{1} \) are symmetric, independent and identically distributed with mean 0 and variance 1. As in Carroll and Ruppert (1984) we compute the asymptotic distribution for the estimators when \( (\lambda, \theta) \) are fixed and then again when \( (\lambda, \theta) \) are estimated. We find that there is no effect on the asymptotic distribution of the estimated regression parameters.

In Chapter 3 we study the effect of measurement error in the transform-both-sides setting. Here the predictors \( x_1 \) are measured with error where the observed values \( (y_1, X_1) \) are such that \( X_1 = x_1 + \sigma_2 v_1 \) for some random error \( v_1 \). The resulting estimates for the parameters are biased. We calculate the asymptotic distribution of the estimators and study this bias for the transform-both-sides Errors-in-Variables model. We compute also bias-corrected estimators and determine the effect of measurement error on the corrected estimates.

In Chapter 4 we explore how the ACE algorithm performs in the transform-both-sides framework. Since we are interested in the modified power transformation family applied to continuous data we do not consider the MORALS procedure. We show that for \( \lambda = 0 \), the transform-both-sides parameterized version of the ACE criterion does not yield consistent estimates for \( \lambda \). Data generated according to
model (1.5) with \( \lambda = 0 \) are used to study the performance of ACE in the transform-both-sides setting. We then use ACE as a tool for selecting the better of two or more competing models. Finally we construct two case deletion diagnostics and study the effectiveness of ACE as a diagnostic tool for identifying influential points.
Chapter II

TRANSFORM-BOTH-SIDES AND POWER-OF-THE-MEAN WEIGHTING

2.0 Introduction

Suppose that in the absence of random error the relationship between x and y is such that \( y_i = f(x_i, \beta) \). Consider the model

\[
y_i^{(\lambda)} = f^{(\lambda)}(x_i, \beta) + \sigma f^{(\theta)}(x_i, \beta) \epsilon_i,
\]

where the response and the regression function are both transformed by a Box-Cox power transformation and the variance is modeled as a power-of-the-mean. \( \varnothing = \varnothing(\beta, \sigma, \lambda, \theta) \) is a vector of unknown parameters and \( f(x_i, \beta) \) is a known function. Assume the \( \epsilon_i \) are independent and symmetric with mean 0 and variance 1. The way in which error enters the model is controlled by the choice of \((\lambda, \theta)\).

In the study of this transformation model we employ the now standard small-\( \sigma \) asymptotic distribution theory of Bickel and Doksum (1981). Calculations made in computing the asymptotic distribution when \( \sigma \) is fixed are complex and yield little insight into the problem, whereas the small-\( \sigma \) asymptotic theory, where \( N \to \infty \) and the \( \sigma \to 0 \), lead to major simplifications allowing a tractable theory to be developed in the transformation setting. In data sets where the data agree closely with the model, i.e., there is little variation about the mean, the small-\( \sigma \) assumption is reasonable. See for example Bickel and Doksum (1981) and Carroll and Ruppert (1987). Carroll and
Ruppert (1987) suggest using the median coefficient of variation, median $(s_i/|\hat{\mu}_i|)$, to determine the adequacy of modeling the variance as a power-of-the-mean. Based on their experience they find the median coefficient of variation is often less than .35 when power-of-the-mean variance is used. Since the small-\(\sigma\) assumption in model (2.1) yields a power-of-the-mean equivalent model we use the median coefficient of variation to indicate the appropriateness of the small-\(\sigma\) assumption in a given situation. For model (2.1), the median coefficient of variation for a given $(\lambda, \theta)$ combination is defined as median $[\hat{\sigma}_f^\theta(x_i, \hat{\beta})/\hat{f}^\lambda(x_i, \hat{\beta})]$.

In Section 2.1 we ascertain the effect of fixing $(\lambda, \theta)$ on the asymptotic distribution of $\hat{\beta}$. We employ the small-\(\sigma\) asymptotic theory where $N \to \infty$ and then $\sigma \to 0$. In Section 2.2 the methods discussed in Section 2.1 are applied to three examples. Section 2.3 is the appendix for Chapter 2 and Section 2.4 contains all the tables and plots for the examples.

2.1 Theoretical Analysis

As discussed in Chapter 1, the parameter $\beta$ has physical meaning whether $\lambda$ is known or unknown, that is, $f(x, \beta)$ is the median of $y$ regardless of the value of $\lambda$. Thus conditioning on the value of $\lambda$ should not lead to the same controversy that arises in the Box-Cox model where only the response is transformed.

We can estimate all the elements of $\Theta$ simultaneously using maximum likelihood, but this is complex numerically. By treating $(\lambda, \theta)$ as fixed and known and conditioning on these values, standard nonlinear regression software can be used to estimate $(\beta, \sigma)$. We
estimate here the effect of treating \((\lambda, \theta)\) as known by computing the asymptotic distribution of \(\hat{\Theta}(\hat{\beta}, \hat{\sigma}, \hat{\lambda}, \hat{\theta})\) and \(\hat{\Theta}(\hat{\beta}, \hat{\sigma}|\lambda, \theta)\). Carroll and Ruppert (1984) have done this for the case where \(\theta = 0\). As they did in order to get manageable expressions, we employ the small-\(\sigma\) asymptotic theory to derive and compare the asymptotic distributions of \(\hat{\Theta}(\hat{\beta}, \hat{\sigma}, \hat{\lambda}, \hat{\theta})\) and \(\hat{\Theta}(\hat{\beta}, \hat{\sigma}|\lambda, \theta)\).

We make the following assumptions to be used throughout the arguments contained in this section.

(A.1) \((\varepsilon_i, x_i)\) are independent and identically distributed, \((\varepsilon_i)\) and \((x_i)\) are independent, and

\[
y_i^{(\lambda)} = f^{(\lambda)}(x_i, \beta) + \sigma f^{\theta}(x_i, \beta) \varepsilon_i.
\]

(A.2) \(x_i\) are uniformly bounded.

(A.3) There exists \(0 < M_1, M_2 < \infty\) such that for all \(x, \| \beta - \beta \| \leq M_2\) implies that \(M_1^{-1} \leq f(x_i, \beta) \leq M_1\).

(A.4) The standard convergence result for normal theory maximum likelihood holds, (Serfling, 1980, Chapter 4, and Rao, 1973, section 5f.2), for each \(\sigma\), i.e., as \(N \to \infty\),

\[
N^{1/2} \begin{bmatrix} \hat{\beta} - \beta \\ \hat{\sigma} - \sigma \\ \hat{\theta} - \theta \\ \hat{\lambda} - \lambda \end{bmatrix} \Rightarrow N(0, I_\beta^{-1}(\beta, \sigma, \theta, \lambda))
\]

and if \(\ell\) is the loglikelihood function,

\[
E(\partial \ell(\beta, \sigma, \theta, \lambda)/\partial (\beta, \sigma, \theta, \lambda)^T) = 0, \quad \text{and}
\]

\[
I_\beta = -E(\partial^2 \ell(\beta, \sigma, \theta, \lambda)/\partial (\beta, \sigma, \theta, \lambda) \partial (\beta, \sigma, \theta, \lambda)^T).
\]
(A.5) The standard maximum likelihood asymptotic distribution theory holds if \( \lambda \) is known, \( \theta \) is known, or both, i.e., the limiting covariance matrix of the remaining estimates is the inverse of the appropriate information matrix.

(A.6) If \( h(x, \beta, \lambda) = f(\lambda)(x_1, \beta) \) and if \( h_\beta(x, \beta, \lambda) = \partial h(x, \beta, \lambda) / \partial \beta \) then

\[
E \left[ \frac{h_\beta(x, \beta, \lambda)}{f(x_1, \beta)^{2\theta}} \right]
\]

exists and is positive definite.

(A.7) The first two derivatives of \( f(x_1, \beta) \) exist, are continuous, and are uniformly bounded.

Remark: Assumption (A.4) holds strictly only for \( \lambda = 0 \). This is because \( y^{(\lambda)} \) is undefined for \( y < 0 \), and if \( P(y > 0) = 1 \) then, except for \( \lambda = 0 \), exact transformation to normality is impossible. There is a long history starting with Box and Cox (1964) of ignoring this problem. One way to make the transformation exact is to replace \( y^{(\lambda)} \) by \( (\text{sign}|y|^{\lambda} - 1)/\lambda \). We have instead made the decision to do our calculations in the usual way, so that strictly speaking, they are exact only for \( \lambda = 0 \).

Now for each fixed \( \sigma \), write

\[
I_\sigma = \begin{bmatrix}
I_{\beta \beta} & I_{\beta \sigma} & I_{\beta \theta} & I_{\beta \lambda} \\
I_{\sigma \beta} & I_{\sigma \sigma} & I_{\sigma \theta} & I_{\sigma \lambda} \\
I_{\theta \beta} & I_{\theta \sigma} & I_{\theta \theta} & I_{\theta \lambda} \\
I_{\lambda \beta} & I_{\lambda \sigma} & I_{\lambda \theta} & I_{\lambda \lambda}
\end{bmatrix}
\]
Define

\[ F_\star = \begin{bmatrix}
\sigma^2_{\beta \beta} & \sigma^2_{\beta \theta} & \sigma_{\beta \lambda} \\
\sigma^2_{\beta \theta} & \sigma^2_{\theta \theta} & \sigma_{\theta \lambda} \\
\sigma_{\theta \lambda} & \sigma_{\theta \lambda} & I_{\theta \theta} & I_{\theta \lambda} \\
\sigma_{\lambda \beta} & \sigma_{\lambda \sigma} & I_{\lambda \theta} & I_{\lambda \lambda}
\end{bmatrix} \]

**Lemma 2.1:** For fixed \( \sigma \),

\[
N^{1/2} \begin{bmatrix}
(\hat{\beta} - \beta)/\sigma \\
(\hat{\sigma} - \sigma)/\sigma \\
\hat{\theta} - \theta \\
\hat{\lambda} - \lambda
\end{bmatrix} \Rightarrow N(0, F_\star^{-1})
\]

**Proof:** Matrix algebra using (A.4).

The loglikelihood is given by

\[
\ell(\theta) = -\log \sigma - \theta \log[f(x_1, \beta)] + (\lambda - 1) \log y_1
\]

\[
- \frac{1}{2} \left[ \frac{y_1^{(\lambda)} - f^{(\lambda)}(x_1, \beta)}{\sigma f^{\theta}(x_1, \beta)} \right]^2
\]

\[
= -\log \sigma - \theta \log[f(x_1, \beta)] + (\lambda - 1) \log y_1
\]

\[
- \frac{1}{2} \left[ \frac{h(y_1, \lambda) - h[f(x_1, \beta), \lambda]}{\sigma f^{\theta}(x_1, \beta)} \right]^2
\]

Write \( h(x, \beta, \lambda) = f^{(\lambda)}(x_1, \beta) \). We proceed on a case-by-case inspection of \( F_\star \) as \( \sigma \to 0 \). The strong boundedness conditions (A.2), (A.3) and (A.7) make these calculations legitimate.
Lemma 2.2: Write

\[ A = E \left[ \frac{h(x, \beta, \lambda) \ h^T(x, \beta, \lambda)}{f(x_1, \beta)^{2\theta}} \right] \]

where \( v_j = E[\log f(x_1, \beta)] \). Then

\[ \lim_{\sigma \to 0} F^\sigma = F_\infty = \begin{bmatrix} A & 0 \\ 0 & 2M_\infty \end{bmatrix} \]

where

\[ M_\infty = \begin{bmatrix} 1 & v_1 & -v_1 \\ v_1 & v_2 & -v_2 \\ -v_1 & -v_2 & v_2 \end{bmatrix} \]

Proof: This is just a case by case examination of \( F^\sigma \). We consider the terms \( \sigma^2 I_{\alpha \sigma} \), \( \sigma^2 I_{\beta \beta} \), and \( \sigma^2 I_{\beta \sigma} \) and omit the details for the others.

Now,

\[ \ell_\sigma = \frac{-1}{\sigma} + \frac{-1}{\sigma^3} \frac{[h(y, \lambda) - h(x, \beta, \lambda)]^2}{f^{2\theta}(x, \beta)} \]

and thus

\[ \ell_{\beta \sigma} = \frac{-2}{\sigma^3} \frac{[h(y, \lambda) - h(x, \beta, \lambda)] h_\beta(x, \beta, \lambda)}{f^{2\theta}(x, \beta)} \]

\[ + \frac{-2\theta}{\sigma^3} \frac{[h(y, \lambda) - h(x, \beta, \lambda)]^2 f_\beta(x, \beta)}{f^{2\theta+1}(x, \beta)} \]
Since \([h(y, \lambda) - h(x, \beta, \lambda)] = \sigma \int \theta(x, \beta) \, \epsilon\).

\[
\ell_{\beta \sigma} = \frac{-\epsilon}{\sigma^2} \frac{h_\beta(x, \beta, \lambda)}{f^\theta(x, \beta)} - \frac{2\theta}{\sigma} \frac{\epsilon^2 f_\beta(x, \beta)}{f(x, \beta)}.
\]

Therefore

\[
\sigma^2 I_{\beta \sigma} = -\sigma^2 \mathbb{E}[\ell_{\beta \sigma}]
\]

\[= 2 \sigma \theta \mathbb{E}[f_\beta(x, \beta)/f(x, \beta)].\] See (A.3).

Thus, \(\lim_{\sigma \to 0} \sigma^2 I_{\beta \sigma} = 0\).

The derivative of \(\ell_{\sigma}\) with respect to \(\sigma\) yields

\[
\ell_{\sigma \sigma} = \frac{1}{\sigma^2} - \frac{3}{\sigma^4} \frac{[h(y, \lambda) - h(x, \beta, \lambda)]^2}{f^{2\theta}(x, \beta)}
\]

\[= \frac{1}{\sigma^2} - \frac{3}{\sigma^2} \epsilon^2.\]

Thus \(\sigma^2 I_{\sigma \sigma} = 2\).

Note that

\[
\ell_{\beta} = \frac{-f_\beta(x, \beta)}{f(x, \beta)} + \frac{1}{\sigma^2} \frac{[h(y, \lambda) - h(x, \beta, \lambda)] h_\beta(x, \beta, \lambda)}{f^{2\theta}(x, \beta)}
\]

\[+ \frac{\theta}{\sigma^2} \frac{[h(y, \lambda) - h(x, \beta, \lambda)]^2 f_\beta(x, \beta, \lambda)}{f^{2\theta+1}(x, \beta)}
\]

\[= \frac{\theta}{\sigma^2} \frac{f_\beta(x, \beta)}{f(x, \beta)} \left[\frac{[h(y, \lambda) - h(x, \beta, \lambda)]^2}{f^{2\theta}(x, \beta)} - 1\right]
\]

\[+ \frac{1}{\sigma^2} \frac{[h(y, \lambda) - h(x, \beta, \lambda)] h_\beta(x, \beta, \lambda)}{f^{2\theta}(x, \beta)}.\]
Now
\[ \ell_{\beta} = \frac{\theta f_{\beta}(x, \beta)}{\sigma^2 f(x, \beta)} \left[ \frac{[h(y, \lambda) - h(x, \beta, \lambda)]^2}{f^2(x, \beta)} - 1 \right] \]
\[- \frac{\theta f_{\beta}(x, \beta) f_{\beta}^T(x, \beta)}{\sigma^2 f^2(x, \beta)} \left[ \frac{[h(y, \lambda) - h(x, \beta, \lambda)]^2}{f^2(x, \beta)} - 1 \right] \]
\[- \frac{2\theta f_{\beta}(x, \beta)}{\sigma^2 f(x, \beta)} \left[ \frac{[h(y, \lambda) - h(x, \beta, \lambda)] h_{\beta}(x, \beta, \lambda)}{f^2(x, \beta)} \right] \]
\[- \frac{2\theta}{\sigma^2 f^2(x, \beta)} \left[ \frac{[h(y, \lambda) - h(x, \beta, \lambda)]^2 f_{\beta}(x, \beta)}{f^{2\theta + 1}(x, \beta)} \right] \]
\[- \frac{h_{\beta}(x, \beta, \lambda) h_{\beta}^T(x, \beta, \lambda)}{\sigma^2 f^{2\theta}(x, \beta)} + \frac{[h(y, \lambda) - h(x, \beta, \lambda)] h_{\beta}(x, \beta, \lambda) f_{\beta}^T(x, \beta)}{\sigma^2 f^{2\theta}(x, \beta)} \]
\[ + \frac{2\theta}{\sigma^2 f^{2\theta}(x, \beta)} \left[ \frac{[h(y, \lambda) - h(x, \beta, \lambda)] h_{\beta}(x, \beta, \lambda) f_{\beta}^T(x, \beta)}{f^{2\theta + 1}(x, \beta)} \right]. \]

Rewrite this as
\[ \ell_{\beta} = \frac{\theta f_{\beta}(x, \beta)}{f(x, \beta)} \left[ \epsilon^2 - 1 \right] \]
\[- \frac{\theta f_{\beta}(x, \beta) f_{\beta}^T(x, \beta)}{f^2(x, \beta)} \left[ \epsilon^2 - 1 \right] \]
\[- \frac{2\theta f_{\beta}(x, \beta)}{\sigma f^{\theta + 1}(x, \beta)} \left[ \epsilon h_{\beta}(x, \beta, \lambda) - 2\sigma \epsilon^2 f_{\beta}(x, \beta) \right] \]
\[- \frac{h_{\beta}(x, \beta, \lambda) h_{\beta}^T(x, \beta, \lambda)}{\sigma^2 f^{2\theta}(x, \beta)} + \frac{1}{\sigma} \epsilon h_{\beta}(x, \beta, \lambda) \]
\[ + \frac{2\theta}{\sigma} \frac{\epsilon h_{\beta}(x, \beta, \lambda) f_{\beta}^T(x, \beta)}{f^{\theta + 1}(x, \beta)}. \]
Therefore

\[
\lim_{\sigma \to 0} \sigma^2 I_{\beta} = \mathbb{E} \left[ \frac{h_\beta(x, \beta, \lambda) h_\beta^T(x, \beta, \lambda)}{f^{2\theta}(x_1, \beta)} \right] = A.
\]

The other terms are similar but are no more difficult.

The next lemma shows that for small \( \sigma \), if either \( \lambda \) or \( \theta \) is known, the maximum likelihood estimator for \( \beta \) has approximately the same distribution as if both \( \lambda \) and \( \theta \) are known.

**Definition:** A statistic \( T_N \) will be said to satisfy \( T_N \Rightarrow N(\Omega, \Sigma) \) as \( N \to \infty \) and then \( \sigma \to 0 \) if for all sets \( A \),

\[
\lim_{\sigma \to 0} \lim_{N \to \infty} P(T_N \in A) = P(N(\Omega, \Sigma) \in A).
\]

**Lemma 2.3:** Let \( \hat{\beta}(\lambda, \theta) \) be the maximum likelihood estimator for \( \beta \) at \((\lambda, \theta)\). Then as \( N \to \infty \) and then \( \sigma \to 0 \),

\[
N^{1/2} (\hat{\beta}(\lambda, \theta) - \beta) / \sigma \Rightarrow N(\Omega_A^{-1})
\]

\[
N^{1/2} (\hat{\beta}(\lambda, \theta) - \beta) / \sigma \Rightarrow N(\Omega_A^{-1})
\]

\[
N^{1/2} (\hat{\beta}(\lambda, \theta) - \beta) / \sigma \Rightarrow N(\Omega_A^{-1}).
\]
Proof: We verify only the last case, the others being similar. Let $F_0^*(\beta, \sigma, \lambda)$ be that part of $F^*$ with row and column corresponding to known $\theta$ left out. As in Lemma 2.1,

$$N^{1/2} \begin{bmatrix} \hat{\beta} - \beta \\ \hat{\sigma} - \sigma \\ \hat{\lambda} - \lambda \end{bmatrix} \Rightarrow N(0, F_0^*(\beta, \sigma, \lambda)^{-1}).$$

Write $b = (a \ 0 \ 0)^T$. Then

$$P[a^T N^{1/2} (\hat{\beta} - \beta)/\sigma \leq v] = P[b^T N^{1/2} \begin{bmatrix} \hat{\beta} - \beta \\ \hat{\sigma} - \sigma \\ \hat{\lambda} - \lambda \end{bmatrix} \leq v].$$

$$\xrightarrow{N \to \infty} P[b^T N(0, F_0^*(\beta, \sigma, \lambda)^{-1}) \leq v].$$

But as $\sigma \to 0$, $F_0^*(\beta, \sigma, \lambda) \to \begin{bmatrix} A & 0 & 0 \\ 0 & 2 & -v_1 \\ 0 & -v_1 & v_2 \end{bmatrix} = F_0^*(\beta, \sigma, \lambda)$. Since $F_0^*$ is of full rank and hence invertible

$$F_0^*(\beta, \sigma, \lambda)^{-1} \to F_0^*(\beta, \sigma, \lambda)^{-1}$$

$$b^T F_0^*(\beta, \sigma, \lambda)^{-1} b \to b^T F_0^*(\beta, \sigma, \lambda)^{-1} b$$

$$= a^T A^{-1} a.$$

Therefore

$$\lim_{\sigma \to 0} P[b^T N(0, F_0^*(\beta, \sigma, \lambda)^{-1}) \leq v] = P[b^T N(0, A)^{-1} \leq v]$$

$$= P[a^T N(0, A^{-1}) \leq v].$$
We must now consider what happens if neither $\lambda$ nor $\theta$ is known. In this case the entire matrix $F^{-\infty}_\infty$ in Lemma 2.2 is no longer positive definite. The intuitive reason is that as $\sigma \to 0$, transformation and weighting are no longer distinguishable. The technical problem is that if $C^{-\infty}_\infty$ is a generalized inverse of $F^{-\infty}_\infty$, we need not necessarily have $(F^{-\infty}_\infty)^{-1} \to C^{-\infty}_\infty$ as $\sigma \to 0$. A simple counterexample is $C_{\sigma} = \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix}$.

While $C_{\sigma} \to \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ as $\sigma \to 0$, with a generalized inverse, $C_{\sigma}^{-1} \to \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, and $C_{\sigma}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/\sigma \end{bmatrix} \to \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ as $\sigma \to 0$.

**Lemma 2.4:** Make the additional assumption that for any $c > 0$,

$$\lim_{\sigma \to 0} \lim_{N \to \infty} P\{\|F^{-\infty}_\infty - F^{-\infty}_\infty\| \geq c\} = 0$$

where

$$T_N = N^{1/2} \begin{bmatrix} \hat{\beta} - \beta \\ \hat{\sigma} - \sigma \\ \hat{\theta} - \theta \\ \hat{\lambda} - \lambda \end{bmatrix}, \quad \hat{\beta} = \hat{\beta}(\hat{\lambda}, \hat{\theta})$$

and $\| \cdot \|$ is the maximum element. Then as $N \to \infty$ and then $\sigma \to 0$,

$$N^{1/2}(\hat{\beta} - \beta)/\sigma \Rightarrow N(0, A^{-1})$$

**Proof:** By Lemma 2.1, for fixed $\sigma$, $T_N \Rightarrow N(0, F^{-\infty}_\infty^{-1})$. Therefore, $F^{-\infty}_\infty T_N \Rightarrow N(0, F^{-\infty}_\infty)$. By the assumption, according to the definition, this means that for any vector $a$,

$$\lim_{\sigma \to 0} \lim_{N \to \infty} P\{a^T F^{-\infty}_\infty T_N \leq v\} = P\{a^T N(0, F^{-\infty}_\infty) \leq v\}$$. 
To see this, note that for any $c > 0$,

\[
P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v) = P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v + a^T(\mathcal{F}^\mathcal{M} - \mathcal{F}^\mathcal{M}_\infty)N \leq v) \\
\leq P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v + c \text{ and } |a^T(\mathcal{F}^\mathcal{M} - \mathcal{F}^\mathcal{M}_\infty)N| \leq c) \\
+ P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v + c \text{ and } |a^T(\mathcal{F}^\mathcal{M} - \mathcal{F}^\mathcal{M}_\infty)N| > c) \\
\leq P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v + c) + P(|a^T(\mathcal{F}^\mathcal{M} - \mathcal{F}^\mathcal{M}_\infty)N| > c).
\]

Also,

\[
P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v) \geq P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v - c \text{ and } |a^T(\mathcal{F}^\mathcal{M} - \mathcal{F}^\mathcal{M}_\infty)N| \leq c) \\
= P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v - c) - P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v - c \text{ and } |a^T(\mathcal{F}^\mathcal{M} - \mathcal{F}^\mathcal{M}_\infty)N| > c) \\
\geq P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v - c) - P(|a^T(\mathcal{F}^\mathcal{M} - \mathcal{F}^\mathcal{M}_\infty)N| > c).
\]

Now let $N \to \infty$ and then $\sigma \to 0$ and by the assumption of the lemma for any $c > 0$

\[
\lim_{\sigma \to 0} \lim_{N \to \infty} P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v) \leq \lim_{\sigma \to 0} \lim_{N \to \infty} P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v + c),
\]

and

\[
\lim_{\sigma \to 0} \lim_{N \to \infty} P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v) \geq \lim_{\sigma \to 0} \lim_{N \to \infty} P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v - c).
\]

Now let $c \to 0$ so that

\[
\lim_{\sigma \to 0} \lim_{N \to \infty} P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v) = \lim_{\sigma \to 0} \lim_{N \to \infty} P(a^{\mathcal{F}^\mathcal{M}}_{cN} \leq v).
\]
Thus
\[ \lim_{\sigma \to 0} \lim_{N \to \infty} \Pr\{a_{\omega}^{T}F_{N}^{*} \leq v\} = \Pr\{a_{N(\omega,F_{\infty}^{*})}^{T} \leq v\}. \]
and \( F_{\omega N}^{*} \to N(0,F_{\infty}^{*}) \). Therefore \( N^{1/2}(\hat{\beta} - \beta)/\sigma \to N(0,A^{-1}) \) because of the form of \( F_{\infty}^{*} \) (See Lemma 2.2).

While the small- \( \sigma \) asymptotic distribution theory allows for a manageable theory, the results are of somewhat limited value since we are interested in all values of \( \sigma \). Carroll and Ruppert (1984) show for the transform-both-sides model with fixed \( \sigma \) that the asymptotic relative efficiency of the MLE for \( \hat{\beta} \) when \( \lambda \) is estimated, \( \hat{\beta}(\lambda) \), and that for \( \hat{\beta} \) when \( \lambda \) is known, \( \hat{\beta}(\lambda) \), is at least \( 2/\pi \), that is, \( \text{ARE}[\hat{\beta}(\lambda),\hat{\beta}(\lambda)] \geq 2/\pi \). Our model differs in that we have also included power-of-the-mean variance which introduces the additional parameter \( \theta \). This however does not change the Carroll and Ruppert result and we have that the \( \text{ARE}[\hat{\beta}(\lambda,\hat{\theta}),\hat{\beta}(\lambda,\theta)] \geq 2/\pi \). See Section 2.3 for a heuristic sketch of the proof. This result together with Lemma 2.4 tells us that the estimates from a nonlinear procedure where we condition on the value of \( (\lambda,\theta) \) will be only moderately in error.

It is generally difficult to compute the maximum likelihood estimators in this problem. For this reason, we will use a generalized least squares estimate in all of our practical calculations. That is, for fixed \( (\lambda,\theta) \), we will solve the generalized least squares equations. (See Carroll and Ruppert, (1987), Chapter 2.)
\begin{equation}
0 = \sum_{i=1}^{N} \left[ \frac{y_i^{(\lambda)} - f^{(\lambda)}(x_i, \beta)}{f^{\theta}(x_i, \beta)} \right] \frac{f_\beta(x_i, \beta)f^{\lambda-1}(x_i, \beta)}{f^{\theta}(x_i, \beta)}.
\end{equation}

Perhaps the greatest justification for replacing $\hat{\beta}_{\text{MLE}}$ by $\hat{\beta}_{\text{GLS}}$ at each step of the calculations is that in the small- $\sigma$ asymptotic distribution theory ($N \to \infty$, then $\sigma \to 0$) $\hat{\beta}_{\text{MLE}}(\lambda, \theta)$ and $\hat{\beta}_{\text{GLS}}(\lambda, \theta)$ have the same limit distributions in this sense:

\begin{equation}
\lim_{\sigma \to 0} \lim_{N \to \infty} \left[ P\{N^{1/2}/\sigma[\hat{\beta}_{\text{MLE}}(\lambda, \theta) - \beta] \leq v\} - P\{N^{1/2}/\sigma[\hat{\beta}_{\text{GLS}}(\lambda, \theta) - \beta] \leq v\} \right] = 0,
\end{equation}

for all $v$. A sketch which indicates that this result is reasonable is given in Section 2.3.2 of the appendix.

2.2 **Examples**

In this section we apply the transform-both-sides power-of-the-mean model to three sets of data. This section demonstrates how the method can be applied in real situations and numerically illustrates how skewness and heteroscedasticity are addressed by the parameters $(\lambda, \theta)$.

In the first example, $N$ is large ($N = 297$) and $\sigma$ is small (the median coefficient of variation is approximately .1). In this example the assumed model can be linearized by transformation but the transform-both-sides power-of-the-mean analysis makes it clear that linearization is not necessary.

The second and third examples arise from fisheries management. In both examples $N$ is small, 26 and 27 respectively, and the median
coefficients of variation are approximately .4 and .9. These examples are interesting in that we shall see the interplay between skewness and heteroscedasticity and the subsequent effect on the parameters \((\lambda, \theta)\).

2.2.1 Capacitance Data

These data originate from the National Bureau of Standards (Thurber, Lowney and Phillips, 1985), and are used in the characterization of deep-level defect centers in semiconductor depletion regions through the use of transient capacitance techniques. There are two observed variables: \(x = \text{time and } C(x) = \text{capacitance measured at time } x\). Capacitance ratio, \(y = C_r(T,x)\), is a function of a fixed measurement temperature \((T)\) and time \((x)\), such that

\[
y = C_r(T,x) = \frac{[C_b^2(T) - C_f^2(T)][C_f^2(T) - C^2(x)]}{[C_b^2(T) - C^2(x)][C_f^2(T) - C_i^2(T)]}
\]

where

\(C_f(T) = \) the capacitance at reverse voltage, \(V_r\), at temperature \(T\)

\(C_b(T) = \) the capacitance at charging voltage, \(V_c\), at temperature \(T\)

\(C_i(T) = \) the capacitance when reverse voltage, \(V_r\), is restored at temperature \(T\) and at time \(x = 0\)

\(C(x) = \) the capacitance at time \(x\).

See Figure 2.1 for a schematic.

For these data

\[
y = 5.805346 \frac{[667628.747 - C^2(x)]}{785925.600447 - C^2(x)}
\]
where $C_b(T) = 886.5245$, $C_f(T) = 817.0655$, and $C_i(T) = 801.8797$. There are 11,872 original data points from which a total of 297 points are chosen by selecting every fortieth observation.

The relationship between time ($x$) and capacitance ratio ($y$) is assumed to be exponential. The data are plotted in Figure 2.2 where an exponential relationship is indicated. The log transformation is the linearizing transformation. Plotting log($y$) against $x$ (Figure 2.3) shows a heteroscedastic trend not as apparent in Figure 2.2 and suggests that the log scale is not appropriate. We assume model (2.1) where $f(x_1, \beta) = \beta_1 \exp(\beta_2 x_1)$ and the parameters $(\beta_1, \beta_2, \sigma, \lambda, \theta)$ are unknown. The parameters $(\beta_1, \beta_2)$ are estimated for fixed $(\lambda, \theta)$ using an iterative generalized least squares procedure. See Carroll and Ruppert (1987) for a discussion of generalized least squares procedures. A grid of values for $(\lambda, \theta)$ is selected ranging from -0.25 to 1.75 for $\lambda$ and -1 to 1 for $\theta$. Table 2.1 contains the values of the loglikelihoods computed at $(\hat{\beta}, \hat{\sigma}, \lambda, \theta)$. See Section 2.3.3 for a discussion of how to obtain confidence regions for this problem.

From Table 2.1, the parameters for which the maximum value of the loglikelihood is attained over the grid are $(\lambda, \theta) = (1, 0)$, the parameters corresponding to no transformation. The $(\lambda, \theta)$ pairs that fall within the 99% confidence region are such that for any given $\theta$, $\lambda = 1 + \theta$, that is, a change in one parameter corresponds to a similar change in the other. This indicates that the model is overparametrized and supports the small-$\sigma$ assumption.

In addition to the no-transform model we consider two other models, $(\lambda, \theta) = (0, 0)$ and $(0, -1)$. The $(\lambda, \theta) = (0, 0)$ model, the linearized model one would have had the original errors been
homoscedastic and entered the model in a multiplicative fashion, yields

$$\log(y_i) = \log \beta_1 + \beta_2 x_i + \sigma \varepsilon_i.$$  

This model ignores the possibility of heteroscedastic errors or the creation of such errors by the transformation. If the true model is such that no transformation is needed (the errors are homoscedastic and additive) but a log transformation is incorrectly applied to both sides, then heteroscedasticity is induced and should be taken into account in the model by using $(\lambda, \theta) = (0, -1)$. For example, if $y_i = f(x_i, \beta) + \sigma \varepsilon_i$ is the correct model, applying logs to both sides yields

$$\log y_i = \log[f(x_i, \beta) + \sigma \varepsilon_i] = \log f(x_i, \beta) + \frac{\sigma \varepsilon_i}{f(x_i, \beta)}.$$  

Weighting by the inverse power-of-the-mean removes the heteroscedasticity induced by the log transformation. This is the model typically used to model these data.

As illustrated in Table 1, neither $(\lambda, \theta) = (0, 0)$ nor $(0, -1)$ fall in the 99% confidence region but $(\lambda, \theta) = (0, -1)$ is on the same ridge of values where for a given $\lambda$ the maximum is attained over the $\theta$'s. The loglikelihood value for $(\lambda, \theta) = (0, 0)$ is small relative to the maximum over the grid, so one might expect it to do poorly in relationship to the other two.

The residual plot for $(\lambda, \theta) = (1, 0)$ (Figure 2.4) reveals no evidence of heteroscedasticity or skewness whereas for $(\lambda, \theta) = (0, 0)$ there is clear heteroscedasticity and the data are somewhat skewed to
the left (See Figure 2.5). The nonlinearity of the normal probability
plot (Figure 2.6) indicates that the assumption of normality may be
inappropriate for these data, although it is not clear how to
interpret such plots when the residuals are heteroscedastic. In the
case where \((\lambda, \theta) = (0, -1)\), the residuals do not exhibit any
heteroscedastic pattern but there are twenty fewer positive residuals
indicating a slight right skewness.

See Table 2.2 for computed values of the skewness, kurtosis and
Spearman's rank correlation coefficient \((\rho)\). These values are
consistent with the residual plots. The \((\lambda, \theta) = (0, 0)\) model yields
residuals that are skewed and heteroscedastic and have a large value
for kurtosis. The \((\lambda, \theta) = (0, -1)\) model competes favorably with the no
transformation model as far as heteroscedasticity is concerned but is
more skewed and has a somewhat inflated value for kurtosis. The
skewness and kurtosis for the no transformation model are closer to 0
than the other two and do as well as the \((\lambda, \theta) = (0, -1)\) model in
controlling the heteroscedasticity.

Analyzing the data in the original scale with no weighting yields
residuals that are symmetric about zero with no evidence of heterosce-
dasticity. Applying a log transformation with weights formed by
\(f^{-1}(x_1, \beta)\) removes the induced heteroscedasticity but causes the
residuals to be skewed. Simply taking a log transformation without
accounting for the induced heteroscedasticity produces residuals that
are not only heteroscedastic, but clearly do not satisfy the normal
assumptions. This example demonstrates that it is not necessary to
linearize nonlinear models even when it is possible to do so.
2.2.2 The Skeena River Sockeye Salmon Data

These data originate from Ricker and Smith (1975) and represent the stock-recruitment of the Skeena River sockeye salmon from 1940 to 1967. For each year there are two variables, \( x \) = spawners (mature fish) and \( y \) = the number of recruits into the fishery. A preliminary analysis indicated that observations 12 and 16 were extreme points and are removed for this analysis. Both observations have small \( x \) values. Justification for their removal is that in 1951 (observation 12) there was a rockslide that blocked the river and prevented the fish from getting upriver to spawn. Since it takes about four years for the sockeye salmon to mature and return to the river to spawn, the observation in 1955 (observation 16) is also severely affected. Removal of these two points leaves 26 points for the analysis. The data are listed in Table 2.3 and are plotted in Figure 2.7.

We fit model (2.1) over a grid of values for \((\lambda, \theta)\) where we use the Ricker model, \( f(x_1, \beta) = \beta_1 x_1 \exp(\beta_2 x_1) \). Table 2.4 illustrates the 95% confidence region for the loglikelihood where \( \lambda \) ranges from \(-.75\) to \(1.5\) and \( \theta \) ranges from \(-.5\) to \(2.25\). Note that the confidence region in this example is much larger than that in the previous example. This is due in part to the small sample size and in part to the range of \( x \) and \( y \) relative to the model. When \( x \) and \( y \) vary over a wide range of values, the parameters are better determined by the data than when the range of \( x \) and \( y \) is narrow, which consequently results in smaller confidence regions.

The maximum likelihood estimators for \((\lambda, \theta)\) over the grid are \((.5, 1.25)\). In Figure 2.7 the mean regression line corresponding to \((\lambda, \theta) = (1, 0)\) and the median regression line, \((\lambda, \theta) = (\lambda, 0.75)\) are
plotted. The Ricker model is such that as $x$ increases $y$ also increases until $x$ reaches a maximum at $-1/\beta_2$. After that point, increasing $x$ results in decreasing $y$. A possible explanation for this behavior is that after a certain number of fish enter the system overcrowding occurs and food and oxygen supplies are depleted resulting in fewer recruits. It is important for those managing the fishery to know where this maximum will occur so that the amount of fish can be regulated if necessary. Thus we are interested in estimating $\beta_2$ as efficiently as possible. For these two models $\hat{\beta}_2(\lambda = 1, \theta = 0) = -0.00085$ and $\hat{\beta}_2(\lambda = .5, \theta = 1.25) = -0.00064$. This corresponds to maximum values on the Ricker curve of 1176.5 and 1562.5 respectively for the mean and the median regression.

The no-transformation model, $(\lambda, \theta) = (1,0)$, does not fall within the confidence region. Taylor series expansions about $\sigma$ yield the following small-$\sigma$ equivalent models: $(-.75,0)$ - transform-both-sides with no weighting, $(1,1.75)$ - weight only, and $(0,0.75)$ - linearize the model with the log transformation but weight to offset the induced heteroscedasticity. It is also of interest to look at the $(\lambda, \theta) = (0,0)$ model to see how badly one does if the model is linearized but the log induced heteroscedasticity is not taken into account. Table 2.5 contains the values of skewness, kurtosis and Spearman's rank correlation coefficient for these six models. It is interesting to see how well the maximum likelihood estimators transform to both symmetry and homoscedasticity, at least in terms of these numerical measures. As expected the residual plots exhibit no heteroscedasticity, whereas the residual plots for the no transformation model show marked heteroscedasticity. See Figures 2.8 and 2.9.
for the residual plots for these two models.

In several models, observation 4 is associated with a large residual. Carroll and Ruppert (1986) in their robust analysis found this point to be influential; see also Section 4.3. However in this analysis the leverage is not large. See Table 2.3 for computed leverages for the no transformation model and the maximum likelihood model. Whereas most of the nearby points are fit well, observation 4 is poorly fit. See Table 2.6 for the effect on the summary statistics when this point is removed from the analysis. In Table 2.6, \( \theta \) has been set equal to 0 and \( \lambda \) varies. Since observation 4, the third smallest \( x \) value, has a relatively large residual and the general heteroscedastic trend of the residuals is to grow as the mean increases, it is not surprising that the value for Spearman's \( \rho \) increases (more heteroscedastic) when observation 4 is removed from the analysis. Also the value for skewness generally decreases but the differences between the models stay the same.

It is interesting that the model \((\lambda, \theta) = (-.75, 0)\) given by small-\( \sigma \) Taylor manipulations is not the best \( \theta = 0 \) model when one considers how well the model accounts for both heteroscedasticity and skewness. It appears that this model is trying to account for heteroscedasticity by overtransforming at the expense of skewness. Closer investigation of the likelihood surface indicates that given \( \theta = 0 \), the loglikelihood is maximized for \( \lambda \) between 0 and \(-.25\), which agrees with the summary statistics given in Table 2.6. This is a further indication that care should be used when one is not in a small-\( \sigma \) setting.

In Table 2.7, \( \lambda \) has been fixed at 1 and \( \theta \) ranges from 0 to 2.25.
Heteroscedasticity is adequately accounted for when $\theta = 1.75$ ($\rho = .06$) or $\theta = 2$ ($\rho = -.02$). It is interesting to note that as $\theta$ changes the skewness remains fairly constant while $\theta$ accounts for the heteroscedasticity in the model. This is not the case when $\theta$ is set equal to 0 and $\lambda$ varies (Table 2.6). When $\lambda$ is the only parameter in the model used to transform to approximate normality, it plays the dual role of trying to account for both skewness and heteroscedasticity, whereas $\theta$ alone attempts to model only the heteroscedasticity. The $(\lambda, \theta) = (1, 1.75)$ model is the model obtained through the small-$\sigma$ Taylor expansions and appears to be adequate given $\lambda = 1$ unlike the small-$\sigma$ derived model for $\theta = 0$.

These data have been extensively analyzed by Carroll and Ruppert (1986, 1987) and Ruppert and Carroll (1985). They use the Ricker model and do a robust analysis of the full data set and also analyze the data with observation 12 deleted. From the robust analysis they find observations 4, 5, 12, 16, and 18 to be influential points, but they only remove observation 12 (the rockslide year) for the standard nonrobust analysis. Observation 16 has an exceptionally small number of spawners since the spawning stock that year came from the recruitment of the rockslide year, but since observation 16 had normal recruitment it was retained for their analysis. Carroll and Ruppert (1987) find for the transform-both-sides model with no weighting ($\theta = 0$) that $\hat{\lambda}_{\text{MLE}} = -.2$, supporting our results which indicate a value between 0 and -.25. They have also calculated $\hat{\lambda}_{\text{sk}}$, the value of $\lambda$ that produces symmetric residuals, and $\hat{\lambda}_{\text{het}}$ which transforms to homoscedasticity, to be .45 and -.86 respectively. According to our analysis (see Table 2.6) $\hat{\lambda}_{\text{sk}} \approx .5$ and $\hat{\lambda}_{\text{het}} \approx -1$. Note that our table
stops with $\lambda = -1$, the actual estimate for $\lambda_{het}$ may be smaller. It is interesting that as in Carroll and Ruppert (1987) the transform-both-sides model with power weighting yields $\hat{\lambda}_{MLE}$ to be close to $\hat{\lambda}_{sk}$, indicating that skewness alone is determining $\hat{\lambda}$. While our estimates differ slightly from those of Carroll and Ruppert (1987) in that we chose to delete observation 16 as well as observation 12, the essential characteristics of the analysis agree.

2.2.3 Population A Sockeye Salmon Data

This is another example of spawner-recruit data from the fisheries industry. Since permission to identify the stock was refused by the original source, Ruppert and Carroll (1985) and Carroll and Ruppert (1987) refer to these data as the "Population A" data. There are 28 years of data which are listed in Table 2.8 and plotted in Figure 2.10. The number of recruits for the last observation year is extremely small relative to the other observations. Carroll and Ruppert (1987) suspected that the collection for that year was not complete and subsequently deleted the observation. We do the same here.

As in Ruppert and Carroll (1985) we look at both the Ricker model $f(x_1, \beta) = \beta_1 x_1 \exp(\beta_2 x_1)$, and the Beverton-Holt model, $f(x_1, \beta) = 1/(\beta_1 + \beta_2/x_1)$. See Table 2.9 for 95% confidence regions for the two models. The maximum likelihood estimator for the Ricker model is $(\lambda, \theta) = (0.25, 1)$ while that for the Beverton-Holt model is $(\lambda, \theta) = (0.25, 3)$. For each model we consider the no transformation model, the linearizing transformation, the maximum likelihood and the log transformation. See Table 2.10. As with the Skeena River data, it is
interesting to see how well the maximum likelihood estimates simultaneously account for both skewness and heteroscedasticity. The data under either model are somewhat insensitive to the amount of weighting used but are greatly affected by the value of $\lambda$. This is not surprising since the confidence regions are narrow for $\lambda$ but long for $\theta$, especially for the Beverton-Holt model. The Beverton-Holt median regression line (Figure 2.10) quickly approaches its asymptote and is nearly constant over most of the range of $x$. Therefore a power-of-the-mean variance model is effectively a constant variance model and the parameter $\theta$ is not well determined.

For these data the median coefficient of variation is approximately .9 indicating that the small- $\sigma$ assumption may not be appropriate. This is further supported in that the relationship $\theta \approx 1 - \lambda$ does not appear to hold based on the loglikelihood grid. It is reasonable to assume that if $\lambda$ and $\theta$ are both included in the model then $\lambda$ will model the skewness and $\theta$ will control the heteroscedasticity.

Consider the Beverton-Holt model. The Spearman's $\rho$ value for the $(\lambda, \theta)$ combinations (.25, 0) and (.25, 3) are respectively .20 and .18. According to this numerical measure both models seem to do a comparable job accounting for heteroscedasticity. Skewness and kurtosis are closer to 0 for the maximum likelihood estimators but the difference between the two models is slight. As was anticipated by the extremely long confidence interval, the inclusion of $\theta$ does not produce significantly different results than the $\theta = 0$ model. This analysis seems to indicate that skewness is the predominant feature of the data rather than heteroscedasticity. Compare Figure 2.10 with the
Skeena River data (Figure 2.7). Relative to the plotted regression lines, the Skeena River data exhibit a clear heteroscedastic trend not as apparent in the Population A data where the data appear to be skewed relative to the plotted regression lines. This skewness is further evidenced by the median regression being smaller than the mean regression.

As in the Skeena River example, the managers of the fishery are interested in the efficient estimation of the parameters. The behavior of the Beverton-Holt model is quite different from that of the Ricker model. The Beverton-Holt curve does not exhibit the downward trend as the Ricker curve does. In the Beverton-Holt model the number of recruits seems to stabilize at an asymptote determined by the parameters. We see from Figure 2.10 that this asymptote is very different depending on whether or not we use transformations.

2.2.4 Discussion

We see from these examples that the transform-both-sides power-of-the-mean model is important in nonlinear regression. While it may effectively reduce to a transform-both-sides model, or a power-of-the-mean model, or a no-transformation model, the transform-both-sides power-of-the-mean model is useful in discovering characteristics of the data that might otherwise go undetected in a model that is initially limited to one or the other.

We saw that linearizing a nonlinear model may not always be appropriate (Example 1) and the transform-both-sides power-of-the-mean approach was able to recognize that no transformation was necessary. In Examples 2 and 3 we saw that skewness and heteroscedasticity were
major factors of the data. The choice of $(\lambda, \theta)$ allows error to enter
the model in different ways and the effect of skewness and heteroscedasticity is better understood. By choosing a $(\lambda, \theta)$ combination such
that the errors are approximately normally distributed we can achieve
more efficient estimators than would be possible if the model were not
transformed. Improved estimators, while in addition to having nice
statistical properties, can lead to savings in cost and resources and
may even change management policies as in the fisheries examples.
Thus the transform-both-sides power-of-the-mean model is a tool
whereby insight into the error behavior can be gained and more
efficient modeling can occur.

2.3 Appendix

2.3.1 Heuristic Sketch of the Carroll-Ruppert Efficiency Result

We sketch here a proof of Carroll and Ruppert's (1984) result
adapted to the transform-both-sides regression model with power-of-
the-mean variance. Basically the idea is to compute a third estimator
for $\beta_0$, the true parameter, and compute its efficiency with respect to
$(\hat{\beta} | \lambda, \theta \text{ known})$. This gives a bound on the efficiency of the maximum
likelihood estimators.

Let $w_1, \ldots, w_n$ be positive numbers and let $\hat{\beta}_1$ be any point that
minimizes $\sum w_i |y_i - f(x_i, \hat{\beta}_1)|$. Under (2.1), $f(x_i, \beta_0)$ is the unique
median of $y_i$. so $\hat{\beta}_1$ will be consistent under certain regularity
conditions. If $w_i$ is set equal to the Jacobian of the transformation
evaluated at the true parameters, $\beta_0$, $\lambda_0$, $\theta_0$, and $y = f(x_i, \beta_0)$, that
is, $w_i$ is proportional to the density of $[y_i - f(x_i, \beta_0)]$ at its
median, zero, then \( \sqrt{N} \frac{\hat{\beta}_1 - \beta_0}{\sigma_0} \) \( \Rightarrow \) \( N(0, \pi/2 S_1^{-1}) \), where

\[
S_1 = N^{-1} \sum_{i=1}^{N} \frac{\frac{\partial}{\partial \beta} f^{(\lambda)}(x_i, \beta) \frac{\partial}{\partial \beta} f^T(x_i, \beta)}{f^2(\beta)(x_i, \beta)}.
\]

From (2.5), the \( \text{Cov}(\hat{\beta} | \lambda, \theta) = S_1^{-1} \). The asymptotic optimality of maximum likelihood estimates shows that \( S_1^{-1} \leq \text{Cov}(\hat{\beta}) \leq \pi/2 S_1^{-1} \), where the inequalities are in the sense of positive definiteness.

2.3.2 A Heuristic Derivation of (2.3)

Fix \( \lambda \) and \( \theta \). We rely on the calculations of Chapter 7 of Carroll and Ruppert (1987), who indicate by Taylor series that for fixed \((\lambda, \theta)\)

\[
N^{1/2} \frac{\hat{\beta}_{\text{GLS}} - \beta}{\sigma} - N^{-1/2} S_1^{-1} \sum_{i=1}^{N} \varepsilon_i^1 \frac{\partial f^{(\lambda)}(x_i, \beta)}{\partial \beta} f^\theta(x_i, \beta),
\]

where

\[
S_1 = N^{-1} \sum_{i=1}^{N} \frac{\frac{\partial}{\partial \beta} f^{(\lambda)}(x_i, \beta) \frac{\partial}{\partial \beta} f^T(x_i, \beta)}{f^2(\beta)(x_i, \beta)}.
\]

McCullagh (1983) has a similar calculation. By formal likelihood theory, for fixed \((\sigma, \lambda, \theta)\),

\[
N^{1/2} \left[ \begin{array}{c}
\hat{\beta} - \beta \\
\hat{\sigma} - \sigma
\end{array} \right] \approx \left[ \sigma^2 I(\beta, \sigma) \right]^{-1} \sigma N^{-1/2} \left[ \begin{array}{c}
\varepsilon_{\beta}(\beta, \sigma) \\
\varepsilon_{\sigma}(\beta, \sigma)
\end{array} \right],
\]

where \( I(\beta, \sigma) \) is the information matrix for a single observation assuming the \( x \)'s are random. From previous calculations
\[ \sigma^2 I(\beta, \sigma) \rightarrow \begin{bmatrix} S_1 & 0 \\ 0 & 2 \end{bmatrix}. \]

\[ A_1 = \sigma N^{-1/2} \frac{\partial}{\partial \beta} \ell(\beta, \sigma) \]
\[ = N^{-1/2} \sum_{i=1}^{N} \epsilon_i f_\beta(x_i, \beta)[f(x_i, \beta)]^{\lambda-\theta-1} \]
\[ + \sigma \theta N^{-1/2} \sum_{i=1}^{N} \frac{(\epsilon_i^2 - 1) f_\beta(x_i, \beta)}{f(x_i, \beta)}. \]

and
\[ A_2 = \sigma N^{-1/2} \frac{\partial}{\partial \sigma} \ell(\beta, \sigma) \]
\[ = N^{-1/2} \sum_{i=1}^{N} (\epsilon_i^2 - 1). \]

Thus
\[ N^{1/2} \begin{bmatrix} \hat{\beta}_{\text{MLE}} - \beta \\ \hat{\sigma}_{\text{MLE}} - \sigma \end{bmatrix} \approx [\sigma^2 I(\beta, \sigma)]^{-1} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \]
\[ = \begin{bmatrix} S_1 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \]
\[ + \left[ (\sigma^2 I(\beta, \sigma))^{-1} - \begin{bmatrix} S_1 & 0 \\ 0 & 1/2 \end{bmatrix} \right] \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}. \]

Since \( A_1 = O_p(1) \) and \( A_2 = O_p(1) \), and since
\[ \left[ (\sigma^2 I(\beta, \sigma))^{-1} - \begin{bmatrix} S_1 & 0 \\ 0 & 1/2 \end{bmatrix} \right] \xrightarrow{p} 0 \]
we get
\[ N^{1/2} \begin{bmatrix} \hat{\beta}_{\text{MLE}} - \beta \\ \hat{\sigma}_{\text{MLE}} - \sigma \end{bmatrix} \approx \begin{bmatrix} S_1 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}. \]
Therefore
\[ N^{1/2}(\hat{\beta}_{\text{MLE}} - \beta)/\sigma \approx S_1^{-1}A_1, \]
and,
\[ N^{1/2}(\hat{\beta}_{\text{MLE}} - \hat{\beta}_{\text{GLS}})/\sigma \approx \sigma \theta S_1^{-1} N^{-1/2} \sum_{i=1}^{N} (\epsilon_i^2 - 1) f_{\beta}(x_i, \beta)/f(x_i, \beta). \]

Now let \( \sigma \to 0 \) and get a stronger form of (2.3) since
\[ N^{-1/2} \sum_{i=1}^{N} (\epsilon_i^2 - 1) f_{\beta}(x_i, \beta)/f(x_i, \beta) = O_p(1). \]

\[ \blacksquare \]

2.3.3 \textbf{Approximate Confidence Regions for } (\lambda, \theta) \textbf{)

We compute an approximate \((1-\alpha)100\% \) confidence region in the following way. Define
\[ L_{\text{max}}(\lambda, \theta) = \ell(\beta(\lambda, \theta), \sigma(\lambda, \theta), \lambda, \theta) \]
to be the loglikelihood maximized with respect to \( \beta \) and \( \sigma \) for fixed \( \lambda \) and \( \theta \). The loglikelihood ratio is given by
\[ LR(\lambda_0, \theta_0) = L_{\text{max}}(\hat{\lambda}, \hat{\theta}) - L_{\text{max}}(\lambda_0, \theta_0) \]
where \((\hat{\lambda}, \hat{\theta})\) are those values of \((\lambda, \theta)\) over the grid for which the loglikelihood is maximized. In the usual maximum likelihood estimation scheme (Rao, 1973, section 6.e), \( 2LR(\lambda_0, \theta_0) \) is asymptotically \( \chi^2_2 \) with level \( \alpha \) under the null hypothesis that \((\lambda, \theta) = (\lambda_0, \theta_0)\). We form an approximate confidence region by finding all those values of \((\lambda_0, \theta_0)\) such that
\[ 2[L_{\text{max}}(\hat{\lambda}, \hat{\theta}) - L_{\text{max}}(\lambda_0, \theta_0)] \leq \chi^2_2(1 - \alpha). \]

Two points need to be addressed. First we have not computed maximum likelihood estimators but instead have used a generalized least squares estimation scheme. Although the two estimators are asymptotically equivalent as \( \sigma \to 0 \), it is not an automatic consequence that the
loglikelihood ratio evaluated at the generalized least squares estimators is asymptotically $\chi^2$. A second point is that the maximum is attained over a grid and is thus not exact.

Therefore we must qualify our usage of the normal theory confidence regions. It can be shown (see below) that as $N \to \infty$ and then $\sigma \to 0$,

$\begin{align*}
2[L_{\text{max}}(\hat{\lambda}, \hat{\theta}) - L_{\text{max}}(\lambda_0, \theta_0)] & \Rightarrow \chi^2_1.
\end{align*}$

It is not clear what effect the overparamterization and subsequent loss of a degree of freedom signifies. For the examples considered there is little or no difference in the confidence region when based on one or two degrees of freedom. A difference would be observed if our estimation of $\lambda$ and $\theta$ were not limited to a grid of values.

Calculations for (2.5):

Recall that $\theta = (\beta, \sigma, \lambda, \theta)$. The usual normal theory likelihood ratio test $2[\ell(\hat{\theta}_{\text{MLE}}) - \ell(\theta)]$ is asymptotically $\chi^2$ under the null hypothesis. Cox and Hinkley (1974) show that the normal theory likelihood ratio test $2[\ell(\hat{\theta}_{\text{MLE}}) - \ell(\hat{\theta}_0)]$ is asymptotically $\chi^2$ with degrees of freedom given by the number of parameters estimated under the null hypothesis. In order to show that $2[\ell(\hat{\theta}_{\text{GLS}}) - \ell(\hat{\theta}_0)] \sim \chi^2_1$ when $N \to \infty$ and then $\sigma \to 0$, rewrite

$\ell(\hat{\theta}_{\text{GLS}}) - \ell(\hat{\theta}_0) = [\ell(\hat{\theta}_{\text{GLS}}) - \ell(\hat{\theta}_{\text{MLE}})] + [\ell(\hat{\theta}_{\text{MLE}}) - \ell(\theta)]$

$\quad - [\ell(\hat{\theta}_0) - \ell(\theta)].$

We must show

(a) $\ell(\hat{\theta}_{\text{GLS}}) - \ell(\hat{\theta}_{\text{MLE}}) \overset{p}{\longrightarrow} 0$, and

(b) $2[\ell(\hat{\theta}_{\text{MLE}}) - \ell(\theta)] - 2[\ell(\hat{\theta}_0) - \ell(\theta)] \Rightarrow \chi^2_1$. 

Part (a) is an immediate consequence of the discussion of Section 2.2.3. In order to show (b) we first define additional notation. Let all estimates be maximum likelihood estimates, and define \( \theta_1 = (\beta, \sigma) \), \( \theta_2 = (\lambda, \theta) \), and \( \hat{\theta}_{01} = (\beta_0, \sigma_0 | \lambda = \lambda_0, \theta = \theta_0) \). Let \( \mathcal{J} \) be the information matrix, partitioned such that

\[
\begin{align*}
\mathcal{J}_{11} &= E \begin{bmatrix} \ell_{\beta\beta} & \ell_{\beta\lambda} \\ \ell_{\beta\lambda} & \ell_{\lambda\lambda} \end{bmatrix}, & \mathcal{J}_{12} &= E \begin{bmatrix} \ell_{\beta\lambda} & \ell_{\beta\theta} \\ \ell_{\beta\theta} & \ell_{\lambda\theta} \end{bmatrix}, \\
\mathcal{J}_{21} &= E \begin{bmatrix} \ell_{\lambda\beta} & \ell_{\lambda\lambda} \\ \ell_{\lambda\lambda} & \ell_{\lambda\theta} \end{bmatrix}, & \mathcal{J}_{22} &= E \begin{bmatrix} \ell_{\lambda\theta} & \ell_{\lambda\lambda} \\ \ell_{\lambda\lambda} & \ell_{\theta\theta} \end{bmatrix},
\end{align*}
\]

and let \( \text{Cov}(\theta) = \Sigma^* \). Thus

\[
\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} \sim N \left( \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \mathcal{J}^{-1} \Sigma^* \mathcal{J}^{-1} = \Sigma \right)
\]

and \( (\hat{\theta}_{01} | \theta_2) \sim N(\hat{\theta}_1 - \Sigma^{1/2}_{12} \Sigma^{1/2}_{22}(\hat{\theta}_2 - \theta_2), \Sigma_Y) \) for some covariance matrix \( \Sigma_Y \). (See Rao 1973, p. 522.) Write \( \ell(\hat{\theta}_\text{MLE}) - \ell(\theta) - [\ell(\hat{\theta}_0) - \ell(\theta)] = [\ell(\hat{\theta}_1, \hat{\theta}_2) - \ell(\theta_1, \theta_2)] - [\ell(\hat{\theta}_{01}, \theta_2) - \ell(\theta_{01}, \theta_2)] \). Expand in a Taylor series,

\[
[\ell(\hat{\theta}_1, \hat{\theta}_2) - \ell(\theta_1, \theta_2)] \approx \frac{1}{2} \begin{bmatrix} (\hat{\theta}_1 - \theta_1) \\ (\hat{\theta}_2 - \theta_2) \end{bmatrix}^T \begin{bmatrix} \ell(\theta_1, \theta_2) \end{bmatrix}
\]

and

\[
[\ell(\hat{\theta}_{01}, \theta_2) - \ell(\theta_{01}, \theta_2)] \approx \frac{1}{2} \begin{bmatrix} (\hat{\theta}_{01} - \theta_{01}) \\ 0 \end{bmatrix}^T \begin{bmatrix} \ell(\theta_1, \theta_2) \end{bmatrix}
\]

where \( \ell(\theta_1, \theta_2) = \sigma^2 \ell(\theta) / \partial \theta \partial \theta^T \). Combining the above we get...
(2.6) \[ 2[\ell(\hat{\theta}_{\text{MLE}}) - \ell(\hat{\theta}_0)] \approx \begin{bmatrix} (\hat{\theta}_1 - \theta_1)^T \\ (\hat{\theta} - \theta_2)^T \\ (\hat{\theta} - \theta_2)^T \end{bmatrix} \cdot \begin{bmatrix} (\hat{\theta}_1 - \theta_1)^T \\ (\hat{\theta} - \theta_2)^T \end{bmatrix} \\
- \begin{bmatrix} (\hat{\theta}_{01} - \theta_1)^T \\ 0^T \end{bmatrix} \cdot \begin{bmatrix} (\hat{\theta}_{01} - \theta_1)^T \\ 0^T \end{bmatrix}. \]

Replace \( \hat{\theta}_{01} \) with \( \hat{\theta}_1 - \bar{\Sigma}_{122} (\hat{\theta}_2 - \theta_2) \). \( \bar{\Sigma} \) is not of full rank and must be calculated using generalized inverses. Substitution and simplification reduce (2.6) to \( (\kappa + 2)/2 (\hat{\theta}_2 - \theta_2)^T \bar{\Sigma}_{22} (\hat{\theta}_2 - \theta_2) \), where \( \kappa = E[\epsilon_1^4] - 3 \) represents the kurtosis. Thus \( 4/\kappa + 2 \) [\( \ell(\hat{\theta}_{\text{MLE}}) - \ell(\hat{\theta}_0) \)] is asymptotically \( \chi^2 \) with one degree of freedom, the dimension of \( \bar{\Sigma}_{22} \).

2.4 Tables and Figures

The tables and figures for the Chapter 2 examples follow.
Table 2.1: Loglikelihood computed at \((\beta, \lambda, \sigma)\) for Capacitance Data.

<table>
<thead>
<tr>
<th>(\lambda)</th>
<th>(\theta)</th>
<th>-1</th>
<th>-.75</th>
<th>-.5</th>
<th>-.25</th>
<th>0</th>
<th>.25</th>
<th>.5</th>
<th>.75</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-.25</td>
<td></td>
<td>525.4</td>
<td>471.7</td>
<td>405.3</td>
<td>329.7</td>
<td>246.5</td>
<td>155.0</td>
<td>51.4</td>
<td>-75.2</td>
<td>***</td>
</tr>
<tr>
<td>0.00</td>
<td></td>
<td>578.9</td>
<td>550.1</td>
<td>502.0</td>
<td>440.2</td>
<td>368.4</td>
<td>288.6</td>
<td>200.7</td>
<td>102.1</td>
<td>-14.5</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>590.7</td>
<td>590.3</td>
<td>566.4</td>
<td>522.9</td>
<td>465.1</td>
<td>396.9</td>
<td>320.6</td>
<td>236.7</td>
<td>143.5</td>
</tr>
<tr>
<td>0.50</td>
<td></td>
<td>569.2</td>
<td>594.4</td>
<td>596.9</td>
<td>576.4</td>
<td>536.4</td>
<td>481.8</td>
<td>416.8</td>
<td>343.7</td>
<td>263.7</td>
</tr>
<tr>
<td>0.75</td>
<td></td>
<td>524.8</td>
<td>569.6</td>
<td>595.9</td>
<td>600.0</td>
<td>581.7</td>
<td>544.2</td>
<td>492.1</td>
<td>429.6</td>
<td>359.5</td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td>465.6</td>
<td>524.1</td>
<td>568.9</td>
<td>595.6</td>
<td>600.5</td>
<td>583.5</td>
<td>547.5</td>
<td>497.2</td>
<td>436.8</td>
</tr>
<tr>
<td>1.25</td>
<td></td>
<td>396.9</td>
<td>464.6</td>
<td>522.8</td>
<td>567.3</td>
<td>593.8</td>
<td>598.9</td>
<td>582.3</td>
<td>547.2</td>
<td>498.2</td>
</tr>
<tr>
<td>1.50</td>
<td></td>
<td>321.1</td>
<td>395.7</td>
<td>463.2</td>
<td>521.0</td>
<td>565.0</td>
<td>590.8</td>
<td>595.4</td>
<td>578.6</td>
<td>543.9</td>
</tr>
<tr>
<td>1.75</td>
<td></td>
<td>240.9</td>
<td>320.3</td>
<td>394.5</td>
<td>461.6</td>
<td>518.7</td>
<td>561.8</td>
<td>586.7</td>
<td>590.2</td>
<td>572.8</td>
</tr>
</tbody>
</table>

*** Convergence not obtained within time limit
— 99% confidence region
== maximum value for loglikelihood over grid

Table 2.2: Summary statistics for Capacitance Data

<table>
<thead>
<tr>
<th>(\lambda)</th>
<th>(\theta)</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>(\rho)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0</td>
<td>-.115</td>
<td>.4154</td>
<td>.134</td>
<td>.0210</td>
<td></td>
</tr>
<tr>
<td>0 0</td>
<td>1.774</td>
<td>10.3921</td>
<td>.673</td>
<td>.0001</td>
<td></td>
</tr>
<tr>
<td>0 -1</td>
<td>.624</td>
<td>1.6006</td>
<td>.133</td>
<td>.0221</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.1: Capacitance Data. Schematic of voltage and waveforms associated with repetively changing the voltage and analyzing the resulting capacitance ratio.
Figure 2.2: Capacitance Data. Scatterplot of capacitance ratio by time.
Figure 2.3: Capacitance Data. Scatterplot of $\log(\text{capacitance ratio})$ by time.
Figure 2.4: Capacitance Data. Plot of studentized residuals from the no transformation model, $(\lambda, \theta) = (1, 0)$.
Figure 2.5: Capacitance Data. Plot of studentized residuals from the log transformation model, $(\lambda, \theta) = (0, 0)$. 
Figure 2.6: Capacitance Data. Normal probability plot for the log transformation model.
Table 2.3: Skeena River Sockeye Salmon Data

<table>
<thead>
<tr>
<th>Observation</th>
<th>Year</th>
<th>Spawners</th>
<th>Recruits (λ, θ)=(1.0)</th>
<th>(λ, θ)=(.5,1.25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1940</td>
<td>963</td>
<td>2215</td>
<td>.1469</td>
</tr>
<tr>
<td>2</td>
<td>1941</td>
<td>572</td>
<td>1334</td>
<td>.0501</td>
</tr>
<tr>
<td>3</td>
<td>1942</td>
<td>305</td>
<td>800</td>
<td>.0680</td>
</tr>
<tr>
<td>4</td>
<td>1943</td>
<td>272</td>
<td>438</td>
<td>.0648</td>
</tr>
<tr>
<td>5</td>
<td>1944</td>
<td>824</td>
<td>3071</td>
<td>.0751</td>
</tr>
<tr>
<td>6</td>
<td>1945</td>
<td>940</td>
<td>957</td>
<td>.1318</td>
</tr>
<tr>
<td>7</td>
<td>1946</td>
<td>486</td>
<td>934</td>
<td>.0592</td>
</tr>
<tr>
<td>8</td>
<td>1947</td>
<td>307</td>
<td>971</td>
<td>.0681</td>
</tr>
<tr>
<td>9</td>
<td>1948</td>
<td>1066</td>
<td>2257</td>
<td>.2304</td>
</tr>
<tr>
<td>10</td>
<td>1949</td>
<td>480</td>
<td>1451</td>
<td>.0599</td>
</tr>
<tr>
<td>11</td>
<td>1950</td>
<td>393</td>
<td>686</td>
<td>.0679</td>
</tr>
<tr>
<td>12</td>
<td>1951</td>
<td>176</td>
<td>127</td>
<td>**</td>
</tr>
<tr>
<td>13</td>
<td>1952</td>
<td>237</td>
<td>700</td>
<td>.0593</td>
</tr>
<tr>
<td>14</td>
<td>1953</td>
<td>700</td>
<td>1381</td>
<td>.0489</td>
</tr>
<tr>
<td>15</td>
<td>1954</td>
<td>511</td>
<td>1393</td>
<td>.0563</td>
</tr>
<tr>
<td>16</td>
<td>1955</td>
<td>87</td>
<td>363</td>
<td>**</td>
</tr>
<tr>
<td>17</td>
<td>1956</td>
<td>370</td>
<td>668</td>
<td>.0690</td>
</tr>
<tr>
<td>18</td>
<td>1957</td>
<td>448</td>
<td>2067</td>
<td>.0634</td>
</tr>
<tr>
<td>19</td>
<td>1958</td>
<td>819</td>
<td>644</td>
<td>.0734</td>
</tr>
<tr>
<td>20</td>
<td>1959</td>
<td>799</td>
<td>1747</td>
<td>.0671</td>
</tr>
<tr>
<td>21</td>
<td>1960</td>
<td>273</td>
<td>744</td>
<td>.0649</td>
</tr>
<tr>
<td>22</td>
<td>1961</td>
<td>936</td>
<td>1087</td>
<td>.1296</td>
</tr>
<tr>
<td>23</td>
<td>1962</td>
<td>558</td>
<td>1335</td>
<td>.0514</td>
</tr>
<tr>
<td>24</td>
<td>1963</td>
<td>597</td>
<td>1981</td>
<td>.0483</td>
</tr>
<tr>
<td>25</td>
<td>1964</td>
<td>848</td>
<td>627</td>
<td>.0842</td>
</tr>
<tr>
<td>26</td>
<td>1965</td>
<td>619</td>
<td>1099</td>
<td>.0472</td>
</tr>
<tr>
<td>27</td>
<td>1966</td>
<td>397</td>
<td>1532</td>
<td>.0677</td>
</tr>
<tr>
<td>28</td>
<td>1967</td>
<td>616</td>
<td>2086</td>
<td>.0473</td>
</tr>
</tbody>
</table>

** not included in the analysis
Table 2.4: ($\lambda, \theta$) Confidence Region for Skeena River Sockeye Salmon

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>-1.50</th>
<th>-1.25</th>
<th>-1.00</th>
<th>-0.75</th>
<th>-0.50</th>
<th>-0.25</th>
<th>0.00</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
<th>1.25</th>
<th>1.50</th>
<th>1.75</th>
<th>2.00</th>
<th>2.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>1.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>-0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>-0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>-0.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

+ 90% Confidence Region
*** MLE over the grid

Table 2.5: Summary Statistics for Skeena River Data

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\theta$</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Spearman's $\rho$ (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.75</td>
<td>-.406</td>
<td>-.121</td>
<td>.061 (.7665)</td>
</tr>
<tr>
<td>1.00</td>
<td>0.00</td>
<td>-.466</td>
<td>.486</td>
<td>.562 (.0028)</td>
</tr>
<tr>
<td>.50</td>
<td>1.25</td>
<td>.005</td>
<td>-.501</td>
<td>.107 (.6029)</td>
</tr>
<tr>
<td>0.00</td>
<td>.75</td>
<td>.396</td>
<td>-.500</td>
<td>.203 (.3189)</td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>.352</td>
<td>-.390</td>
<td>.389 (.0493)</td>
</tr>
<tr>
<td>-.75</td>
<td>0.00</td>
<td>.929</td>
<td>.114</td>
<td>.179 (.3821)</td>
</tr>
</tbody>
</table>
Table 2.6: Summary statistics for Skeena River data. \( \theta \) has been set to 0.

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>SKEWNESS</th>
<th>KURTOSIS</th>
<th>SPEARMAN’S ( p ) (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>-.47 [-.46]</td>
<td>.486 [.454]</td>
<td>.56 (.0028) [.57 (.0028)]</td>
</tr>
<tr>
<td>.75</td>
<td>-.26 [-.26]</td>
<td>.063 [.075]</td>
<td>.52 (.0063) [.57 (.0028)]</td>
</tr>
<tr>
<td>.50</td>
<td>-.05 [-.06]</td>
<td>-.214 [-.160]</td>
<td>.48 (.0135) [.53 (.0062)]</td>
</tr>
<tr>
<td>.25</td>
<td>.15 [.13]</td>
<td>-.359 [-.267]</td>
<td>.42 (.0311) [.50 (.0105)]</td>
</tr>
<tr>
<td>0.00</td>
<td>.35 [.31]</td>
<td>-.390 [-.266]</td>
<td>.39 (.0493) [.46 (.0219)]</td>
</tr>
<tr>
<td>-.25</td>
<td>.54 [.48]</td>
<td>-.319 [-.177]</td>
<td>.32 (.1043) [.38 (.0644)]</td>
</tr>
<tr>
<td>-.50</td>
<td>.74 [.65]</td>
<td>-.152 [-.020]</td>
<td>.28 (.1596) [.39 (.0524)]</td>
</tr>
<tr>
<td>-.75</td>
<td>.93 [.80]</td>
<td>.114 [.185]</td>
<td>.18 (.3821) [.35 (.0875)]</td>
</tr>
<tr>
<td>-1.00</td>
<td>1.09 [.95]</td>
<td>.494 [.421]</td>
<td>.13 (.5347) [.31 (.1350)]</td>
</tr>
</tbody>
</table>

[ ] represent values when observation 4 is deleted

Table 2.7: Summary statistics for Skeena River data. \( \lambda \) has been set to 1.

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>SKEWNESS</th>
<th>KURTOSIS</th>
<th>SPEARMAN’S ( p ) (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>-.466</td>
<td>.486</td>
<td>.56 (.0028)</td>
</tr>
<tr>
<td>.25</td>
<td>-.459</td>
<td>.307</td>
<td>.54 (.0043)</td>
</tr>
<tr>
<td>.50</td>
<td>-.452</td>
<td>.127</td>
<td>.47 (.0146)</td>
</tr>
<tr>
<td>.75</td>
<td>-.445</td>
<td>-.042</td>
<td>.38 (.0566)</td>
</tr>
<tr>
<td>1.00</td>
<td>-.438</td>
<td>-.181</td>
<td>.29 (.1444)</td>
</tr>
<tr>
<td>1.25</td>
<td>-.432</td>
<td>-.264</td>
<td>.26 (.1944)</td>
</tr>
<tr>
<td>1.50</td>
<td>-.424</td>
<td>-.256</td>
<td>.18 (.3859)</td>
</tr>
<tr>
<td>1.75</td>
<td>-.406</td>
<td>-.121</td>
<td>.06 (.7665)</td>
</tr>
<tr>
<td>2.00</td>
<td>-.365</td>
<td>.165</td>
<td>-.02 (.9037)</td>
</tr>
</tbody>
</table>
Figure 2.7: Skeena River Data. Scatterplot of recruits by spawners. The solid line represents the median regression line and the dotted line represents the mean regression line for the Ricker model.
SKEENA RIVER DATA
No Transformation Model

Figure 2.8: Skeena River Data. Scatterplot of studentized residuals by predicted values for the no transformation model.
SKEENA RIVER DATA
MLE Model

Figure 2.9: Skeena River Data. Scatterplot of studentized residuals by predicted values for the MLE model.
Table 2.8: Population A Sockeye Salmon Data

<table>
<thead>
<tr>
<th>Observation</th>
<th>Spawners</th>
<th>Recruits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19431</td>
<td>131635</td>
</tr>
<tr>
<td>2</td>
<td>12725</td>
<td>54928</td>
</tr>
<tr>
<td>3</td>
<td>32539</td>
<td>182836</td>
</tr>
<tr>
<td>4</td>
<td>12856</td>
<td>116935</td>
</tr>
<tr>
<td>5</td>
<td>28050</td>
<td>10933</td>
</tr>
<tr>
<td>6</td>
<td>8989</td>
<td>217870</td>
</tr>
<tr>
<td>7</td>
<td>28137</td>
<td>232492</td>
</tr>
<tr>
<td>8</td>
<td>21636</td>
<td>72378</td>
</tr>
<tr>
<td>9</td>
<td>8690</td>
<td>21572</td>
</tr>
<tr>
<td>10</td>
<td>20237</td>
<td>8801</td>
</tr>
<tr>
<td>11</td>
<td>38439</td>
<td>30715</td>
</tr>
<tr>
<td>12</td>
<td>8363</td>
<td>39208</td>
</tr>
<tr>
<td>13</td>
<td>7033</td>
<td>4623</td>
</tr>
<tr>
<td>14</td>
<td>4246</td>
<td>57472</td>
</tr>
<tr>
<td>15</td>
<td>15824</td>
<td>47854</td>
</tr>
<tr>
<td>16</td>
<td>14469</td>
<td>161915</td>
</tr>
<tr>
<td>17</td>
<td>1196</td>
<td>24962</td>
</tr>
<tr>
<td>18</td>
<td>13521</td>
<td>32377</td>
</tr>
<tr>
<td>19</td>
<td>13360</td>
<td>12377</td>
</tr>
<tr>
<td>20</td>
<td>19720</td>
<td>30940</td>
</tr>
<tr>
<td>21</td>
<td>2407</td>
<td>43587</td>
</tr>
<tr>
<td>22</td>
<td>41716</td>
<td>46276</td>
</tr>
<tr>
<td>23</td>
<td>6108</td>
<td>63126</td>
</tr>
<tr>
<td>24</td>
<td>2482</td>
<td>20372</td>
</tr>
<tr>
<td>25</td>
<td>15193</td>
<td>72928</td>
</tr>
<tr>
<td>26</td>
<td>27806</td>
<td>34050</td>
</tr>
<tr>
<td>27</td>
<td>40032</td>
<td>41873</td>
</tr>
<tr>
<td>28</td>
<td>11308</td>
<td>1541</td>
</tr>
</tbody>
</table>
Table 2.9: 95% Confidence Regions for $(\lambda, \theta)$ for Population A Data

**Reverted-Holt Model**

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\theta$</th>
<th>0.0</th>
<th>.50</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
<th>4.0</th>
<th>4.5</th>
<th>5.0</th>
<th>5.5</th>
<th>6.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>.50</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>.25</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

**Ricker Model**

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\theta$</th>
<th>-.25</th>
<th>0.0</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>1.0</th>
<th>1.25</th>
<th>1.50</th>
<th>1.75</th>
<th>2.0</th>
<th>2.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>.50</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>.25</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

*** MLE over the grid  
+ 95% confidence region
Table 2.10: Summary statistics for Population A Data

<table>
<thead>
<tr>
<th>Model</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Spearman's $\rho$ (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beverton-Holt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(\lambda, \theta) = (1.0)$ no transformation</td>
<td>-1.36</td>
<td>1.12</td>
<td>.44 (.02)</td>
</tr>
<tr>
<td>$(\lambda, \theta) = (-1.0)$ linearizing</td>
<td>2.90</td>
<td>9.69</td>
<td>.04 (.85)</td>
</tr>
<tr>
<td>$(\lambda, \theta) = (.25,3)$ MLE</td>
<td>.12</td>
<td>-.08</td>
<td>.18 (.36)</td>
</tr>
<tr>
<td>$(\lambda, \theta) = (.5,0)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$= (.25,0)$</td>
<td>-.68</td>
<td>-.11</td>
<td>.31 (.11)</td>
</tr>
<tr>
<td>$= (0,0)$</td>
<td>-.20</td>
<td>-.36</td>
<td>.20 (.31)</td>
</tr>
<tr>
<td></td>
<td>.38</td>
<td>-.01</td>
<td>.09 (.64)</td>
</tr>
<tr>
<td>Ricker</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(\lambda, \theta) = (1.0)$ no transformation</td>
<td>-1.25</td>
<td>1.05</td>
<td>.42 (.03)</td>
</tr>
<tr>
<td>$(\lambda, \theta) = (0,0)$ linearizing</td>
<td>.59</td>
<td>-.08</td>
<td>.05 (.80)</td>
</tr>
<tr>
<td>$(\lambda, \theta) = (.25,1)$ MLE</td>
<td>-.20</td>
<td>-.37</td>
<td>-.15 (.44)</td>
</tr>
</tbody>
</table>
Figure 2.10: Population A Data. Scatterplot of recruits by spawners. The solid line represents the mean regression line for the Beverton-Holt model. The dashed line represents the median regression model for the Beverton-Holt model.
CHAPTER III

ERRORS-IN-VARIABLES AND THE TRANSFORM-BOTH-SIDES MODEL

3.0 Introduction

We assume the transform-both-sides model

\[ y_1^{(\lambda)} = f^{(\lambda)}(x_1, \beta) + \sigma_1 \epsilon_1 \]  

but the observed data \((y_1, X_1)\) are measured with error such that \(X_1 = x_1 + \sigma_2 v_1\). If the observed variables \(X_1\)'s are substituted for the \(x_1\)'s into (3.1) and the measurement error is ignored, the estimated parameters are biased. By varying the rates of convergence we are able to gain insight into the nature of this bias through the investigation of asymptotic distributions. In section 3.1 we consider two types of asymptotics and compute the asymptotic distribution of the maximum likelihood estimator, \(\hat{\Theta}\), based on each. We study the bias terms and suggest corrected estimators for \(\hat{\Theta}\). The calculations here are deliberately left on a heuristic basis as we are trying to understand the form of the asymptotic bias. Recall that \(\Theta = (\beta, \sigma, \lambda)\).

3.1 Asymptotic Theory

We assume model (3.1), but the observed data \((y_1, X_1)\) are measured with error such that \(X_1 = x_1 + \sigma_2 v_1\). We assume further that the \(\epsilon_1\)'s and the \(v_1\)'s are independent, symmetric random variables with mean 0 and variance 1.
We consider two cases of the asymptotics used by Amemiya and Fuller (1985) in their study of the nonlinear functional relationship. In their work, to obtain consistency and asymptotic normality, the measurement error must tend to 0 as \( N \) increases, that is, \( N^{1/2} \sigma_2 \to \xi \), \( 0 \leq \xi < \infty \). How fast \( \sigma_2 \) approaches 0 relative to \( \sigma_1 \) is controlled by the choice of \( \zeta \) where \( \sigma_2 / (\sigma_1^{1+\zeta}) \to \delta \), for \( 0 \leq \delta < \infty \) and \( \zeta \geq 0 \), for some \( \zeta \). We compute the asymptotic distribution of the estimators based on two cases of these asymptotics where in case (A), \( \zeta = 1 \), and in case (B), \( \zeta = 0 \). Specifically we let

\[
(3.4) \quad \text{Case (A)} \quad N^{1/2} \sigma_2 \to \xi \quad 0 \leq \xi < \infty \\
\frac{\sigma_2}{(\sigma_1)^2} \to \delta \quad 0 \leq \delta < \infty 
\]

\[
(3.5) \quad \text{Case (B)} \quad N^{1/2} \sigma_2 \to \xi \quad 0 \leq \xi < \infty \\
\frac{\sigma_2}{\sigma_1} \to \gamma \quad 0 \leq \gamma < \infty 
\]

In each case \( N \to \infty \), \( \sigma_1 \to 0 \), and \( \sigma_2 \to 0 \). In case (A), the measurement error variance \( \sigma_2^2 \) is very small relative to the equation error variance \( \sigma_1^2 \), whereas in case (B) they are proportional. The case (B) rates of convergence lead to more interesting results than those of case (A) since under rates (B) the model is farther from the non-EIV model than in case (A).

In order to simplify the presentation, let subscripts \( A \) and \( B \) represent the asymptotics employed and subscripts \( \beta, \sigma, \) and \( \lambda \) refer to the designated parameter.
Result 3.1: Assume the rates of convergence given by (3.4). Let
\[ N \to \infty, \sigma_1 \to 0, \text{and } \sigma_2 \to 0 \text{ simultaneously, then} \]
\[ N^{1/2} D^{-1} (\hat{\theta} - \theta) \Rightarrow N(\Delta_A, \tilde{\Sigma}_A). \]

Under the rates given by (3.5),
\[ N^{1/2} D^{-1} (\hat{\theta} - \theta) \Rightarrow N(\Delta_B, \tilde{\Sigma}_B). \]

where \( D = (\sigma_1, \ldots, \sigma_1, 1) \), \( \tilde{\Sigma}_A = \Sigma \) (the covariance matrix when \( \sigma_2 = 0 \)), and \( \Delta_A = (0 \ A_{\alpha \sigma} A_{\alpha \lambda})^T \), \( \Delta_B = (A_{\beta \beta} A_{\beta \sigma} A_{\beta \lambda})^T \), and \( \tilde{\Sigma}_B \) are as defined in Section 3.2.

It is not surprising that there is an asymptotic bias since the estimating equations are biased. See Section 3.2. Under rate (A) convergence the only effect of measurement error is a bias in the estimates, \( \hat{\lambda} \) and \( \hat{\sigma} \). It is interesting that the bias in \( \hat{\lambda} \) does not correspond to a bias in the estimated regression parameters \( \hat{\beta} \). Under the weaker rates (B), where \( \sigma_2 \) is proportional to \( \sigma_1 \), not only is there bias for each of the parameters but the bias in \( (\hat{\lambda}, \hat{\sigma}) \) is not convergent. In addition, the asymptotic variance increases for each of the parameters, that is, \( \tilde{\Sigma}_{B_{11}} > \tilde{\Sigma}_{A_{11}} \).

In order to obtain asymptotically unbiased estimates we subtract from the original estimating equations the estimated bias, where \( \ell(\theta) \) is defined on page 70. By asymptotically unbiased, we mean that
\[ N^{1/2} (\hat{\theta} - \theta) \Rightarrow N(0, \tilde{\Sigma}). \]

i.e., the limit distribution is centered at \( \theta \).

Although the resulting estimating equations produce asymptotically unbiased estimators of the parameters, there is an effect on the
asymptotic covariance matrix since \( \hat{E}[\ell(\theta)] \) will have random elements due to the measurement error \( \sigma_{2 \nu_1} \).

**Result 3.2:** The bias-corrected estimating equations given by

\[
\ell^*(\theta) = \ell(\theta) - \hat{E}[\ell(\theta)]
\]
yield under rates (A),

\[
N^{1/2} D^{-1}(\hat{\theta} - \theta) \Rightarrow N(\theta, \Sigma_A^*)
\]

and under rates (B),

\[
N^{1/2} D^{-1}(\hat{\theta} - \theta) \Rightarrow N(\theta, \Sigma^*)
\]

where \( \hat{E}[\ell(\theta)] \) is \( E[\ell(\theta)] \) evaluated at \( X_1 = x_1 + \sigma_{2 \nu_1} \), and

\[
\Sigma^* = \begin{bmatrix}
\Sigma_\beta^* & 0 & 0 \\
0 & \frac{\Sigma_\beta}{\sigma_\lambda} & \Sigma_\beta^*
\end{bmatrix} = \begin{bmatrix}
\Sigma_\beta & 0 & 0 \\
0 & \Sigma_\beta\Sigma_\beta & \Sigma_\beta
\end{bmatrix}
\]

Thus there is no effect due to measurement error under case (A) when the bias-adjusted estimating equations are used. Under case (B) we can obtain unbiased estimates for \( \hat{\beta} \) but there is a cost due to measurement error in that \( \Sigma_\beta^* = \frac{\Sigma_\beta}{\sigma_\lambda} > \Sigma_\beta \). Since there is no change in the asymptotic covariance matrix, \( \Sigma_\beta = \Sigma_\beta^* \), there is no additional cost due to the bias correction. In obtaining unbiased estimates for \( \lambda \) and \( \sigma \) there is an additional effect due to the correction imposed by the measurement error contained in the correction terms, i.e.,

\[
\Sigma_\beta^* = \frac{\Sigma_\beta}{\sigma_\lambda}
\]

### 3.2 Detailed Calculations

In this section we sketch results 3.1 and 3.2. To simplify the
proofs we adopt additional notation. Let \( h(y_1, \lambda) = y_1^{(\lambda)} \), \( g(x_1, \beta, \lambda) = f^{(\lambda)}(x_1, \beta) \) and \( D = (\sigma_1, \ldots, \sigma_1, 1) \). Subscripts on \( g \) and \( f \) indicate the derivative taken with respect to that parameter. Superscript \( o \) indicates that \( \sigma_2 \) has been set to \( 0 \) after the last \( \sigma_2 \) differentiation. Once \( X_1 \) has been replaced by \( x_1 + \sigma_2 \nu_1 \) and the subsequent expansion about \( \sigma_2 = 0 \) occurs, the arguments are suppressed, i.e., \( g_{\beta}(x_1, \beta, \lambda) = g_{\beta} \). If there is any ambiguity the arguments will be retained.

Calculations for Result 3.1:

Let \( \ell(\theta) = N^{-1} \) (derivative of the loglikelihood with respect to \( \theta \) evaluated at the observed data). We assume that

\[
N^{1/2}(\beta - \beta)/\sigma_1 = o_p(1)
\]

\[
N^{1/2}(\sigma_1 - \sigma_1)/\sigma_1 = o_p(1)
\]

\[
N^{1/2}(\lambda - \lambda) = o_p(1)
\]

Let \(-E[\ell_\lambda(\theta)]\) be the information in the correctly measured data. Thus

\[
0 = N^{1/2} \ell(\theta)
\]

\[
= N^{1/2} \ell(\theta) + N^{1/2} \ell(\theta) \hat{\theta} - \theta
\]

\[
= N^{1/2} \ell(\theta) + N^{1/2} E[\ell(\theta)] \hat{\theta} - \theta
\]

\[
+ N^{1/2}(\ell(\theta) - E[\ell(\theta)]) \hat{\theta} - \theta.
\]

By the form of \( D \), if \( A = E[\ell(\theta)] \) then
\[ 0 = N^{1/2} D^{-1} \epsilon(\theta) + N^{1/2} E[\epsilon(\theta)] D^{-1} (\hat{\theta} - \theta) + N^{1/2} (\epsilon(\theta_\ast) - E[\epsilon_\ast(\theta)]) D^{-1} (\hat{\theta} - \theta) \]

\[ \approx N^{1/2} D^{-1} \epsilon(\theta) + N^{1/2} \Lambda D^{-1} (\hat{\theta} - \theta). \]

When set equal to zero, the estimating equations based on the normal likelihood for model (3.1) are:

\[ \epsilon(\beta) = \frac{1}{N \sigma_1} \sum_{i=1}^{N} \left( \frac{h(y_i, \lambda) - g(X_i, \beta, \lambda)}{\sigma_1} \right) g_\beta(X_i, \beta, \lambda) \]

\[ \epsilon(\sigma) = \frac{1}{N \sigma_1} \sum_{i=1}^{N} \left( \frac{h(y_i, \lambda) - g(X_i, \beta, \lambda)}{\sigma_1} \right)^2 - \frac{1}{\sigma_1} \]

\[ \epsilon(\lambda) = -\frac{1}{N \sigma_1} \sum_{i=1}^{N} \left( \frac{h(y_i, \lambda) - g(X_i, \beta, \lambda)}{\sigma_1} \right) \left( \frac{h_\lambda(y_i, \lambda) - g_\lambda(X_i, \beta, \lambda)}{\sigma_1} \right) \]

\[ + \frac{1}{N} \sum_{i=1}^{N} \log(y_i) \]

Replacing \( X_i \) with \( x_i + \sigma_2 \nu_i \) and expanding in a Taylor series about \( \sigma_2 = 0 \) gives

\[ \epsilon(\beta) = \frac{1}{N \sigma_1} \sum_{i=1}^{N} \left[ \epsilon_1 g_\beta - \left( \frac{\sigma_2}{\sigma_1} \right) v_1 g_\sigma_2 g_\beta + \sigma_2 v_1 \epsilon_1 v_1 g_\beta \sigma_2 \right. \]

\[ \left. - \left[ \frac{\sigma_2^2}{(2 \sigma_1)} \right] v_1^2 g_\sigma_2^2 g_\beta - \left( \frac{\sigma_2}{\sigma_1} \right) v_1^2 g_\sigma_2 g_\beta \sigma_2^2 \right. \]

\[ + \left( \frac{\sigma_2^2}{2} \right) \epsilon_1 v_1^2 g_\beta \sigma_2^2 + O(\sigma_2^3) \]
\[ \ell(\sigma) = \frac{1}{N\sigma} \sum_{i=1}^{N} \left[ \epsilon_{i}^2 - 2(\sigma_{2}/\sigma_{1}) \nu_{i} \epsilon_{i} \bar{g}_{\sigma_{2}^o} - (\sigma_{2}/\sigma_{1})^2 \nu_{i} \epsilon_{i} \bar{g}_{\sigma_{2}^o} \right. \\
+ \left. (\sigma_{2}/\sigma_{1})^2 \nu_{i} \bar{g}_{\sigma_{2}^o} \right] - \frac{1}{\sigma_{1}} \]

\[ \ell(\lambda) = -\frac{1}{N\sigma} \sum_{i=1}^{N} \left[ \epsilon_{i}(h_{\lambda} - g_{\lambda}) - (\sigma_{2}/\sigma_{1}) \nu_{i} (h_{\lambda} - g_{\lambda}) g_{\sigma_{2}^o} - \sigma_{2} \epsilon_{i} \nu_{i} g_{\lambda} \sigma_{2} \\
- [\sigma_{2}/(2\sigma_{1})] \nu_{i} (h_{\lambda} - g_{\lambda}) g_{\sigma_{2}^o} \sigma_{2} + (\sigma_{2}/\sigma_{1}) \nu_{i} g_{\sigma_{2}^o} g_{\lambda} \sigma_{2} \\
- (\sigma_{2}/2) \epsilon_{i} \nu_{i} g_{\lambda} \sigma_{2} \sigma_{2} + 0(\sigma_{2}^3) \right] + \frac{1}{N} \sum_{i=1}^{N} \log (y_{i}). \]

A Taylor expansion about \( \sigma_{1} = 0 \) and taking expectations yields

\[ (3.6) \]
\[ E[\ell(\beta)] \approx -\sigma_{2}^2/(N\sigma_{1}^2) \left[ \sum_{i=1}^{\Sigma} g_{\sigma_{2}^o} g_{\sigma_{2}^o} \bar{g}_{\beta} - \sum_{i=1}^{\Sigma} g_{\sigma_{2}^o} g_{\sigma_{2}^o} \bar{g}_{\sigma_{2}^o} \right] \]

\[ E[\ell(\sigma_{1})] \approx \sigma_{2}^2/(N\sigma_{1}^3) \left[ \sum_{i=1}^{\Sigma} g_{\sigma_{2}^o} g_{\sigma_{2}^o} \right] \]

\[ E[\ell(\lambda)] \approx -\sigma_{2}^2/(N\sigma_{1}^2) \left[ \sum_{i=1}^{\Sigma} g_{\sigma_{2}^o} g_{\sigma_{2}^o} \bar{g}_{\lambda} \sigma_{2} \right] \]

We compute \( A = E[\partial/\partial \theta \ell(\theta)] \) and pre-multiply by \( D \) to obtain

\[ DA = \begin{bmatrix} -S_{11}/\sigma_{1} + (\sigma_{2}/\sigma_{1}) S_{0} & -(\sigma_{2}/\sigma_{1}) S_{12} & (\sigma_{2}/\sigma_{1}) S_{13} \\
(\sigma_{2}/\sigma_{1}) S_{12} & -2/\sigma_{1} - 3(\sigma_{2}/\sigma_{1})^3 S_{22} & 2\bar{v} + (\sigma_{2}/\sigma_{1}) S_{23} \\
(\sigma_{2}/\sigma_{1}) S_{13} & (2/\sigma_{1})\bar{v} + (\sigma_{2}/\sigma_{1}) S_{23}^T & -2\bar{w} - (\sigma_{2}/\sigma_{1}) S_{33} \end{bmatrix} \]
where \( \bar{v} = N^{-1} \sum \log f, \bar{w} = N^{-1} \sum (\log f)^2 \) and all sums \( S_{ij} \) are from 1 to \( N \) and are defined in Section 3.3.

Up to this point we have not yet had to consider the differing rates of convergence. The asymptotic bias, given by \( N^{1/2}(DA)^{-1}E[\ell(\theta)] \), under case (A) tends to \( \Delta_A = (0 \Delta_{\alpha} \Delta_{\lambda})^T \) where

\[
\Delta_{\alpha} = -\xi \delta / 2 (\bar{w} S_{22} - \bar{v} S_{23}) / (\bar{w} - \bar{v})^2 \\
\Delta_{\lambda} = -\xi \delta / 2 (\bar{v} S_{22} - S_{23}) / (\bar{w} - \bar{v})^2.
\]

In case (B), \( \Delta_B = (\Delta_{\beta \beta} \Delta_{\beta \sigma} \Delta_{\beta \lambda})^T \) where

\[
\Delta_{\beta \beta} = -\xi \gamma^2 / 4 S_{11}^{-1} S_{12} + \xi \gamma^4 / D^\ast \ S_{11}^{-1} \left( (\bar{w} + \gamma^2 S_{33}) S_{12} S_{22} + (\bar{v} + \gamma^2 S_{23}) 2 S_{13} S_{23} \right) - (2 + 3 \gamma^2 S_{22}) S_{13} S_{23} \right]
\]

\[
\Delta_{\beta \sigma} = -2N^{1/2} / D^\ast \left( (\bar{w} + \gamma^2 S_{33}) S_{22} + (\bar{v} + \gamma^2 S_{23}) \right)
\]

\[
\Delta_{\beta \lambda} = -N^{1/2} / D^\ast \left( 2 \gamma^2 (\bar{v} + \gamma^2 S_{23}) S_{22} + (2 + 3 \gamma^2 S_{22}) \right)
\]

and \( D^\ast = 2 (\bar{w} + \gamma^2 S_{33})(2 + 3 \gamma^2 S_{22}) - 4 (\bar{v} + \gamma^2 S_{23})^2 \). We also compute

\[
E[\ell(\theta)] \quad \text{where}
\]

\[
\text{DE}[\ell(\theta)] D = \begin{bmatrix}
-S_{11} + \sigma_2^2 S_0 & (\sigma_2^2 / \sigma_1) S_{12} & -(\sigma_2^2 / \sigma_1) S_{13} \\
(\sigma_2^2 / \sigma_1) S_{12}^T & -2 - 3(\sigma_2^2 / \sigma_1^2) S_{22} & 2\bar{v} + 2(\sigma_2^2 / \sigma_1^2) S_{23} \\
-(\sigma_2^2 / \sigma_1) S_{13}^T & 2\bar{v} + 2(\sigma_2^2 / \sigma_1^2) S_{23}^T & -2\bar{w} - (\sigma_2^2 / \sigma_1^2) S_{33}
\end{bmatrix}
\]
Under case (A),

\[
\begin{bmatrix}
-S_{11} & 0 & 0 \\
0 & -2 & 2v \\
0 & 2v & -2w
\end{bmatrix}
\]

\[\text{DE}[\hat{\varepsilon}(\theta)]D \rightarrow \]

and under Case (B),

\[
\begin{bmatrix}
-S_{11} & 0 & 0 \\
0 & -2-3v^2s_{22} & 2v + 2v^2s_{23} \\
0 & 2v + 2v^2s_{32} & -2w - v^2s_{33}
\end{bmatrix}
\]

\[
\Sigma \text{ is the covariance matrix for } \varepsilon(\theta) \text{ and we have under case (A)}
\]

\[
\begin{bmatrix}
S_{11} & 0 & 0 \\
0 & Ee_1^4 - 1 & -(Ee_1^4 - 1) v \\
0 & -(Ee_1^4 - 1) v & (Ee_1^4 - 1) w
\end{bmatrix}
\]

and under case (B), \(\text{ND}_B \Sigma = \text{ND}_A \Sigma + \Sigma_A\) where

\[
\Sigma_A = \\
\left[\begin{array}{ccc}
\gamma^2s_{10} & 0 & 0 \\
\gamma^2s_{22} & \gamma^4(Ee_1^4 - 1)s_{20} & [-2\gamma^2(s_{20} + s_{23}) - (Ee_1^4 - 1)\gamma^4s_{30}] \\
\gamma^4(Ee_1^4 - 1)s_{50} + \gamma^2s_{60}
\end{array}\right]
\]
Therefore we can compute the asymptotic covariance matrix for case (A), \( \tilde{\Sigma}_A = N(DAD)^{-1} \Sigma_A D(DAD)^{-1} \), which can be written as

\[
\tilde{\Sigma}_A = \begin{bmatrix}
S_{11}^{-1} & 0 & 0 \\
0 & \frac{(E\varepsilon_1^4 - 1)\bar{w}}{4(\bar{w} - \bar{v}^2)} & -\frac{(E\varepsilon_1^4 - 1)\bar{v}}{4(\bar{w} - \bar{v}^2)} \\
0 & -\frac{(E\varepsilon_1^4 - 1)\bar{v}}{4(\bar{w} - \bar{v}^2)} & \frac{E\varepsilon_1^4}{4(\bar{w} - \bar{v}^2)}
\end{bmatrix}
\]

Under the asymptotics for case (B), the terms of the asymptotic covariance matrix, \( \tilde{\Sigma}_B = N(DAD)^{-1}(\Sigma_B D)(DAD)^{-1} \), are unwieldy. The variance term for \( \hat{\beta} \) is given by

\[
\tilde{\Sigma}_{B\beta} = S_{11}^{-1} + \gamma^2 S_{11}^{-1} \left[ \sum_i \left( g_{\sigma_2}^o \right)^2 g_{\sigma_2} g_{\beta}^T \right] S_{11}^{-1}
\]

**Calculations for Result 3.2:**

The calculations parallel those of Result 3.1. Let \( \ell_*(\theta) = \ell(\theta) - \hat{E}[\ell(\theta)] \), where \( \hat{E}[\ell(\theta)] \) is \( E[\ell(\theta)] \) evaluated at \( X_1 = x_1 + \sigma_2 v_1 \). Expanding in a Taylor series about \( \sigma_2 = 0 \) yields

\[
\ell_*(\beta) = \ell(\beta) + \left[ \frac{\sigma_2^2}{(N\sigma_1^2)} \right] \sum_1 \left[ g_{\sigma_2} g_{\beta} - g_{\sigma_2} g_{\beta}^o \right] \\
+ \left[ \frac{\sigma_2^3}{(N\sigma_1^2)} \right] \sum_1 v_1 \left[ g_{\sigma_2} g_{\beta} - g_{\sigma_2} g_{\beta}^o \right] + O(\sigma_2^4)
\]
\[ \ell_\lambda(\sigma_1) = \ell(\sigma_1) - \left[ \sigma_2^2 / (N \sigma_1^3) \right] \sum_1 \left[ g_{\sigma_2 \sigma_2} g_{\sigma_2} \right] \\
- 2 \left[ \sigma_2^3 / (N \sigma_1^4) \right] \sum_1 \nu_1 \left[ g_{\sigma_2 \sigma_2} g_{\sigma_2} \right] + O(\sigma_2^4) \]

\[ \ell_\lambda(\lambda) = \ell(\lambda) + \left[ \sigma_2^2 / (N \sigma_1^2) \right] \sum_1 \left[ g_{\sigma_2 \sigma_2} g_{\lambda \sigma_2} \right] \\
+ \left[ \sigma_2^3 / (N \sigma_1^2) \right] \sum_1 \nu_1 \left[ g_{\sigma_2 \sigma_2} g_{\lambda \sigma_2} \right] \\
+ \left[ (\sigma_2^3 / (N \sigma_1^2) \right] \sum_1 \nu_1 \left[ g_{\sigma_2 \sigma_2} g_{\lambda \sigma_2} \right] + O(\sigma_2^4). \]

\[ \text{Var}[\ell_\lambda(\lambda)] \equiv \text{Var}[\ell(\lambda)] \text{ but } E[\ell_\lambda(\lambda)] \neq E[\ell(\lambda)] \text{ and will affect the asymptotic covariance of the estimated parameters. We have under either (A) or (B) that} \]

\[ \begin{bmatrix}
-S_{11} & 0 & 0 \\
0 & -2 & 2\nu \\
0 & 2\nu & -2w
\end{bmatrix} \]

Thus \( \Sigma_\beta \) is not affected but \( \Sigma_{\lambda \lambda} \) is different from \( \tilde{\Sigma}_b \).

3.3 Appendix

The following sums are used throughout Chapter 3. All sums are from 1 to \( N \).

\[ S_{11} = N^{-1} \sum_1 g_{\beta}^T g_{\beta} \]

\[ S_{12} = -2N^{-1} \sum_1 \left[ g_{\lambda \sigma_2} g_{\beta} + 2g_{\sigma_2 \sigma_2} g_{\beta} \right] \]

\[ S_{13} = -N^{-1/2} \sum_1 \left[ g_{\lambda \sigma_2} g_{\beta} + g_{\sigma_2 \sigma_2} g_{\beta \lambda} + g_{\sigma_2 \lambda \sigma_2} + g_{\sigma_2} g_{\beta \sigma_2} \right] \]
\[ \begin{align*}
S_{22} &= N^{-1} \sum_{i} g_{\sigma_2}^\circ g_{\sigma_2}^\circ \\
S_{23} &= N^{-1} \sum_{i} g_{\sigma_2}^\circ g_{\sigma_2}^{\alpha} \\
S_{33} &= N^{-1} \sum_{i} (g_{\sigma_2}^\circ)^2 \\
S_{10} &= N^{-1} \sum_{i} (g_{\sigma_2}^\circ)^2 g_{\beta}^T g_{\beta} \\
S_{30} &= N^{-1} \sum_{i} (g_{\sigma_2}^\circ)^3 g_{\lambda \sigma_2} \\
S_{50} &= N^{-1} \sum_{i} (g_{\sigma_2}^\circ g_{\lambda \sigma_2})^2 \\
S_0 &= N^{-1} \sum_{i} [2g_{\sigma_2}^\circ g_{\beta}^T g_{\beta} + g_{\sigma_2}^\circ g_{\beta} g_{\beta} + g_{\sigma_2}^\circ g_{\lambda \sigma_2}^T + 2g_{\sigma_2}^\circ g_{\beta} g_{\lambda \sigma_2}^T] \\
S_{21} &= S_{12} \\
S_{31} &= S_{13} \\
S_{32} &= S_{23} \\
S_{20} &= N^{-1} \sum_{i} (g_{\sigma_2}^\circ)^4 \\
S_{40} &= N^{-1} \sum_{i} (\log f)(g_{\sigma_2}^\circ)^2 \\
S_{60} &= N^{-1} \sum_{i} (g_{\sigma_2}^\circ \log f + g_{\lambda \sigma_2}^\circ) \\
\end{align*} \]
CHAPTER 4

THE ACE ALGORITHM AND TRANSFORM-BOTH-SIDES REGRESSION

4.0 Introduction

The ACE algorithm of Breiman and Friedman (1985a, b) is designed to find transformations for each of the regression variables such that the best-fitting additive model is constructed. By "best-fitting" they mean the model where the fraction of variance not explained by the regression of the transformed response onto the transformed predictors is minimized.

The ACE procedure is conceptually simple. Although ACE is sufficiently general for vector-valued responses and predictors, we consider only the bivariate case with one response and one predictor. Let y and x be random variables such that y is the response and x is the predictor, where \( \theta \) and \( \phi \) are respectively the transformation functions of y and x. The ACE algorithm minimizes with respect to \( \theta \) and \( \phi \)

\[
e^2(\theta, \phi) = \frac{E[(\theta(y) - \phi(x))^2]}{\operatorname{Var}[\theta(y)]}.
\]

The basis for the ACE algorithm is

\[
\theta_1(y) = E[\phi(x)|y] / \|E[\phi(x)|y]\|.
\]

which minimizes (4.1) with respect to \( \theta(y) \) for a given function of
\( \phi(x) \). Minimizing (4.1) with respect to \( \phi(x) \) for a given \( \theta(y) \) yields

(4.3) \[ \phi_1(x) = \mathbb{E}[\theta(y) | x]. \]

See Section 4.4 for a derivation. The ACE algorithm is an iterative procedure based on computing and updating (4.2) and (4.3) in order to minimize (4.1). In the bivariate case, the effect of ACE is to estimate the maximal correlation between the two variables, thus minimizing \( e^2 \) is equivalent to maximizing \( R^2 = 1 - e^2 \), the usual correlation coefficient.

The ACE procedure is nonparametric in the sense that the only distributional assumption made is that the data \((y_i, x_i)\) are independent samples drawn from the distribution of \((y, x)\). No assumptions are made about the form of the transformation, i.e., it need not be from a particular parametric family and need not be monotone. However, the algorithm is flexible enough to constrain the transformations to be monotone if so desired.

Our goal in this chapter is to investigate the performance of ACE in the transform-both-sides setting. When little is known about the relationship between the response and the predictors, ACE may be useful in identifying some otherwise unknown structure in the data. The transform-both-sides problem is slightly different. We have an assumed regression model which we want to preserve, but the way in which error enters the system is unknown. We consider transforming both the response and the regression function simultaneously by the same transformation in order to maintain this relationship yet still address problems in the error distribution, i.e., heteroscedasticity and lack of symmetry. Breiman and Friedman (1985) claim that finding
the best-fitting additive model is a more comprehensive goal than the common goals of variance stabilization and symmetrization of errors. Thus it seems the ACE procedure is a likely candidate for estimating the transformation in the transform-both-sides setting.

In the following we address several issues concerning the performance of ACE in the transform-both-sides model. Under weak conditions on the joint distribution of \( y, x_1, \ldots, x_p \), Breiman and Friedman (1985) show that optimal transformations exist and are generally unique up to a change of sign. There may exist suboptimal transformations not identified by ACE, that is, eigenfunction solutions that are approximately equal to the optimal solutions, that will provide different information about the relationship between the response and the predictors. The first question is whether or not maximizing \( R^2 \) is a reliable criterion for model selection. Maximizing \( R^2 \) with respect to the regression parameters in a linear model leads to the least squares estimators. Thus in a linear model under normality we get consistent estimates using \( R^2 \) as our selection criterion. This criterion, under the nonlinear transform-both-sides model, does not clearly lead to consistent estimates. We investigate this in Section 4.1. In transform-both-sides we assume that we have a regression model based on some prior understanding about the underlying physical process. In Section 4.2 we look at ACE as a goodness-of-fit tool for the assumed model, \( y = f(x, \beta) \). We investigate the ability of ACE to select the more appropriate model in two examples where there are two or more competing models. Breiman and Friedman (1985) state that ACE does not perform well in the presence of outliers, where by outliers they mean points that are located at extreme points. These points are
not necessarily bad points but have high leverage by virtue of their location apart from the bulk of the data. In section 4.3 we investigate how ACE responds to influential points that are not outliers in the Breiman-Friedman sense and determine if ACE can be used as a diagnostic tool to identify such points.

The numerical results were obtained using an ACE procedure programmed by R. N. Rodriguez that implements the ACE algorithm given in Breiman and Friedman (1982) and the data smoother found in Friedman and Stuetzle (1982). The smoother used by this program is not the same smoother used by Breiman and Friedman (1985).

4.1 The ACE Criterion in the Transform-Both-Sides Setting

4.1.1 Theoretical Analysis

Since ACE is a nonparametric procedure for estimating transformations its performance will be no more efficient than a parametric procedure where the distribution and model assumptions are correct. For the sake of comparison we formulate the ACE criterion as a parametric estimation problem where the errors are assumed to be independent and identically distributed normal random variables and the assumed model is correct. We consider the transformations $\theta$ and $\phi$ to be the same and to be members of the Box-Cox modified power transformation family indexed by the parameter $\lambda$. We define $\hat{\lambda}_{R_2}$ to be the value of the parameter $\lambda$ that minimizes

\begin{equation}
(4.4) \quad e^2(\lambda) = 1 - R^2(\lambda) = \frac{1/N \sum \limits_i [y_i^{(\lambda)} - f(x_i, \beta)^{(\lambda)}]^2}{1/N \sum \limits_i [y_i^{(\lambda)} - \bar{y}^{(\lambda)}]^2}
\end{equation}
In addition to \( \hat{\lambda}_{R^2} \) we consider the maximum likelihood estimator (\( \hat{\lambda}_{MLE} \)) and an estimator suggested by Breiman and Friedman (1985), (\( \hat{\lambda}_{BF} \)). The estimate \( \hat{\lambda}_{BF} \) is that value of \( \lambda \) for which \( R^2(\lambda) \) is maximized when ACE is applied to \( (y^\lambda, f^\lambda(x, \beta)) \). Breiman and Friedman suggest this in an example where there is a strong linear relationship between the response and the predictors. In that example the ACE procedure does not indicate a transformation but there is reason to believe that a transformation is called for.

We know that at \( \lambda = 0 \), the estimating equations for the maximum likelihood estimators are unbiased in a transform-both-sides model with known \( \beta \), i.e.,

\[
E \left[ \frac{\partial}{\partial \lambda} \log \text{likelihood} \right]_{\lambda=0} = 0. \tag{See below.}
\]

Therefore, by M-estimation theory, for any \( \sigma \) we can expect transform-both-sides to be consistent if \( \lambda = 0 \). Our goal here is different in that we want to study ACE. We want to show in the case where \( \lambda = 0 \), that ACE does not consistently estimate \( \lambda \). Since \( \beta \) is known, we suppress the arguments and write \( f_1 = f(x_1, \beta) \). The loglikelihood for the transform-both-sides model is

\[
\ell(\lambda, \sigma) = -1/2 \log \sigma^2 + (\lambda-1) \log(y_1) - 1/(2\sigma^2)[y(\lambda) - f(\lambda)]^2.
\]

We have

\[
\frac{\partial}{\partial \sigma^2} \ell(\lambda, \sigma) \bigg|_{\lambda=0} = \frac{(y(\lambda) - f(\lambda))^2}{2\sigma^4} - 1/(2\sigma^2)
\]

\[
= \frac{(\log y - \log f)^2 - \sigma^2}{2\sigma^4}
\]

\[
= \frac{\sigma^2(\varepsilon^2 - 1)}{2\sigma^4}.
\]
Therefore
\[ E \left( \frac{\partial}{\partial \sigma^2} \ell(\lambda, \sigma) \right|_{\lambda=0} \right) = 0. \]

Also
\[ \frac{\partial}{\partial \lambda} \ell(\lambda, \sigma) \bigg|_{\lambda=0} = \{ \log y - 1/(\lambda \sigma^2)[y^{(\lambda)} - f^{(\lambda)}][y^{\lambda} \log y - f^\lambda \log f] \]
\[ + 1/(\lambda \sigma^2) [y^{(\lambda)} - f^{(\lambda)}]^2 \bigg|_{\lambda=0} \}
\]

Repeated application of L'Hospital's rule yields
\[ \frac{\partial}{\partial \lambda} \ell(\lambda, \sigma) \bigg|_{\lambda=0} = \log y - 1/(2\sigma^2)[\log y - \log f] [(\log y)^2 - (\log f)^2] \]
\[ = \log f + \sigma \epsilon - \epsilon/(2\sigma) [(\log f + \sigma \epsilon)^2 - (\log f)^2] \]
\[ = \log f + \sigma \epsilon - \epsilon^2 \log f - \sigma \epsilon^3/2. \]

This also has mean 0, which shows that the estimating equation for the maximum likelihood estimator is unbiased.

Assume that \((f_i, \epsilon_i)\) are independent and identically distributed, \(f_i\) is independent of \(\epsilon_i\), and \(\epsilon_i \sim N(0,1)\). Let \(\log y_i = \log f_i + \sigma \epsilon_i\). Assume also that for some \(M > 0\), \(0 < M^{-1} < f_i < M < \infty\).

**Proposition 1:** For any integer \(k > 0\) and any \(a > 0\),

1. \(E[\log y_i]^k[y_i^a + y_i^{-a}] < b(a,k) < \infty.\)
   
   where \(b(a,k)\) is independent of \(i\).

2. \(N^{-1} \sum_{i=1}^{N} \log y_i]^k[y_i^a + y_i^{-a}] = O_p(1).\)
Proof:

(Part 1):

\[ \log y_i = \log f_1 + \sigma \epsilon_i \]
\[ y_i = f_1 e^{\sigma \epsilon_i} \quad M^{-1} < f_1 < M. \]

Let \( M_2 = \log M \). Therefore

\[
[\log y_i]^k[y_i^a + y_i^{-a}] = [\log f_1 + \sigma \epsilon_i]^k[(f_1 e^{\sigma \epsilon_i})^a + (f_1 e^{\sigma \epsilon_i})^{-a}]
\]
\[
\leq [M_2 + \sigma |\epsilon_i|]^k (M^a + M^{-a}) [e^{a \sigma \epsilon_i} + e^{-a \sigma \epsilon_i}].
\]

Therefore for constants \( d_j \),

\[
\leq \sum_{j=0}^{k} d_j |\epsilon_i|^k (M^a + M^{-a}) [e^{a \sigma \epsilon_i} + e^{-a \sigma \epsilon_i}].
\]

It suffices that for any \( k, a, \epsilon_i \), \( E|\epsilon_i|^k e^{a \epsilon_i} < \infty \), which holds for the normal distribution.

(Part 2): By the Markov inequality, \( P[|T| > a] \leq E|X|/a \). Thus

\( |T| \leq O_p(1) \) if \( E|X| < \infty \). Write \( T = \sum_{i=1}^{N} (\log y_i)^k (y_i^a + y_i^{-a}) \) and apply Part (1).

Proposition 2: If \(|\lambda| \leq a\), then

\[ |y^{(\lambda)} - \log y| \leq |\lambda| [\log y]^2 [y^a + y^{-a}]. \]
Proof: \[ y^{(\lambda)} - \log y = (y^{\lambda} - 1 - \lambda \log y) / \lambda \]
\[ = (e^{\lambda \log y} - 1 - \lambda \log y) / \lambda. \]

Let \( x = \lambda \log y \). Then by a Taylor series expansion about \( x = 0 \) where there exists an \( x_\ast \) between 0 and \( x \) such that \( e^x - 1 - x = 1/2 \, x^2 \, e^x \), we can write
\[
|y^{(\lambda)} - \log y| = |(e^{\lambda \log y} - 1 - \lambda \log y) / \lambda|
\]
\[
= \left| \frac{1/2 \, \lambda^2 (\log y)^2 \, e^{\lambda \log y}}{\lambda} \right|
\]
\[
= |\lambda| / 2 \, (\log y)^2 \, |e^{\lambda \log y}|
\]
\[
= |\lambda| / 2 \, (\log y)^2 \, y^{\lambda_\ast}.
\]

Since \( \lambda_\ast \) is between \( \lambda \) and 0 and \( |\lambda| \leq a, \ |\lambda_\ast| \leq a \), therefore \( y^{\lambda_\ast} < y^a + y^{-a} \). To see this, note that we have the four cases:

- \( \lambda_\ast > 0, \ y > 1 \)
- \( \lambda_\ast > 0, \ y < 1 \)
- \( \lambda_\ast < 0, \ y > 1 \)
- \( \lambda_\ast < 0, \ y < 1 \)

Therefore
\[
|y^{(\lambda)} - \log y| \leq |\lambda| \, (\log y)^2 \, [y^a + y^{-a}].
\]
Proposition 3: If \(|\lambda| \leq a\), then

\[
\frac{\partial}{\partial \lambda} y^{(\lambda)} - 1/2 \left[ \log y \right]^2 \leq 2 |\lambda| |\log y|^3 (y^a + y^{-a}).
\]

Proof:

\[
\frac{\partial}{\partial \lambda} y^{(\lambda)} = -(y^{\lambda} - 1)/\lambda^2 + 1/\lambda y^{\lambda} \log y
\]

\[
= \frac{\lambda \log y - y^{\lambda} + 1}{\lambda^2}
\]

\[
= 1/\lambda \left[ (1 - y^{\lambda})/\lambda \right] + 1/\lambda y^{\lambda} (\log y)
\]

\[
= -y^{(\lambda)}/\lambda + [(y^{\lambda} - 1) \log y + \log y]/\lambda
\]

\[
= y^{(\lambda)} \log y + 1/\lambda \log y - 1/\lambda y^{(\lambda)}
\]

\[
= y^{(\lambda)} \log y - 1/\lambda [y^{(\lambda)} - \log y].
\]

From the proof of Proposition 2

\[
y^{(\lambda)} = \log y + 1/(2\lambda) \left[ \lambda \log y \right]^2 e^{\lambda \log y}.
\]

Therefore

\[
1/\lambda \left( y^{(\lambda)} - \log y \right) = 1/2 \left( \log y \right)^2 e^{\lambda \log y}
\]

\[
= 1/2 \left( \log y \right)^2 + 1/2 \left( \log y \right)^2 [e^{\lambda \log y} - 1].
\]

Thus

\[
\frac{\partial}{\partial \lambda} y^{(\lambda)} = y^{(\lambda)} \log y - 1/\lambda [y^{(\lambda)} - \log y]
\]

\[
= y^{(\lambda)} \log y - 1/2 \left( \log y \right)^2 - 1/2 \left( \log y \right)^2 [e^{\lambda \log y} - 1]
\]

\[
= \left( \log y \right) \left[ (\log y) + 1/(2\lambda) \left( \lambda \log y \right)^2 e^{\lambda \log y} \right]
\]

\[
- 1/2 \left( \log y \right)^2 - 1/2 \left( \log y \right)^2 [e^{\lambda \log y} - 1]
\]

\[
= \left( \log y \right)^2 + \lambda/2 \left( \log y \right)^3 e^{\lambda \log y} - 1/2 \left( \log y \right)^2
\]

\[
- 1/2 \left( \log y \right)^2 [e^{\lambda \log y} - 1]
\]
\[ = \frac{1}{2} (\log y)^2 + \lambda/2 (\log y)^3 e^{\lambda \log y} - 1/2 (\log y)^2 [e^{\lambda \log y} - 1]. \]

Now do a Taylor series expansion about \( \lambda = 0 \).

\[
\frac{\lambda \log y}{e^\lambda - 1} = \frac{\lambda}{y} \log y
\]

for some \( \lambda^* \) between 0 and \( \lambda \). Therefore

\[
|\frac{\partial}{\partial \lambda} y^{(\lambda)} - 1/2 [\log y]^2| = |(\log y)^2/2 + \lambda/2 (\log y)^3 e^{\lambda \log y} - (\log y)^2/2 [e^{\lambda \log y} - 1]| - (\log y)^2/2
\]

\[
= |\frac{\lambda}{2} (\log y)^3 e^{\lambda \log y} - (\log y)^2/2 [e^{\lambda \log y} - 1]| - (\log y)^2/2
\]

\[
\leq |\lambda|/2 |\log y|^3 \frac{\lambda^*}{\lambda} (\log y)^2/2 |\log y| - 1| \leq |\lambda|/2 |\log y|^3 \frac{\lambda^*}{\lambda} (\log y)^2/2 |\lambda^* (\log y)^{\lambda^*}|
\]

\[
\leq |\lambda|/2 |\log y|^3 \frac{\lambda^*}{\lambda} |\lambda^*|/2 |\log y|^3 \frac{\lambda^*}{\lambda}
\]

\[
\leq |\lambda| |\log y|^3 \left[ y^a + y^{-a} \right]
\]

\[
\leq 2 |\lambda| |\log y|^3 \left[ y^a + y^{-a} \right].
\]

\[ \boxed{\]}

**Lemma 1:** Define \( H_{11}(\lambda) = \{y_1^{(\lambda)} - y^{(\lambda)}\}^2 \)

\( H_{12}(\lambda) = \{y_1^{(\lambda)} - f_1^{(\lambda)}\}^2 \)

\( H_{13}(\lambda) = \{y_1^{(\lambda)} - y^{(\lambda)}\} \frac{\partial}{\partial \lambda} \{y_1^{(\lambda)} - y^{(\lambda)}\} \)
\[ H_{ij}(\lambda) = \{ y_i^{(\lambda)} - f_i(\lambda) \} \frac{\partial}{\partial \lambda} \{ y_i^{(\lambda)} - f_i(\lambda) \}. \]

Thus for \( j = 1, 2, 3, 4 \) if \( \lambda \xrightarrow{p} \lambda_0 = 0 \),

\[ N^{-1} \sum_{i=1}^{N} \{ H_{ij}(\hat{\lambda}) - H_{ij}(0) \} \xrightarrow{p} 0. \]

**Proof:** We verify only for \( j = 1 \). The rest are similar [for \( j = 3, 4 \) use also Proposition 3].

\[ H_{i1}(\hat{\lambda}) = \{ y_i^{(\hat{\lambda})} - y(\hat{\lambda}) \}^2 \]

\[ = \{ \log y_i - \log y + y_i^{(\hat{\lambda})} - \log y_i^{(\hat{\lambda})} - (y^{(\hat{\lambda})} - \log y) \}^2. \]

From Proposition 2 \(| y_i^{(\hat{\lambda})} - \log y_i | \leq | \hat{\lambda} | [\log y_i]^2 [y_i^a + y_i^{-a}] \), and

\[ | y_i^{(\hat{\lambda})} - \log y_i | = N^{-1} \sum_{i=1}^{N} | y_i^{(\hat{\lambda})} - \log y_j | \]

\[ \leq N^{-1} \sum_{i=1}^{N} | y_i^{(\hat{\lambda})} - \log y_i | \]

\[ \leq N^{-1} \sum_{i=1}^{N} | \hat{\lambda} | [\log y_i]^2 [y_i^a + y_i^{-a}] \]

\[ = | \hat{\lambda} | [\log y]^2 [y^a + y^{-a}] \].

Consider \(| \hat{\lambda} | \leq a \). Therefore,

\[ | H_{i1}(\hat{\lambda}) - H_{i1}(0) | = | 2(\log y_i - \log y) [ y_i^{(\hat{\lambda})} - \log y_i^{(\hat{\lambda})} - (y^{(\hat{\lambda})} - \log y) ] | \]

\[ - [ y_i^{(\hat{\lambda})} - \log y_i - (y^{(\hat{\lambda})} - \log y) ]^2 | \]
\[
\leq 2 |\hat{\lambda}| \left| \left( \log y_i - \overline{\log y} \right) \right| \\
\times \left[ (\log y_i)^2(y_i^a + y_i^{-a}) + \overline{(\log y)^2(y^a + y^{-a})} \right] \\
+ \lambda^2 \left[ (\log y_i)^2(y_i^a + y_i^{-a}) + \overline{(\log y)^2(y^a + y^{-a})} \right]^2.
\]

\[
= |\hat{\lambda}| A_{11} + \lambda^2 A_{21}.
\]

By Proposition 1 and Cauchy-Schwartz,

\[
\overline{\log y} = O_p(1)
\]

and further application of Proposition 1 yields

\[
N^{-1} \sum_{i=1}^{N} A_{11} = O_p(1) = N^{-1} \sum_{i=1}^{N} A_{21}.
\]

Since \( \hat{\lambda} \xrightarrow{p} 0 \), we have \( N^{-1} \sum_{i=1}^{N} \{ H_{11}(\hat{\lambda}) - H_{11}(0) \} \xrightarrow{p} 0 \), and we are done.

\[\blacksquare\]

**Lemma 2:** For the purpose of bookkeeping we have

\[
H_{11}(0) = (\log y_i - \overline{\log y})^2
\]

\[
H_{12}(0) = (\log y_i - \log f_i)^2
\]

\[
H_{13}(0) = (\log y_i - \overline{\log y}) \left\{ (\log y_i^2 - (\overline{\log y})^2 \right\}
\]

\[
H_{14}(0) = (\log y_i - \log f_i) \left\{ (\log y_i^2 - (\log f_i)^2 \right\}
\]

**Proof:** Notation. \[\blacksquare\]
Lemma 3: a) \( N^{-1} \sum_{i=1}^{N} H_{1i}(0) \xrightarrow{p} \text{Var}[\log f] + \sigma^2 \)

b) \( N^{-1} \sum_{i=1}^{N} H_{12}(0) \xrightarrow{p} \sigma^2 \)

c) \( N^{-1} \sum_{i=1}^{N} H_{13}(0) \xrightarrow{p} \frac{1}{2}(E(\log f)^3 - [E(\log f)]^2E(\log f)^2) + \sigma^2E(\log f) \)

d) \( N^{-1} \sum_{i=1}^{N} H_{14}(0) \xrightarrow{p} \sigma^2 E(\log f) \)

Proof: It is given for \( H_{11}, H_{12}, \) and \( H_{14} \) by strong laws.

a) \( N^{-1} \sum_{i=1}^{N} H_{11}(0) \xrightarrow{p} \text{Var}[\log y] \)

\[= \text{Var}[\log f + \sigma \epsilon] \]

\[= \text{Var}[\log f] + \sigma^2 \]

b) \( N^{-1} \sum_{i=1}^{N} H_{12}(0) = N^{-1} \sum_{i=1}^{N} (\sigma \epsilon_i)^2 \xrightarrow{p} \sigma^2 \)

d) \( N^{-1} \sum_{i=1}^{N} H_{14}(0) = N^{-1/2} \sum_{i=1}^{N} (\sigma \epsilon_i)^2 \left[ (\log f + \sigma \epsilon_i)^2 - (\log f)^2 \right] \)

\[= N^{-1/2} \sum_{i=1}^{N} (\sigma \epsilon_i)^2 \left[ (\sigma \epsilon_i)^2 + 2\sigma \epsilon_i \log f \right] \]

\[= N^{-1/2} \sum_{i=1}^{N} \left[ (\sigma \epsilon_i)^3 + 2\sigma^2 \epsilon_i^2 \log f \right] \]

\( \xrightarrow{p} \sigma^2 E(\log f) \)
c) Since \( \log y = N^{-1} \sum_{i=1}^{N} (\log f + \sigma \epsilon_i) \xrightarrow{P} \text{E} [\log f] \) and
\[
\frac{1}{(\log y)^2} = N^{-1} \sum_{i=1}^{N} (\log f + \sigma \epsilon_i)^2 \xrightarrow{P} \text{E} [\log f]^2 + \sigma^2,
\]
we have from strong laws that
\[
2N^{-1} \sum_{i=1}^{N} H_{13}(0) = N^{-1} \sum_{i=1}^{N} (\log y_i)^3 - \overline{(\log y)(\log y)^2}
\]
\[
= N^{-1} \sum_{i=1}^{N} (\log f_i + \sigma \epsilon_i)^3 - \overline{(\log y)(\log y)^2}
\]
\[
\xrightarrow{P} \text{E} [\log f]^3 + 3\sigma^2 \text{E} [\log f] - \text{E} [\log f][\sigma^2 + \text{E} (\log f)^2]
\]
\[
= \text{E} [\log f]^3 + 2\sigma^2 \text{E} [\log f] - \text{E} [\log f][\text{E} [\log f]^2].
\]

Theorem 4.1: Let \( \hat{\lambda} = \hat{\lambda}_R^2 \). If \( \lambda \xrightarrow{P} 0 \) for every \( \sigma \), then we must have
\[
0 = -\sigma^2/2 [\text{E} (\log f)^3 - [\text{E} (\log f)]^3] + 3/2 \sigma^2 \text{E} [\log f] \text{Var} [\log f].
\]

Proof: ACE minimizes
\[
\log \left[ N^{-1} \sum_{i=1}^{N} \{ y_i(\lambda) - f_i(\lambda) \}^2 \right] - \log \left[ N^{-1} \sum_{i=1}^{N} \{ y_i(\lambda) - \overline{y(\lambda)} \}^2 \right]
\]
Taking derivatives with respect to \( \lambda \), we get
\[
0 = \frac{N^{-1} \sum_{i=1}^{N} H_{14}(\hat{\lambda})}{N^{-1} \sum_{i=1}^{N} H_{12}(\hat{\lambda})} - \frac{N^{-1} \sum_{i=1}^{N} H_{13}(\hat{\lambda})}{N^{-1} \sum_{i=1}^{N} H_{11}(\hat{\lambda})}.
\]
\[ 0 = N^{-1} \sum_{i=1}^{N} H_{14}(\hat{\lambda}) - N^{-1} \sum_{i=1}^{N} H_{11}(\hat{\lambda}) - N^{-1} \sum_{i=1}^{N} H_{13}(\hat{\lambda}) + N^{-1} \sum_{i=1}^{N} H_{12}(\hat{\lambda}). \]

From Lemmas 1 - 3, we get,

\[
0 = \left[\sigma^2 E(\log f)\right]\left[\text{Var}(\log f) + \sigma^2\right] \\
- \sigma^2 \left[1/2 E(\log f)^3 + \sigma^2 E(\log f) - 1/2 E(\log f)E(\log f)^2\right] \\
= \sigma^2 E(\log f) \text{Var}(\log f) + \sigma^4 E(\log f) - \sigma^2/2 E(\log f)^3 \\
- \sigma^4 E(\log f) + \sigma^2/2 E(\log f) E(\log f)^2 \\
= -\sigma^2/2 E(\log f)^3 + \sigma^2 E(\log f) \text{Var}(\log f) \\
+ \sigma^2/2 E(\log f) \text{Var}(\log f) + \sigma^2/2 E(\log f)^3 \\
= -\sigma^2/2 \{E(\log f)^3 - [E(\log f)]^3\} \\
+ 3\sigma^2/2 E(\log f) \text{Var}(\log f).
\]

**Note:** Suppose \(E[\log f] = 0\). Then \(\lambda_{R^2} \rightarrow \lambda_0 = 0\) implies

\(E[\log f]^3 = 0\). This simple counterexample shows that ACE cannot be consistent in general.

### 4.1.2 Examples

At this point it will be illustrative to see how these estimators perform on data. Sixteen sets of data were generated according to the model \(\log(y_i) = \log(x_i) + \sigma \epsilon_i\), where the \(\epsilon_i\) are independent standard normal, the \(x_i\) are independent uniform \((0,10)\), \(\sigma = (1, .3, .5, 1)\) and \(N = (50, 100, 200, 500)\). The values for each of the estimators for
each data set are given in Table 4.1. Table 4.2 contains the summary measures of skewness, kurtosis and Spearman's $\rho$ for the residuals of the three models when $N = 100$. Also included are the summary measures based on the residuals from the no-transformation model, the log transformation model and the residuals based on the ACE procedure itself. These are recorded in the column labeled $\lambda_{ACE}$.

The maximum likelihood estimator $\lambda_{MLE}$ is closer to the true value of $\lambda$ than either $\lambda_{BF}$ or $\lambda_{R2}$ and as $N$ increases $\lambda_{MLE}$ tends to 0. From our calculations in the previous section we would expect $\lambda_{R2}$ to be biased. For these data it appears that $\lambda_{R2}$ consistently underestimates $\lambda$, and as $N$ increases the value of $\lambda_{R2}$ seems to stabilize around $-0.5$. For $\lambda_{BF}$ we record only the results for $\sigma = 0.1$. Breiman and Friedman suggested this estimator when there are reasons to believe that a transformation is needed but because a strong linear relationship exists between the response and the predictor ACE does not identify a transformation. Such a situation occurs in the transform-both-sides model when $\sigma$ is small. As seen in Chapter 2, this model is approximately equivalent to the power-of-the-mean variance model with power $\theta \equiv 1 - \lambda$. Data arising from such a model will exhibit a strong linear trend as well as heteroscedasticity.

Consider the simulated data where $N = 100$ and $\sigma = 0.1$. Figure 4.1 is a scatterplot of the untransformed data, $x$ and $y$. Notice that as $x$ increases the points appear to fan out. When ACE is applied to these data no transformation is indicated (Figure 4.2) yet the ACE residuals, $\hat{\theta}(y) - \hat{\psi}(x)$, are markedly heteroscedastic (Figure 4.3). Following Breiman and Friedman's suggestion, we apply ACE to $(y^\lambda, x^\lambda)$ for $\lambda$ ranging over the grid ($-1$, $0$) by increments of $0.1$. The
resulting estimate for $\hat{\lambda}_{BF}$ is $-0.6$. See Figure 4.4 where the ACE transformation for $y$ is plotted. We also plot scaled versions of the log transformation and the $\hat{\lambda}_{BF} = -0.6$ transformation. It is clear that the $\hat{\lambda}_{BF}$ overtransforms the data while the ACE procedure indicates no transformation.

The question arises as to why $\hat{\lambda}_{BF}$ does so poorly in relationship to $\hat{\lambda}_R^2$. The estimate $\hat{\lambda}_{BF}$ is selected as that value of $\lambda$ over the grid for which the $R^2$ value is maximized when ACE is applied to $(y^{\lambda}, x^{\lambda})$. We compute $R^2(\lambda=0) = 0.9718$ and $R^2(\lambda=-0.6) = 0.9997$. Examination of the transformation plot for $\theta(y^{-0.6})$ versus $y^{-0.6}$ (Figure 4.5) shows that the spacing along the $y^{\lambda}$ axis has been manipulated so that there are two points far out in the $y^{\lambda}$ direction. Similar spacing occurs for the transformed $x$ plots. This occurs because values of $y$ near 0 are being driven to infinity as $\lambda$ decreases. In order to avoid the problem of small observations we multiplied the observations by 100 forcing all observations to be greater than 1. Exactly the same sort of behavior occurs for these scaled data, where two points are being forced to 1 and the rest are near 0. This subsequent spacing of the untransformed variables causes the increased $R^2$ when there is no real increase in model performance.

We have seen that ACE does not perform particularly well in transform-both-sides situations when $\sigma$ is small. We investigate what happens when $\sigma$ is large, that is, the "strong linear relationship" is obscured by the increased variation. Consider the simulated data where $N = 100$. As $\sigma$ increases the ACE transformation plots indicate increasingly severe transformations. The transformed $x$ plots, while also indicating transformations, are in each case milder than those
indicated for the $y$ variable. The ACE transformation at $\sigma = .5$ seems
to lie very close to the log transformation. For $\sigma < .5$ the transfor-
mation is too mild and at $\sigma = 1$ the transformation is much too severe.
See Figures 4.2, 4.6, 4.7 and 4.8 for comparison. If we use the
summary measures from the $\lambda = 0$ model as a reference guide the small-
$\sigma$ residuals seem to be heteroscedastic and approximately symmetric.
For $\sigma = 1$ all three measures are large compared to the reference
model. Figure 4.9 is a scatterplot of the ACE residuals from the
$\sigma = .5$ model whereas Figure 4.10 shows the ACE residuals from the
$\sigma = 1$ model. Figure 4.11 illustrates the ACE transformation, and
scaled versions of the log transformation and the $\lambda_{R^2}$ transformation
for the $\sigma = 1$ model. The $\lambda_{R^2}$ seems to well approximate the ACE
transformation but both are more severe than the underlying trans-
formation.

The difficulty that ACE is encountering with data for large $\sigma$ is
that a number of data points, although small in number relative to the
bulk of the data, can be viewed as extreme data points, "outliers" in
the Breiman and Friedman terminology. In Figure 4.8 all data such
that $y > 10$ are effectively truncated in the transformed values.
Figure 4.12 is a scatterplot of the $(x,y)$ values for the $\sigma = 1$ model.
It is clear that the points $y > 10$ are separated from the bulk of the
data and are outliers in that sense.

It is interesting that ACE seems to do fairly well at $\sigma = .5$ but
not at $\sigma = 1$ or .1. It appears that as $\sigma$ increases $\lambda$ decreases.
Therefore it is reasonable to conjecture that for some $\sigma$ the ACE
transformation will approximate the true transformation. One must
then question whether or not, for a given set of data, the errors are
sufficiently large or small enough for ACE to perform adequately well. The severity of the transformation seems to be dictated by the size of \( \sigma \) rather than the need for transformation. It is clear from these examples that the ACE procedure does not always account for heteroscedasticity, nor does it induce symmetry in the error distribution.

The transformation plots provided by ACE are imprecise and only give a general idea of what \( \lambda_{\text{ACE}} \) should be. If estimation of \( \lambda \) is the primary goal, ACE will provide only a rough estimate of \( \lambda_{\text{ACE}} \) which may be a very biased estimate for the true \( \lambda \). If prediction is the goal, there is a predictive ACE procedure (PACE) developed by Friedman and Owens (1985). Since the ACE procedure yields inconsistent estimates for \( \lambda \) in the transform-both-sides setting when \( \lambda = 0 \), we might question the predictive effectiveness of PACE when applied to data arising from a transform-both-sides model.

The parameterized ACE criterion differs from the maximum likelihood criterion by the scaling factor in the denominator. The ACE criterion scales by the variance of the transformed response while maximum likelihood scales by the Jacobian of the transformation. An area of further research would be to adapt the ACE criterion to be like the maximum likelihood criterion. Since the Jacobian is the derivative of the transformation, with the availability of a data smoother, numerical derivatives can easily be computed for \( \theta(y) \) at each point and the appropriate scaling can be accomplished.
4.2 Using ACE to Determine Model Suitability

In the transform-both-sides setting if there is no error in the system then \( y_1 = f(x_1, \beta) \). In a situation where there are two competing models, say \( f_1 \) and \( f_2 \), plotting \( y \) versus \( f_1 \) and \( f_2 \) would be an easy visual check on the appropriateness of the assumed function if no transformation is required. However if an unknown transformation is involved this method is not applicable and provides a setting where ACE might usefully be employed to determine which is the more appropriate model.

The ACE procedure maximizes a criterion \( \Psi \) such that

\[
\max_{h, g} \Psi[h(y), g(x)] = \Psi[\theta_1(y), \phi_1(x)].
\]

If ACE acts upon \( y \) and \( f(x, \beta) \) rather than \( y \) and \( x \) then

\[
\max_{h, g} \Psi[h(y), g(f(x))] = \Psi[\theta_2(y), \phi_2(f(x))].
\]

Thus \( \phi_1(x) = \phi_2(f(x, \beta)) \), and it follows that \( f(x) \) is approximated by \( \phi_2^{-1} \circ \phi_1(x) \). If \( f(x, \beta) \) is a suitable model then \( \theta_1, \theta_2 \) and \( \phi_2 \) should all approximate the same transformation. Plotting \( \theta_1^{-1} \circ \phi_1(x) \) should yield approximately the same results as plotting \( \theta_2^{-1} \circ \phi_2(f(x, \beta)) \) when \( f(x, \beta) \) is the true model. Comparing plots of \( \theta_2^{-1} \circ \phi_2(f(x, \beta)) \) for different functions \( f(x, \beta) \) with the plot of \( \theta_1^{-1} \circ \phi_1(x) \) should indicate the better fitting model. In order to construct these plots an assumption must be made about the form of the \( \theta_1 \) and \( \theta_2 \) transformations. If the choice is limited to the Box-Cox family, the transformation can be estimated by overlaying on the ACE
transformation plot standardized Box-Cox functions for different $\lambda$ and selecting the one that best approximates the $\theta$ curve.

A similar but slightly different setting is where there are two competing models but the response also differs in each model, say $h_1(y,z) = f_1(x,\beta)$ and $h_2(y,z) = f_2(x,\beta)$, where $z$ may contain some of the covariates. Again we wish to determine the better fitting model. By applying ACE to each model, the transformation plots should indicate similar transformations for the better model and somewhat different transformations if the model is inappropriate.

4.2.1 The Atlantic Menhaden Fishery Data

An area of science where transform-both-sides has been successfully used is in the area of fisheries management. See Carroll and Ruppert (1984) and Ruppert and Carroll (1985). An important issue in the management of fisheries is the spawner-recruit relationship. The way in which this relationship is modeled can have a large impact on the subsequent management policies of the fishery. Let $y$ = the number of recruits and $x$ = the number of spawners. There are two models often used to model the spawner-recruit relationship:

$$y = 1/(\beta_1 + \beta_2/x) \quad \text{Beverton-Holt (1957)}$$

$$y = \beta_1 x \exp(\beta_2 x) \quad \text{Ricker (1954)} .$$

In the Beverton-Holt model, large numbers of spawners lead to an asymptote in recruitment, whereas the Ricker model implies that increasing spawners ultimately result in decreasing recruitment. See Carroll and Ruppert (1984) for a more complete discussion of these
models.

The data are from the Atlantic menhaden fishery and are listed and analyzed in Carroll and Ruppert (1984). Although their results indicate that both models are reasonable since the fitted curves are similar over the observed range, they decided upon the Beverton-Holt model because there was no real evidence for asymptotically decreasing recruitment. We investigate if ACE can be useful in determining the better model.

We applied ACE to \((y, x)\), \((y, f_{BH}(x, \hat{\beta}))\) and \((y, f_{R}(x, \hat{\beta}))\), where \(\hat{\beta}\) are the estimated parameters from an untransformed fit for each model, and the subscripts BH and R indicate which spawner-recruit model is used. The transformations \(\theta_1\), \(\theta_2\) and \(\theta_3\) are approximately the same, say \(\theta\), and \(\hat{\lambda}\) based on these plots is approximately \(-1\). See Figure 4.13. Carroll and Ruppert (1984) find \(\hat{\lambda}\) to be approximately \(-.7\) for both models. Thus \(\lambda = -1\) seems a reasonable choice. Note also that \(\lambda = -1\) linearizes the Beverton-Holt model but not the Ricker model.

We then plot \(\theta^{-1} \circ \phi_1(x)\), \(\theta^{-1} \circ \phi_2(f_{BH})\) and \(\theta^{-1} \circ \phi_3(f_{R})\) and compare the results. See Figure 4.14. The \(\theta^{-1} \circ \phi_1(x)\) curve compares favorably with the Beverton-Holt curve and not the Ricker model. In this example ACE appears to select the Beverton-Holt model as being the more appropriate model. This agrees with the conclusion of Carroll and Ruppert but is not based on any considerations of the problem as was their decision.
4.2.2 Chemical Reaction Data

The data set of Carr (1960) on the isomerization of pentane is a good candidate for demonstrating ACE's ability to identify the better model in the second setting. The data are listed and the model is discussed in Box and Hill (1974). Pritchard, Downie and Bacon (1977) and Carroll and Ruppert (1984) further discuss and analyze the data. Our purpose here is not to reanalyze the data but to see if the ACE procedure when applied to different models will select the more appropriate model. One proposed model for these data is

\[
\text{Model 1: } y = \frac{\beta_1 \beta_3 (x_2 - x_3/1.632)}{1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3}
\]

This model can be linearized by transform-both-sides with \( \lambda = -1 \) and yields

\[
\text{Model 2: } 1/y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3
\]

where the \( \alpha \)'s are the appropriate functions of the original parameters. Carr (1960) chose to linearize the model by the following transformation:

\[
\text{Model 3: } (x_2 - x_3/1.632)/y = 1/\beta_1 \beta_3 + (\beta_2/\beta_1 \beta_3) x_1 + (1/\beta_1) x_2 + (\beta_4/\beta_1 \beta_3) x_3.
\]

In the previously mentioned analyses the inverse transformed model 2 induced marked heteroscedasticity while there was slight heteroscedasticity for the original untransformed model. Box and Hill (1974) and Carroll and Ruppert (1984) find values for \( \hat{\lambda} \) to be .8 and .71, respectively, indicating a slight transformation for the original
nonlinear model 1. Carr (1960) rejects model 3 as inadequate. It will be interesting to see how ACE performs on these models.

In each of the three models we form $f_i(x, \hat{\beta})$ where the $\hat{\beta}$ are estimated values of the parameters and the subscript $i = 1, 2,$ and 3 refers to the corresponding model. We then apply ACE to $(y, f_1)$, $(y^{-1}, f_2)$ and $(h(y,x), f_3)$. Figures 4.15, 4.16, and 4.17 show the transformation plots for each model. We see that model 1 calls for no transformation and model 2 shows a slight convexity, thereby supporting the claim that the inverse transformation is too strong. Model 3 indicates a need for transformation but it is not the transform-both-sides situation.

Model 1 ($\lambda = 1$) and model 2 ($\lambda = -1$) belong to the same transform-both-sides transformation family while model 3 does not. Supporting this, the ACE transformation plots indicate that models 1 and 2 are candidates for the transform-both-sides model, whereas model 3 is completely inadequate from a transform-both-sides point of view.

We note here that ACE's performance on model 1 does not indicate a transformation but previous parametric analysis did. As seen before ACE does not control for heteroscedasticity. Thus it is not surprising that the slight transformation indicated by parametric methods is missed by ACE. It is encouraging to see, at least in this example, that ACE does behave differently when confronting two very different models and seems to correctly indicate the more appropriate model.
4.3 **ACE as a Diagnostic Tool**

Often in parametric settings there are influential points which have a large effect on the value of the estimated parameters. Robust methods exist for estimation that minimize the effect of the extreme points, and case deletion diagnostics are useful in identifying such points.

Breiman and Friedman (1985) state that the ACE procedure does not perform well in the presence of extreme data points. The question then arises as to whether or not ACE can locate points that are not "outliers" in the Breiman-Friedman sense yet are influential in determining the estimated transformation.

In some numerical examples the deletion of a single point causes the ACE transformation plots to change dramatically whereas in other examples the deletion of a known influential point hardly affects the plotted transformation. We need a diagnostic that does not depend so heavily on a subjective assessment of the ACE plots yet is able to locate influential points even when the plotted transformation does not appear to change much.

In trying to form a diagnostic based on the ACE procedure we are limited in the amount of available information. There is no theoretical distributional framework so methods based on likelihood distance or influence functions are not available.

Since ACE maximizes $R^2$, the diagnostic $R^2_{D_i} = | R^2 - R^2_{(i)} |$ is considered, where the subscript $i$ indicates that the $i$th point is deleted and the procedure is reapplied to the reduced data set. A second diagnostic, $DACE_{i}$, is in the spirit of Cook's (1977) distance
and DFFITS (Belsley, Kuh, and Welsh, 1980) both of which measure the effect of case deletion in a linear model based on the least squares fitted values. See also Cook and Weisberg (1982) for a discussion of single case deletion diagnostics. Our situation is a little more complicated. The model is nonlinear with the estimated transformations occurring on both sides of the equation. Since our goal is to see how much a single point affects the estimated transformation we decided to compute \[ \left[ \theta(y_j) - \theta(1)(y_j) \right] \text{ and } \left[ \phi(x_j) - \phi(1)(x_j) \right] \] where \( \theta(y) \) and \( \phi(x) \) are the ACE transformed values based on the full data set and \( (1) \) represents the reduced data set where the \( i \)th point has been removed. Notice that we are unable to use \( \theta(y_i) \) and \( \phi(x_i) \) since the \( i \)th point has been deleted and there is no corresponding \( \theta(1)(y_j) \) and \( \phi(1)(x_j) \) available to form the DACE\textsubscript{1} diagnostic. In order to combine the information about \( \theta \) and \( \phi \) we form the following diagnostic

\[
DACE_1 = \sum_{j=1, j\neq i}^{n} \left[ \left( \theta(y_j) - \theta(1)(y_j) \right) - \left( \phi(x_j) - \phi(1)(x_j) \right) \right]^2
\]

\[
= \sum_{j=1, j\neq i}^{n} \left[ \left( \theta(y_j) - \phi(x_j) \right) - \left( \theta(1)(y_j) - \phi(1)(x_j) \right) \right]^2
\]

\[
= \sum_{j=1, j\neq i}^{n} \left[ RES_j - RES_{(1)j} \right]^2
\]

\[
= \left[ RES - RES_{(1)} \right]^T \left[ RES - RES_{(1)} \right]
\]

where \( RES_j \) represents the \( j \)th ACE residual and \( RES_{(1)j} \) represents the
jth ACE residual for the reduced data.

Consider the performance of these two estimators on the Skeena River data set discussed in Chapter 2. In that analysis we delete observations 12 and 16. Carroll and Ruppert (1986) analyze these data using robust procedures. They find observation 12, the year in which the rockslide occurred, to be an outlier. Observation 4, whose influence is masked by observation 12, is also determined to be an influential point. They find observations 5, 16, 19 and 25 to be somewhat influential but with the removal of 12 and the downweighting of 4 the other points seem to stabilize. We compare these results with the outcome of our two diagnostics.

See Figure 4.18 for plots of the diagnostics based on the full data set. The $R^2D_1$ diagnostic does not indicate that observation 12 is an influential point. Based on this plot it seems to have only moderate influence and is smaller than that of several other observations. This is not the case for the $DACE_1$ diagnostic. The value of $DACE_1$ when 12 is deleted is twice as large as any of the other values. It is interesting that the value of both diagnostics for observations 8, 13, 16, 19 and 25 is moderately large but observation 4 is not flagged by either diagnostic.

We delete observation 12 and recompute the diagnostics. See Figure 4.19. Points 4 and 16 are clearly flagged for both diagnostics. As in Carroll and Ruppert's analysis the influence of observation 4 is masked by observation 12. For $DACE_1$ the points 8, 13, 19 and 25 again show moderate influence, whereas observations 6, 13, and 22 have rather large values for $R^2D_1$.

We recompute the diagnostics when observations 12 and 4 are
removed and then again when observations 12 and 4 are deleted. In the first case when observations 4 and 12 are deleted both diagnostics are stable except at the points 11, 17 and 25. The ACE procedure constructs transformations such that there is a large residual at point 16. The value for $R^2_{(1)}$ is not computed by the ACE procedure for these points and the corresponding value for $DACE_1$ becomes very large. When observations 12 and 16 are deleted a large residual occurs at the fourth observation for $\theta_{(21)}(y)$. Observation 8 also shows influence.

The results are not conclusive. Observation 4 when 12 and 16 are deleted and observation 16 when 12 and 4 are deleted continue to cause points to be flagged as influential even though the points themselves do not appear to be identified as influential by either of the diagnostics. It appears that retaining observation 16 creates more problems than retaining observation 4. Using also the rationale that observation 16 occurs four years after the rock slide (observation 12), the approximate time for which it takes the recruits to mature and return to the river to spawn, we have some additional justification for deleting observation 16 from the analysis as well as observation 12, as we did for the example in Chapter 2.

It is important to realize that case deletion diagnostics in the ACE setting are of limited value. Nonparametric regression techniques should be applied to large sets of data, but as the number of observations increases, case deletion methods become impractical. However it is interesting to see that in this example for a relatively small data set, the performance of $DACE_1$ essentially agrees with the robust analysis of Carroll and Ruppert (1986).
4.4 Discussion

The ACE procedure and the transform-both-sides regression model are alike in that both estimate unknown transformations. In transform-both-sides regression the transformation parameter $\lambda$ controls the way in which error enters the model and does not contribute to the structure of the regression relationship. The primary goal in transform-both-sides is to model the data in such a way to obtain the most efficient regression estimators. In ACE the transformations $\theta(.)$ and $\varphi(.)$ create the regression relationship between $x$ and $y$ and in this sense ACE is a prediction model rather than an estimation model.

We saw in Section 4.1 that the parameterized transform-both-sides ACE criterion leads to inconsistent estimates for $\lambda$, when $\lambda = 0$. ACE when applied to the generated data of Section 4.1.2, is generally not sensitive to the skewness and heteroscedasticity of the data and the subsequent ACE transformations are not always the transformations used to generate the data, i.e., the log transformation, $\lambda = 0$.

In Section 4.2, based on two examples, ACE worked well as a goodness-of-fit tool to determine the better of two or more competing models. Since ACE is designed to deal with the structure of regression models we found ACE to be useful in identifying the model that previous analyses in the literature indicate as appropriate for the data.

The goals of ACE and transform-both-sides regression apparently differ in that ACE is a prediction tool and transform-both-sides regression is an estimation model. While the ACE algorithm may be a useful tool when trying to build a model or discriminate between
models that are structurally different, it is not clear at this point
how adequate ACE is for estimation. In order to further study the
effectiveness of ACE for estimation it is necessary to gain more
understanding about certain aspects of the ACE procedure such as the
relationship between sample size, variability and the tuning constants
for the smoothing procedure.

4.5 Appendix

We sketch the derivation of (4.2) and (4.3). Minimizing
\[ e^2 = E[\theta(y) - \varphi(x)]^2 \]
with respect to \( \varphi(x) \) given \( \theta(y) \) is the same as
finding the best mean square error predictor for \( \varphi(x) \). That is,
\[ E[\theta(y) - \varphi(x)]^2 = E[\theta(y) - E[\theta(y)|x]]^2 + E[\varphi(x) - E[\theta(y)|x]]^2 \]
which is minimized by \( \varphi(x) = E[\theta(y)|x] \). Similarly \( \theta(y) = E[\varphi(x)|y]/\|E[\varphi(x)|y]\| \)
minimizes \( e^2 \) with respect to \( \theta(y) \) given \( \varphi(x) \) where \( \theta(y) \) is scaled such
that \( \text{Var}[\theta(y)] = 1 \).

4.6 Tables and Figures

All tables and figures for the examples in Chapter 4 follow.
Table 4.1: Simulated Data.

Parameter estimates for $\hat{\lambda}_{\text{MLE}}, \hat{\lambda}_{R^2}, \hat{\lambda}_{BF}$.

<table>
<thead>
<tr>
<th>N</th>
<th>σ</th>
<th>$\hat{\lambda}_{BF}$</th>
<th>$\hat{\lambda}_{R^2}$</th>
<th>$\hat{\lambda}_{\text{MLE}}$</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>.1</td>
<td>-1.1</td>
<td>-0.86</td>
<td>-0.26</td>
<td>(-.44, -.04)</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>-</td>
<td>-0.88</td>
<td>-0.22</td>
<td>(-.41, .02)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>-</td>
<td>-0.92</td>
<td>-0.16</td>
<td>(-.37, .08)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-</td>
<td>-1.08</td>
<td>-0.03</td>
<td>(-.20, .15)</td>
</tr>
<tr>
<td>100</td>
<td>.1</td>
<td>-0.6</td>
<td>-0.45</td>
<td>-0.08</td>
<td>(-.18, .04)</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>-</td>
<td>-0.45</td>
<td>-0.06</td>
<td>(-.16, .05)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>-</td>
<td>-0.45</td>
<td>-0.04</td>
<td>(-.14, .07)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-</td>
<td>-0.36</td>
<td>-0.02</td>
<td>(-.09, .08)</td>
</tr>
<tr>
<td>200</td>
<td>.1</td>
<td>-1.7</td>
<td>-0.52</td>
<td>-0.04</td>
<td>(-.13, .05)</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>-</td>
<td>-0.57</td>
<td>-0.04</td>
<td>(-.12, .05)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>-</td>
<td>-0.49</td>
<td>-0.03</td>
<td>(-.10, .04)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-</td>
<td>-0.36</td>
<td>-0.02</td>
<td>(-.08, .04)</td>
</tr>
<tr>
<td>500</td>
<td>.1</td>
<td>-1.0</td>
<td>-0.55</td>
<td>0.02</td>
<td>(-.03, .08)</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>-</td>
<td>-0.57</td>
<td>0.02</td>
<td>(-.03, .08)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>-</td>
<td>-0.56</td>
<td>0.02</td>
<td>(-.03, .08)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-</td>
<td>-0.45</td>
<td>0.02</td>
<td>(-.02, .06)</td>
</tr>
</tbody>
</table>
Table 4.2: Summary measures for simulated data, N = 100

<table>
<thead>
<tr>
<th>N = 100</th>
<th>$\lambda = 1$</th>
<th>$\lambda = 0$</th>
<th>$\hat{\lambda}_{\text{MLE}}$</th>
<th>$\hat{\lambda}_{\text{ACE}}$</th>
<th>$\hat{\lambda}_{R^2}$</th>
<th>$\hat{\lambda}_{\text{BF}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Skewness</strong></td>
<td>$\sigma = .1$</td>
<td>.65</td>
<td>-.21</td>
<td>-.29</td>
<td>-.01</td>
<td>-2.64</td>
</tr>
<tr>
<td></td>
<td>$\sigma = .3$</td>
<td>1.70</td>
<td>-.21</td>
<td>-.31</td>
<td>-.11</td>
<td>-2.19</td>
</tr>
<tr>
<td></td>
<td>$\sigma = .5$</td>
<td>2.74</td>
<td>-.21</td>
<td>-.30</td>
<td>-.38</td>
<td>-1.69</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 1$</td>
<td>5.36</td>
<td>-.21</td>
<td>-.25</td>
<td>-.83</td>
<td>-0.52</td>
</tr>
<tr>
<td><strong>Kurtosis</strong></td>
<td>$\sigma = .1$</td>
<td>3.22</td>
<td>.25</td>
<td>.13</td>
<td>1.57</td>
<td>17.29</td>
</tr>
<tr>
<td></td>
<td>$\sigma = .3$</td>
<td>6.23</td>
<td>.25</td>
<td>.18</td>
<td>.76</td>
<td>14.98</td>
</tr>
<tr>
<td></td>
<td>$\sigma = .5$</td>
<td>12.06</td>
<td>.25</td>
<td>.22</td>
<td>.71</td>
<td>12.74</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 1$</td>
<td>37.21</td>
<td>.25</td>
<td>.25</td>
<td>3.59</td>
<td>4.31</td>
</tr>
<tr>
<td><strong>Spearman's $\rho$</strong></td>
<td>$\sigma = .1$</td>
<td>.6904</td>
<td>.1525</td>
<td>.0650</td>
<td>.4114</td>
<td>-.2213</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0001)</td>
<td>(.1298)</td>
<td>(.5203)</td>
<td>(.0001)</td>
<td>(.0269)</td>
</tr>
<tr>
<td></td>
<td>$\sigma = .3$</td>
<td>.6979</td>
<td>.1525</td>
<td>.0866</td>
<td>.1666</td>
<td>-.2236</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0001)</td>
<td>(.1298)</td>
<td>(.3919)</td>
<td>(.0977)</td>
<td>(.0253)</td>
</tr>
<tr>
<td></td>
<td>$\sigma = .5$</td>
<td>.7034</td>
<td>.1525</td>
<td>.1052</td>
<td>.0397</td>
<td>-.2259</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0001)</td>
<td>(.1298)</td>
<td>(.2974)</td>
<td>(.6947)</td>
<td>(.0239)</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 1$</td>
<td>.6969</td>
<td>.1525</td>
<td>.1385</td>
<td>-.4340</td>
<td>-.1799</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0001)</td>
<td>(.1298)</td>
<td>(.1694)</td>
<td>(.0001)</td>
<td>(.0733)</td>
</tr>
</tbody>
</table>
Figure 4.1: Simulated Data. Scatterplot of $y$ versus $x$ for $N = 100$ and $\sigma = .1$. 

SIMULATED DATA
$N = 100$, $\sigma = .1$
Figure 4.2: Simulated Data. The plotted points represent the ACE transformation and the solid line represents a scaled log transformation for $N = 100$ and $\sigma = .1$. 
**Figure 4.3:** Simulated Data. Plot of ACE residuals plotted against $x$ for $N = 100$ and $\sigma = .1$. 
Figure 4.4: Simulated Data. The jagged line represents the ACE transformation, the smooth solid line is the scaled log transformation curve, and the dashed line is the $\lambda = -0.6$ scaled transformation curve.
Figure 4.5: Simulated Data. The plotted ACE transformation when
ACE is applied to \((y^{BF}, f^{BF})\) where \(\hat{\lambda}_{BF} = -.6\) and
\(N = 100\) and \(\sigma = .1\).
Figure 4.6: Simulated Data. The smooth curve is the scaled log transformation curve and the points represent the plotted ACE transformation for $N = 100$ and $\sigma = .3$.
Figure 4.7: Simulated Data. The solid line represents the ACE transformation curve and the dotted line represents the scaled log transformation curve $N = 100$ and $\sigma = .5$. 
Figure 4.8: Simulated Data. The solid line represents the ACE transformation curve and the dotted line represents the scaled log transformation curve for $N = 100$ and $\sigma = 1$. 
Figure 4.9: Simulated Data. Plot of ACE residuals versus time for $N = 100$ and $\sigma = .5$. 
Figure 4.10: Simulated Data. Plot of the ACE residuals against $x$ for $N = 100$ and $\sigma = 1$. 
Figure 4.11: Simulated Data. Dashed line represents the scaled log transformation curve, the solid line represents the ACE transformation curve, and the dotted line represents the scaled $\hat{\lambda}_R^2$ transformation curve for $N = 100$ and $\sigma = 1$. 
SIMULATED DATA
N = 100, Sigma = 1

Figure 4.12: Simulated Data. Scatterplot of y versus x for 
\( \sigma = 1 \) and \( N = 100 \).
Figure 4.13: Atlantic Menhaden Data. The points represent the ACE transformation curve, the solid line is the scaled log transformation curve and the dotted line is the scaled inverse transformation curve.
Figure 4.14: Atlantic Menhaden Data. The solid line represents the ACE transformation, the dotted line represents the Ricker model, and the dashed line represents the Beverton-Holt model.
Figure 4.15: Chemical Reaction Data - Model 1.

ACE transformation plots for y and x.
Figure 4.17: Chemical Reaction Data

ACE transformation plots for $y$ and $f_3$. 
Figure 4.18: Skeena River Data. Diagnostic plots for the full data set, deleting one point at a time. Upper chart is for $R^2D_i$ and the lower is for DACE$_i$. 
Figure 4.19: Skeena River Data. Diagnostic plots for the reduced data set where observation #12 has been removed. Upper chart is for $R^2_D$ and the lower is for $DACE_1$. 
BIBLIOGRAPHY


Carroll, R. J., and Ruppert, D. (1986). Diagnostics when transforming the regression model and the response. manuscript.


