AN ABSTRACT OF THE THESIS OF

Mohammad Fazli Qadir for the degree of Doctor of Philosophy in Statistics presented on August 27, 1993.

Title: Using Percentile Regression for Estimating the Maximum Species Richness Line.

Abstract approved: _____
N. Scott Urquhart

The Index of Biotic Integrity (IBI) has proved useful in evaluating the impact of environmental insults in aquatic environments. The "maximum species richness line" has a central role in its evaluation. Aquatic scientists evaluating IBI have fit this line by eye. A percentile regression line provides a statistically-based estimate of the maximum species richness line. We define percentile regression and explore its estimation in several situations. The form of typical data is modeled by a normal distribution with a mean and standard deviation which each change along lines. For the regression-type line of the form $\beta_0 + \beta_1 X$, we use maximum likelihood on a general linear model to estimate a 100 p^{th} percentile line. Two nonparametric methods also are explored for estimating the 100 p^{th} percentile regression line. A simulation study compares the approaches. These approaches for estimating a percentile line provide practical alternatives for fitting the maximum species richness line. The maximum likelihood approach provides an efficient estimate of regression percentiles, provided the data follow

the assumed model. Otherwise one of the nonparametric regressions provides a more defensible approach.

A method of analysis is required which is insensitive to misspecification of the distribution and/or to possible outliers. We propose an adaptation of robust regression to estimate percentile regression lines. This robust method uses weighted least squares, where the weights are calculated from a beta function. It offers the user of the maximum species richness line a robust alternative to the methods proposed by Fausch et al. (1984) or those mentioned above. We compare this approach with the two earlier approaches. Simulated data sets containing heteroscedasticity are used to compare the approaches; the basis of comparison is the mean-squared error. The maximum likelihood procedure based on a linear model dominates both nonparametric and robust procedures, when the assumptions of the model are satisfactory. Otherwise the robust procedure performs well; it needs to be explored further.

Using Percentile Regression for Estimating The Maximum Species Richness Line

by

Mohammad F. Qadir

A THESIS

submitted to

Oregon State University

In partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Completed August 27, 1993

Commencement June 1994

APPROVED:

Redacted for Privacy

Professor of Statistics in Charge of major

Redacted for Privacy

Head of Department of Statistics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented August 27, 1993

Typed by Mohammad F. Qadir

ACKNOWLEDGMENT

It gives me great pleasure to thank Dr. N. Scott Urquhart, for serving as my advisor and sorting out this thesis topic. I appreciate his patience, guidance, invaluable technical advice and many hours of conversation about the topic and statistics. I also appreciate his support, careful attention and interest that he showed in my work. I consider myself fortunate having him as major professor.

I would like to express my sincere gratitude to my other committee members, Drs. David R. Thomas, Daniel W. Schafer, Paul A. Martaugh, Dawn Petters, and Robert A. Duncan for their assistance and continuous support and valuable suggestions in planning my course work and review of this manuscript. I also appreciate the help and many useful suggestions from Dr. David S. Birkes.

Thanks are extended to Dr. Justus Seely, chairman department of Statistics, faculty members, staff and students for their assistance throughout my stay at Oregon State University.

I reserve a very special thanks for my family. I want to dedicate this thesis to my parents for their love, to my brothers and sisters for their encouragement, and especially to my wife, sons, and daughters, for their great and uncountable sacrifices during my four year stay abroad.

Finally, I would like to extend my Thanks to the Government of Pakistan and US AID for their financial support.

TABLE OF CONTENTS

Chapter			Page
I	INTRODUCTION		1
II	PERCENTILE REGRESSION: A TOOL FOR RELATING SPECIES RICHNESS TO SYSTEM SIZE AND ENVIRONMENTAL IMPACT		5
	1	Abstract	5
	2	Introduction	6
	3	Model and Fitting	8
	4	An Iteratively Reweighted Least Squares Approach	10
	5	A Nonparametric Approach	12
	6	Example	14
	7	Simulation	21
	8	Conclusions	30
	REFERENCES		31
	APPENDIX		33
III	ADAPTING ROBUST REGRESSION TO PERCENTILE REGRESSION FOR ESTIMATING THE MAXIMUM SPECIES RICHNESS LINE		36
	1	Abstract	36
	2	Introduction	37
	3	Development of an Alternative	38
	4	Adaptation for the Problem	48
	5	Simulation Results	50
	6	Conclusions	56

TABLE OF CONTENTS CONTINUED

Chapter		Page
	REFERENCES	57
IV	SUMMARY AND CONCLUSIONS	60
BIBLIOGRAPHY		64

LIST OF FIGURES

Figure		Page
2.1	Plot of Ohio data (species richness vs. drainage area), illustrating violation of the assumption of linearity.	15
2.2	Plot of Ohio data after taking log transformation of "drainage area" . The linearity assumption now appears reasonable.	16
2.3	Normal probability plot of residuals from model of equation 2.4.	17
2.4	Plot of Ohio data with 10^{th}, 50^{th}, and 90^{th} percentile regression lines, obtained from maximum likelihood approach.	18
2.5	Plot of upper 20% of Ohio data, along with OLS line for those 20% data. This line gives nonparametric estimate of maximum species richness line.	19
2.6	Average estimates of the intercept by three different approaches from 2500 simulations.	22
2.7	Average estimates of the slope by three different approaches from 2500 simulations.	23
2.8	Mean-squared error of the estimates of the intercept from 2500 simulations.	25
2.9	Mean-squared error of the estimates of the slope from 2500 simulations.	26
2.10	Data from a mixture of two normal densities with 90^{th} percentile lines, maximum likelihood procedure (solid line) nonparametric (dashed line).	29
3.1	Number of international phone calls from Belgium with OLS fit (solid line) WLS fit (dashed line) and LMS fit (dotted line).	45

LIST OF FIGURES CONTINUED

Figure		Page
3.2	Mean-squared error of the estimate of intercept for 90th percentile line from 2500 simulations.	53
3.3	Mean-squared error of the estimate of slope for 90th percentile line from 2500 simulations.	54
3.4	Three different estimates of maximum species richness line, linear model (solid line), nonparametric (dashed line) and weighted least squares (dotted line).	55

LIST OF TABLES

Table		Page
2.1	Average estimates of the intercept and slope of the 90th percentile line for stated sample sizes and methods of estimation for 1000 and 2500 simulations.	21
2.2	Standard errors of intercept and slope of the 90th percentile line, for 1000 and 2500 simulations.	24
2.3	Average estimates of the intercept and slope of the 90th percentile line from mixture of two normal distributions.	27
2.4	Standard error of the intercept and slope of the 90th percentile line from mixture of two normal distributions.	28
3.1	Residuals from four different fits to the stackloss data.	47
3.2	Values of α and β for different values of tuning constant to find selected percentile regression lines.	51
3.3	Using values from table 3.2, the estimates of intercept and slope for 90th percentile regression line obtained from 1000 simulations.	52

USING PERCENTILE REGRESSION FOR ESTIMATING THE MAXIMUM SPECIES RICHNESS LINE

CHAPTER I

Introduction

Toxic chemicals and other human impacts effect water quality and species richness of fish populations in both lakes and streams. For biological monitoring Karr (1981) proposed a useful tool, the Index of Biotic Integrity (IBI). The IBI is a quantitative assessment that can be used to evaluate human effects on streams and lakes. The IBI was defined to include a range of attributes of fish assemblages. Data are obtained for each of twelve metrics at a given site and evaluated in light of what might be expected at an unimpacted or relatively unimpacted site located in a similar geographical region and on a stream of comparable size. Rating each metric according to the corresponding scoring criterion developed for the site, Karr (1981), Fausch *et al.* (1984), Karr *et al.* (1986) assigned a score of 5, 3, or 1 according to the species richness, compared to undisturbed reference sites. The total score was then calculated by adding the ratings assigned to the twelve metrics. The maximum total (60) indicates a site without perturbation. The minimum score of 12 is possible when all metrics reflect extreme degradation.

Consider a two-dimensional plot of the total number of fish species in a water body against stream size, lake size, or watershed area for sites within a region. Such plots usually display a fan shape of points, since the number of species generally

increases with the size of water body. Such fan-shaped plots reflect the nature of sampling stream fish communities. The upper bound of this fan-shaped plot can be represented by a straight line which forms an upper bound for the fish community in that region. Limnologists traditionally have fit this line "by eye". See Fausch, et al. (1984) and Karr, et al. (1986), for example. Such a line, an upper bound for about 95% of the sites, is called the maximum species richness line. The purpose of calculating species richness versus stream size relationships was to predict the expected total fish species richness, for application of the index of biotic integrity. This line defines an "excellent" fish community for purposes of scoring IBI. Lines delineating other proportions of the plot are used for assigning other score values for IBI.

The problem of estimating the maximum species richness line motivated this thesis. We present here a statistical formulation for estimating the needed line using objective computational methods in place of visual ones. We advance three different approaches: One uses a regression model assuming normally distributed residuals with heteroscedastic variance; the second is a nonparametric method of fitting an ordinary least squares line to a subset of the data; and the third is a robust procedure for estimating the maximum species richness line.

The first method we propose for estimating the line is based on a general linear model. This model assumes a normally distributed response variable with mean and standard deviation that are linear functions of the regressor variable. The proposed model **(2.4)** with four unknown parameters requires an iterative solution for their estimation. The Newton-Raphson method can be used to find maximum likelihood

estimates. A less complicated but more time consuming method of iteratively reweighted least squares also can be applied to a reparametrization of the original model.

The second approach uses nonparametric estimation of the maximum species richness line. Hogg (1975) proposed a nonparametric method for estimating a percentile regression for situations in which the usual distributional assumptions fail. He proposed dividing the data into halves at the median of X and then selected a specified number of observations from each half, where the proportion selected in each half depends on the percentile regression of interest. He extended this nonparametric approach by dividing the data at the quartiles of X and then selecting the required number of observations from each quarter and so on. We propose a simple extension of Hogg's method which can be applied in situations addressed here.

Robust regression has emerged as an alternative to ordinary least squares estimation, Birkes and Dodge (1993). If all the assumptions of the model are satisfied and the data contain no apparent outliers, ordinary least squares produces desirable estimates, but if the assumptions are violated or outliers occur in the data then robust estimation provides a practical alternative to classical methods. For the robust regression method we propose a new objective function for estimation. We adapt this method to estimating of the maximum species richness line.

Chapter II presents two approaches to estimating the maximum species richness line: maximum likelihood based on a linear model and the nonparametric method. The method of calculating the maximum likelihood estimate by iteratively reweighted least squares, using a profile likelihood also is presented as a computational alternative. Three

different selection procedures for the nonparametric method are discussed and compared. Chapter III presents the method of robust regression for estimation of the maximum species richness line based on a new objective function for estimation of multiple regression parameters. Chapter IV presents summary and conclusions.

CHAPTER II

PERCENTILE REGRESSION: A TOOL FOR RELATING SPECIES RICHNESS TO SYSTEM SIZE AND ENVIRONMENTAL IMPACT

by

Mohammad F. Qadir[1] and N. Scott Urquhart[2]

1. Abstract

The Index of Biotic Integrity (IBI) has proved useful in evaluating the impact of environmental insults in aquatic environments. The "maximum species richness line" has a central role in its evaluation. Aquatic scientists evaluating IBI have fit this line by eye. A percentile regression line provides a statistically-based estimate of the maximum species richness line. We define percentile regression and explore its estimation in several situations. The form of typical data is modeled by a normal distribution with a mean and standard deviation which each change along lines. For the regression-type line of the form $\beta_0 + \beta_1 X$, we use maximum likelihood on a general linear model to estimate a 100 p^{th} percentile line. Two nonparametric methods also are explored for estimating the 100 p^{th} percentile regression line. A simulation study compares the approaches. These approaches for estimating a percentile line provide practical alternatives for fitting the maximum species richness

---

[1] Department of Statistics, University of Peshawar, N.W.F.P., Pakistan; formerly a graduate student in the Department of Statistics, Oregon State University, Corvallis, OR 97331-4606, USA.

[2] Department of Statistics, Oregon State University, Corvallis, OR 97331-4606, USA.

line. The basis of comparison is the mean-squared error, the maximum likelihood approach provides an efficient estimate of regression percentiles, provided the data follow the assumed model. Otherwise one of the nonparametric regressions provides a more defensible approach.

2. Introduction

Toxic chemicals and other human factors affect species richness of fish populations in both lakes and streams. Karr (1981) proposed the Index of Biotic Integrity (IBI) as a useful tool for biological monitoring in this context. The IBI was defined to include a range of attributes of fish assemblages.

Consider a two-dimensional plot of the total number of fish species in a water body for sites within a region against stream size, lake size or watershed area. Although the number of species generally increases with the size of the water body, such plots also usually display a fan-shape of points such as in Figure 2.2. The upper bound of this fan-shaped plot represents a straight line forming an upper bound on the data scatter. Sometimes the fan expands down to the horizontal axis giving a data plot shaped like a right triangle. This fan-shape reflects true variation in species richness across the population of water bodies as well as the effects of various environmental insults. A line known as maximum species richness line forms the upper bound for about 95% of the sites. This line is assumed to relate species richness to size of water body in the absence to environmental insults. Potential species richness of a site can be predicted from its size; this line defines an "excellent" fish community for purposes of scoring IBI. Thus, when the number-of-species metric of the IBI is scored, the plot of total species richness

for a given aquatic site is high, medium or low when compared to an "excellent" fish community for the region. Limnologists traditionally have fit this line "by eye". See Fausch, *et al* (1984) and Karr, *et al* (1986), for example. Lines delineating other proportions of the plot are used to identify the high, medium and low categories.

We present here a statistical formulation for this biological problem, and advance methods for estimating the needed lines using objective computational methods in place of subjective visual ones. We advance two different kinds of methods, one a model-based approach assuming approximately normal data with heteroscedastic variance, and a nonparametric approach of fitting a least squares line to the upper $2\times(1-p)\%$ of data points. In applications to IBI, the upper 95% line usually is used, but a median, lower 95% and other percentile lines also are used. Our methods apply to a range of percentages. The first method can be applied to any percentage while the second one allows only certain isolated percentages. Both methods have advantages and disadvantages, as discussed in section 6.

In section 3 we define a 100 p^{th} percentile regression line and use maximum likelihood based on a linear model to estimate it, assuming a heteroscedastic normal model. This statistical model provides a reasonable approximation to situations of the sort to which IBI is applied in aquatic biology. Section 4 presents an iteratively reweighted least squares approach to the computation using the profile likelihood method. Section 5 presents a nonparametric approach for the 100 p^{th} percentile regression, using a relevant part of the data set. In this section three different ways of selecting a fraction of data are considered and compared. The illustration in section 6 is based on real data

from Ohio streams. We present some simulation evaluations in section 7. An appendix covers technical details needed to implement the procedure.

3. Model and Fitting

Suppose that Y_i denotes the i^{th} value of a response variable and X_i its corresponding regressor variable for $i = 1, 2, \ldots . n$. The standard regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, 2, \ldots . n . \tag{2.1}$$

has β_0 and β_1 as the unknown intercept and slope parameters; the ϵ_i denote unobserved random deviations which often are assumed to be distributed identically and independently as normal random variables having a mean of zero and unknown, but constant, variance σ^2. *PERCENTILE REGRESSION* can be defined for this model by reference to the underlying normal distribution. If Z denotes a random variable having a standard normal distribution, then there is a constant $z_{1-\alpha}$ such that Prob $(Z > z_{1-\alpha}) = \alpha$. The line

$$Y = (\beta_0 + z_{1-\alpha}\ \sigma) + \beta_1 X \tag{2.2}$$

is a $100\alpha\%$ upper percentile regression line, because

$$\begin{aligned} & \textit{Prob}\ (Y_i > (\beta_0 + z_{1-\alpha}\ \sigma) + \beta_1 X_i) \\ = & \textit{Prob}\ (\beta_0 + \beta_1 X_i + \epsilon_i > (\beta_0 + z_{1-\alpha}\ \sigma) + \beta_1 X_i) \\ = & \textit{Prob}\ (\epsilon_i > z_{1-\alpha}\ \sigma) = \alpha \quad . \end{aligned} \tag{2.3}$$

Cases such as those depicted by Figure 2.2 can be modeled by an approximately normal model, but with variance which changes as a continuous function of X_i and perhaps some unknown parameters. Heteroscedastic regression models are used for the estimation of percentile line see for example Carroll and Rupert (1982), Devidian and Carroll (1987). Now specifically consider an approximately normal model with heteroscedastic variance of the following form:

$$Y_i = (\alpha_1 + \gamma_1 X_i) + (\alpha_2 + \gamma_2 X_i)e_i \qquad (2.4)$$

for $i = 1, 2, 3, \ldots, n$, and where the e_i are independent and distributed normally with a mean of 0 and a variance of 1. Thus α_1, α_2, γ_1 and γ_2 are four unknown parameters needed to specify the normal distribution. The assumption of a common variance ($\gamma_2 = 0$, and $\alpha_2^2 = \sigma^2$) as in Eq. **2.4** is referred to as the homoscedastic variance assumption. In the case depicted by Eq. **2.4,** $100\alpha\%$ upper percentile regression line would become

$$\alpha_1 + \gamma_1 X + z_{1-\alpha}(\alpha_2 + \gamma_2 X) \quad . \qquad (2.5)$$

Ordinary Least Squares (OLS) provides defensible estimates of the parameters of such a model, assuming homoscedastic and uncorrelated residuals. If, however, variance of the residuals changes across the observations, the heteroscedastic case, a weighted least squares analysis should be performed instead, provided the form of the inequality of variance is known. Such an analysis would apply to the Eq. **2.4** only if $\gamma_2 = 0$; otherwise this model requires a more general approach because the variance depends on two parameters. The log likelihood function for this normal distribution is

$$L = C - \sum_{i=1}^{n} \ln(\alpha_2 + \gamma_2 X_i) - .5 \sum_{i=1}^{n} \left(\frac{Y_i - \alpha_1 - \gamma_1 X_i}{\alpha_2 + \gamma_2 X_i} \right)^2 \qquad (2.6)$$

where C is a constant term. Because the first order partial derivatives of the log likelihood function with respect to α_1, α_2, γ_1 and γ_2 are nonlinear, maximization of this function requires an iterative solution, McCullagh and Nelder (1989). Many iterative methods are presented by Kennedy and Gentle (1980); for example, the Newton-Raphson method commonly has to be used to find maximum likelihood estimates. (See the Appendix for more details.) Although iterative methods require some computational resources, those needed for the present problem are not very limiting relative to modern standards. Once the parameters of Eq. **2.4** have been estimated, the $100p^{th}$ percentile of the distribution of Y can be estimated by the line

$$\hat{Y}_p = \hat{\alpha}_1 + \hat{\gamma}_1 X + z_p(\hat{\alpha}_2 + \hat{\gamma}_2 X) \qquad (2.7)$$

where, as before, z_p is the 100 p^{th} percentile of the standard normal distribution.

4. An Iteratively Reweighted Least Squares Approach

In the previous section we assumed model (**2.4**); another parametrization of that model can be written as

$$Y_i = (\alpha_1 + \gamma_1 X_i) + \sigma(1 + \rho X_i) e_i \quad , \qquad (2.8)$$

for $i = 1, 2, \ldots n$, and where e_i still are distributed independently and identically normal with a mean of 0 and a variance of 1. The parameters of this model relate to those of the previous section through: $\sigma = \alpha_2$ and $\rho = \gamma_2/\alpha_2$. If we let

$$\sqrt{W_i} = \frac{1}{(1 + \rho X_i)}$$

then, conditional on the value of ρ, Y_i is distributed normally with a mean of $(\alpha_1 + \gamma_1 X_i)$ and a variance of σ^2/W_i. The weighted least squares estimates for the parameters α_1, γ_1 and σ are

$$\hat{\gamma}_1 = \frac{\sum W_i(X_i - \bar{X}_w)(Y_i - \bar{Y}_w)}{\sum W_i(X_i - \bar{X}_w)} \quad ,$$

$$\hat{\alpha}_1 = \bar{Y}_w - \hat{\gamma}_1 \bar{X}_w \quad ,$$

$$\hat{\sigma}^2 = \frac{\sum W_i(Y_i - \hat{Y}_i)^2}{n-2} \quad ,$$

where

$$\bar{Y}_w = \frac{\sum W_i Y_i}{\sum W_i} \quad \textit{and} \quad \bar{X}_w = \frac{\sum W_i X_i}{\sum W_i} \quad .$$

When these estimates are inserted into the log likelihood function, the profile log likelihood function for ρ results:

$$L^*(\rho) = -\frac{n}{2} \ln \hat{\sigma}^2 - \sum \ln(1 + \rho X_i) - \frac{n}{2} \quad . \qquad \textbf{(2.9)}$$

The search for the value of ρ which maximizes this profile log likelihood function ($\hat{\rho}$) reduces the four-dimensional search specified in the previous section to a one dimensional search for $\hat{\rho}$, followed by the evaluation of the estimates of α_2, γ_2, and σ in terms of $\hat{\rho}$.

Because the form of the model in this section reflects merely a reparametrization of the original model, the results obtained from weighted least squares and direct maximum likelihood estimation should be the same; the latter method ordinarily would be slower, but could be implemented more easily by someone not familiar with multi-dimensional optimization.

5. A Nonparametric Approach

Hogg (1975) proposed a nonparametric method for estimating a percentile regression for situations in which the usual distributional assumptions fail. He proposed dividing the data into halves at the median of X and then selected the required number of observations from each half, where the proportion selected in each half depended on the percentile regression line of interest. Or similarly, the data could be divided into quarters at the quartiles of X and then selected the required number of observations from each quarter, and so on. For example an 80^{th} percentile line is determined so that exactly 20% of the data points in each interval will be above the line. This method may work for moderate-sized data sets, but becomes impractical for a large data set. Another nonparametric iterative method for simultaneous estimation of percentile curve is proposed by Angers (1978). We propose a simple extension of Hogg's method to the kind of situations addressed by this paper: Suppose we have n observations and the regressor variable X is arranged in ascending order. Divide the data into several subsets, each of size to be discussed below, based on consecutive values of the X variable, or equivalently select intervals of the X variable. Select the data point from each subset for

which the response variable Y has the maximum value; we call this selection of *one observation per interval.* Fit a line to the selected data points. Similarly for selection of *two observations per interval*, divide the data into half as many subsets as before, but then select the two observations with the highest two response values from each interval. Fit a line to these selected points.

The number of intervals and number of points selected per interval to estimate a p-percentile line is based on this: Given n observations, a regressor variable (X) arranged in ascending order and $p > 50$, we need $k = 2n(100 - p)/100$ intervals, where k is increased to the next larger integer, if k is not an integer. From each interval select the largest observation. When two observations are selected per interval, there should be $k = n(100-p)/100$ intervals with k is increased to the next larger integer, if necessary. This process appears to discard the rest of the data, a feature to which some users might object. The process really does not discard data; instead it identifies the linear trend formed by a relevant subset of the data points. Simulation studies show that these observations form a linear trend with a nearly constant variance. An ordinary least squares fit to the selected subset of the data thus gives a straight line analogous to that obtained by the approaches of the previous two sections, but without distributional assumptions necessary there. When the number of observations is not an exact multiple of the number of intervals, put m = next integer above n/k, observations in k - 1 intervals and the rest in another; locate the intervals so the interval having fewer than m observations is in the middle of the range of X to minimize the effect of this partial filled interval, and select no observations from this interval. Estimation of lower

percentile lines would proceed in a completely analogous fashion, except p above would be replaced by $100 - p$, and the lowest, rather than the highest, point(s) would be selected in each interval.

The simulation results presented in Section 7 show, for example, that the selection of two observations from intervals containing 10 observations gives a superior fit for a 90^{th} percentile regression line to selection one observation from intervals of 5 observations. For example suppose we need to estimate a 90^{th} percentile regression line through 1000 observations. In the case of two observations per interval, we need to divide the data into 100 intervals along the X variable and select the 2 highest Y observations from each interval. Using the ordinary least squares to these 200 data points, we can find the required 90^{th} percentile regression line.

6. Example

A real data set illustrates the procedures advanced in earlier sections. The Ohio Environmental Protection Agency has an ongoing program of evaluating the biological condition of streams in that state. The resulting data are available in a public database described by Yoder (1991). Details of the field protocols and allied matters are documented in a User's Guide available from the Ohio EPA (1987). The data used here relates to 245 stream sites in northeastern Ohio selected from that database as representing the small and intermediate sized streams in that region. At each stream site the response variable Y = species richness was evaluated; the regressor variable X is area in square miles drained by the stream above the sampling site. The data are nonlinear

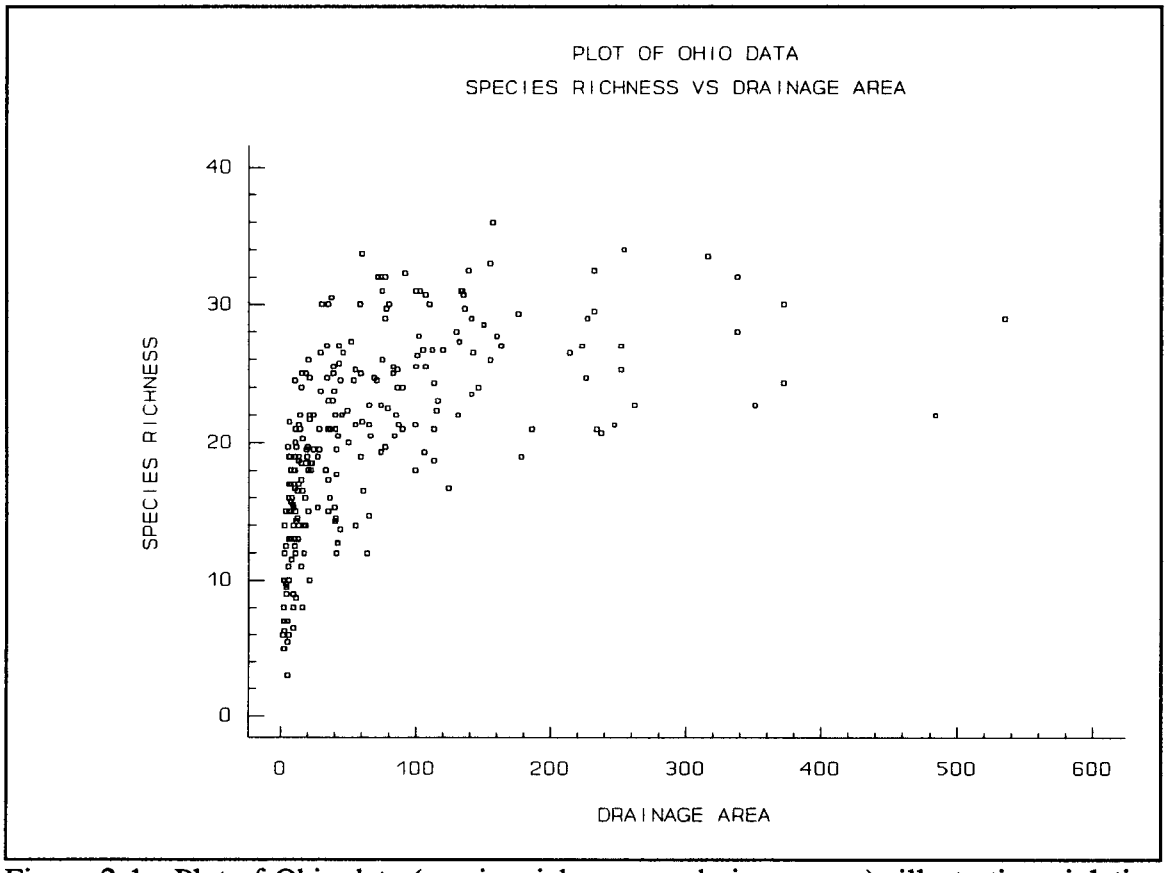

Figure 2.1: Plot of Ohio data (species richness vs. drainage area), illustrating violation of the assumption of linearity.

(see Figure 2.1). To check assumptions, we divided the data set into the same 25 subsets used for the nonparametric regression estimator. The lack of fit test for linearity, Draper and Smith (1981), demonstrated pronounced nonlinearity (F = , P $\approx$ 0.0000). The [3]log transformation for X variable makes the data linear much more nearly linear. See Figure 2.2; the lack of fit test no longer is significant (F = , P > 0.05). We also checked the model assumption that the standard deviation of Y is a linear function of variable X. We

[3]Common logarithm of base 10 is taken, but the process would work equally well if natural log were used throughout.

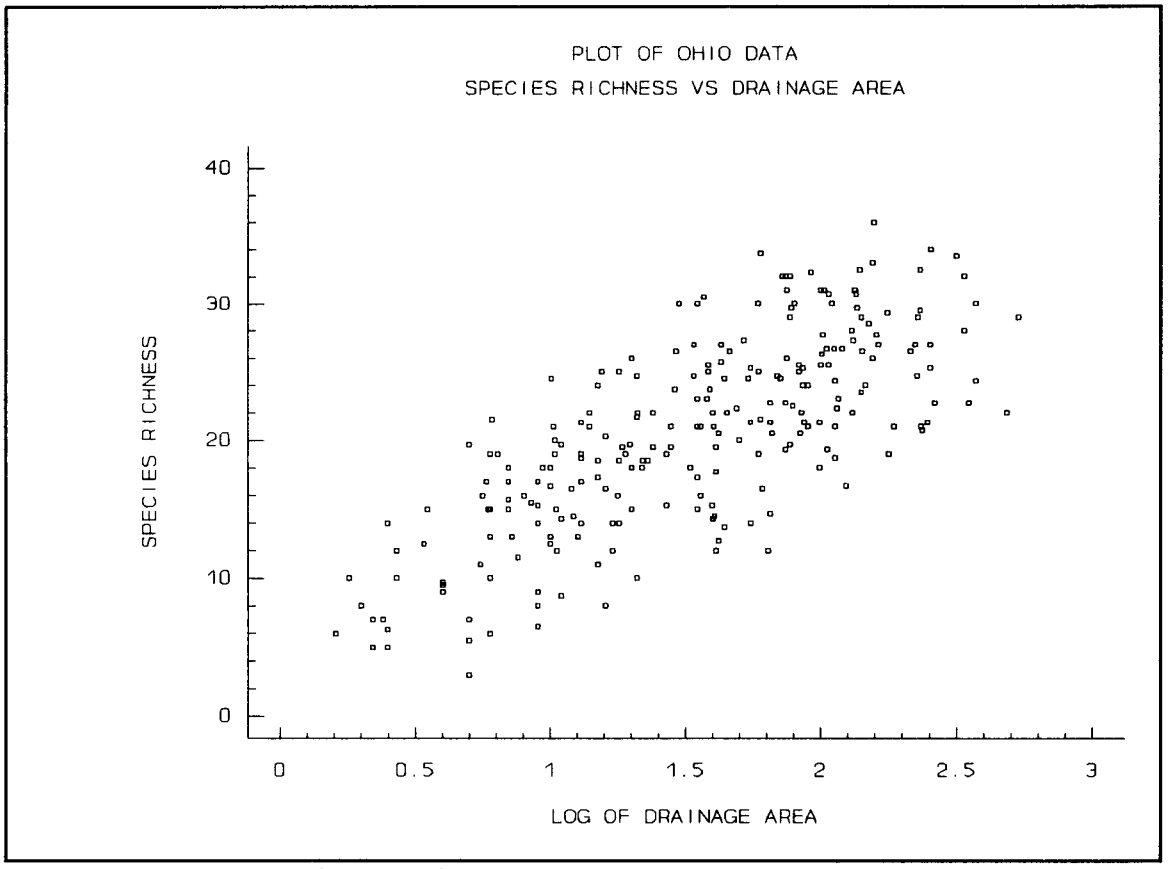

Figure 2.2: Plot of Ohio data after taking log transformation of "drainage area". The linearity assumption now appears reasonable.

evaluated the standard deviation of Y and mean of $\log X$ in each interval, and regressed the standard deviation on the means of $\log X$, giving estimates for the intercept and slope of 3.93 and 0.25, respectively. Because the plot of residuals from this least squares fit has no pattern, the standard deviation of Y approximates a linear function of X. (The values 3.93 and 0.25 can be used as starting values for the parameter α_2 and γ_2 respectively, in iterative procedure, discussed latter.)

Finally we examined the assumption of normality. A normal probability plot provides a tool for assessing the normality assumption. The normal probability plot of

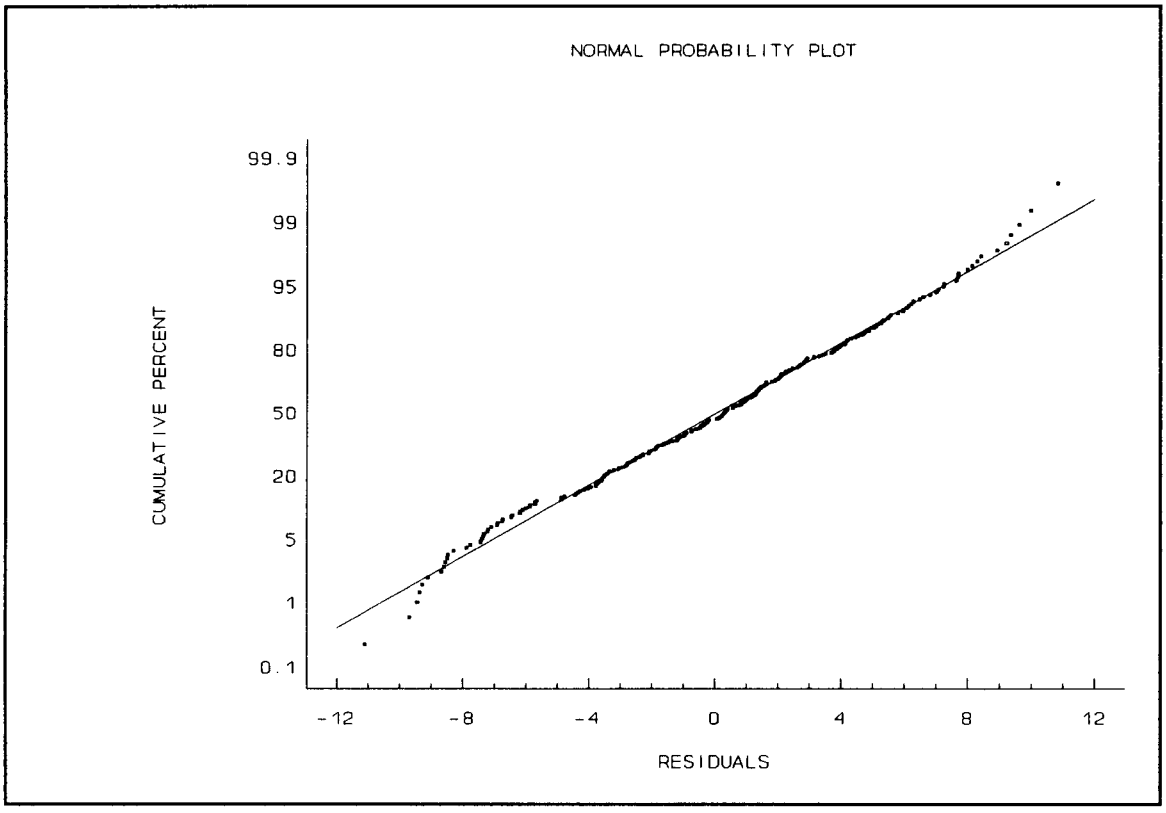

Figure 2.3: Normal probability plot of residuals from model of equation 2.4.

the standardized residuals for the Ohio data after log transformation is shown in Figure 2.3. It looks like a straight line, affirming that the assumption of normality is approximately satisfied. The residuals shown in Figure 2.3 are based on estimating all of the parameters in the model **(2.4)**,

$$\frac{(Y_i - \hat{\alpha}_1 - \hat{\gamma}_1 X_i)}{(\hat{\alpha}_2 + \hat{\gamma}_2 X_i)}$$

because previous checks have shown the need for unequal variances.

The linear model **(2.4)** provides a reasonable approximation to the appropriate model for the transformed data. We used both the Newton-Raphson iteration procedure

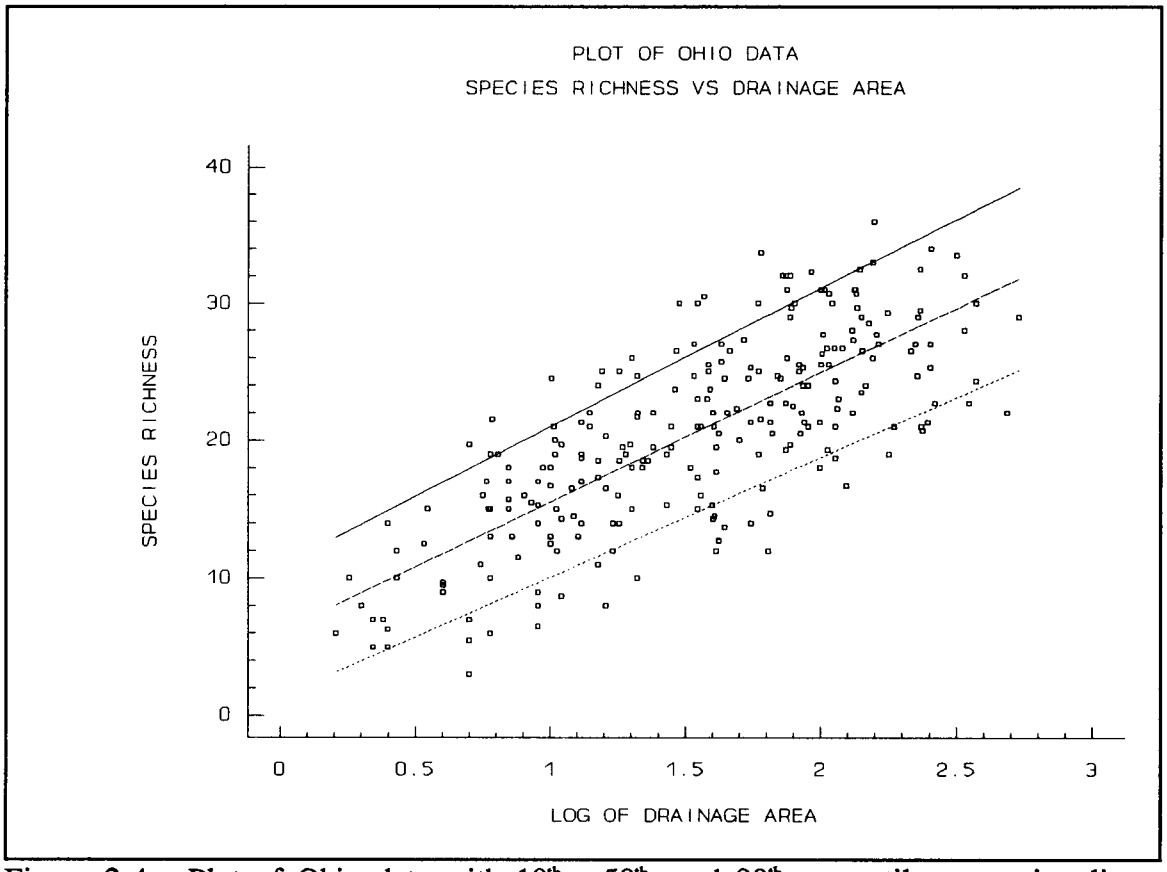

Figure 2.4: Plot of Ohio data with 10^{th} , 50^{th}, and 90^{th} percentile regression lines, obtained from maximum likelihood approach.

and profile likelihood for obtaining the maximum likelihood estimates. A good set of starting values often poses a difficult problem for any iteration procedure. Unfortunately no uniformly applicable rules exist for selecting good starting values, except that they should be as close to the final values as possible. For the present situation we obtained good initial values from the analyses above, values for α_1 and γ_1 from the least squares fit of Y on log X, and for α_2 and γ_2 as described above.

Using the model **(2.4)**, for the transformed data and applying the Newton-Raphson iterative procedure, the maximum likelihood estimates for the parameters are

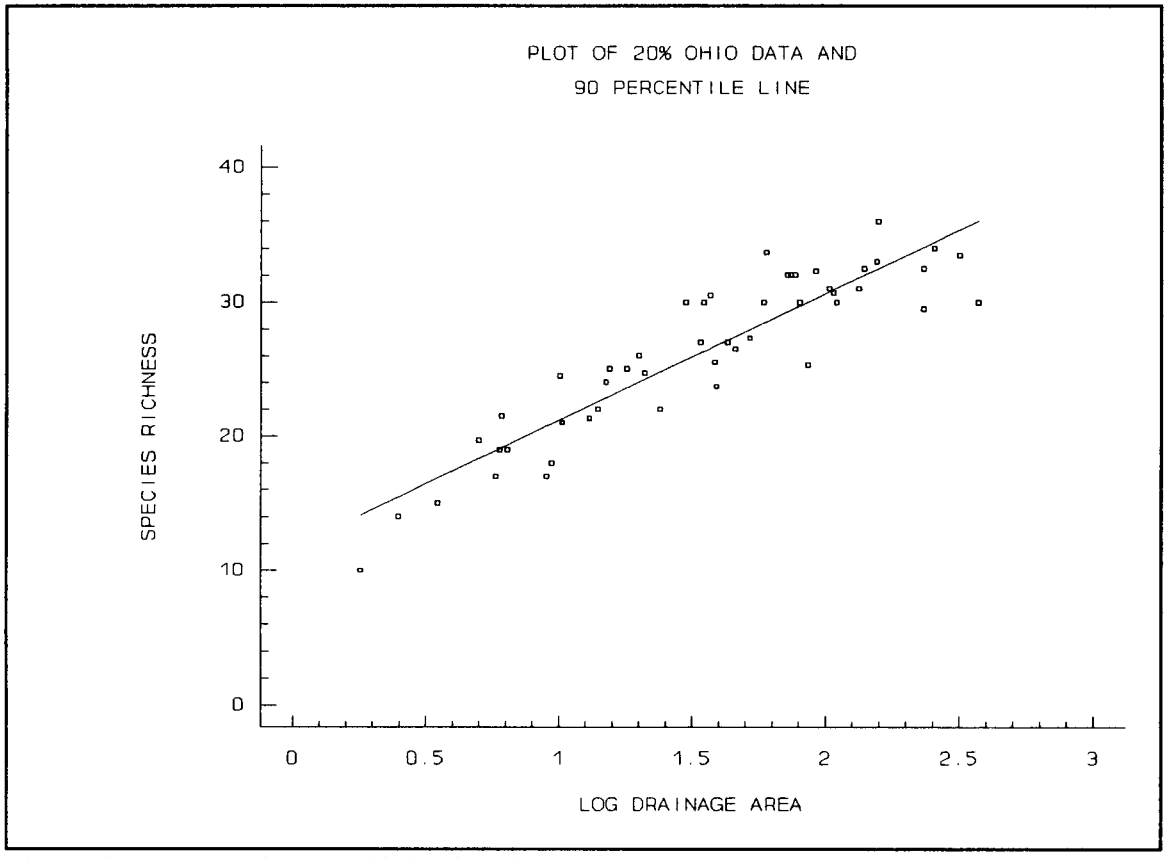

Figure 2.5: Plot of upper 20% of Ohio data, along with OLS line for those 20% data. This line gives nonparametric estimate of maximum species richness line.

$\hat{\alpha}_1 = 6.123$ (0.799) and $\hat{\gamma}_1 = 9.405$ (0.516),

$\hat{\alpha}_2 = 3.702$ (0.605) and $\hat{\gamma}_2 = 0.557$ (0.393),

where the values in the parentheses give the standard errors of the estimates. The parameter estimates were stable to 13 digits after only 7 iterations, using the Newton-Raphson method, but took 29 iterations to achieve the same computational precision for the iterative profile likelihood method. The plot of residuals against the fitted values of Y shows a random pattern around zero with no detectable trend.

Using the normality assumption and applying model **(2.4)** we found that

$$\hat{Y}_{90} = 10.869 + 10.119\,X\,,$$

$$\hat{Y}_{50} = 6.123 + 9.405\,X\,, \text{ and}$$

$$\hat{Y}_{10} = 1.377 + 8.691\,X\,.$$

These lines and the graph of the Ohio data set are presented in Figure 2.4. The upper line in Figure 2.4 is the 90th regression percentile, and, as such, estimates the maximum species richness line.

Next we illustrate nonparametric estimation of the 90th percentile regression line, using only the 20 % observations from Ohio data. The X_i's were arranged in ascending order and divided into 24 interval of 10 observations according to X variable, one interval, the center one, of 5 observations; the two highest Y observations were selected from each interval except one was taken from the middle one. The ordinary least squares procedure gave estimates for intercept and slope of 11.347 and 9.645 respectively and so the 90th percentile regression line is

$$\tilde{Y}_{90} = 11.347 + 9.675\,X\,.$$

The graph of these 49 observations and the 90th percentile regression line is shown in Figure 2.5. This is another estimate of the maximum species richness line. By comparing the two graphs, i.e., Figure 2.4 and Figure 2.5 we can see that these two approaches gave similar estimates of the maximum species richness line.

n	General Linear Model		Two Observations per Interval		One Observation per Interval	
	Intercept	Slope	Intercept	Slope	Intercept	Slope
200	35.0655 35.1043	1.7202 1.7145	35.1481 35.1126	1.7089 1.7096	34.3378 34.2938	1.6647 1.6662
500	35.1971 35.1781	1.7114 1.7118	35.2009 35.1694	1.7048 1.7067	34.3425 34.3070	1.6620 1.6645
700	35.2100 35.1875	1.7097 1.7133	35.1924 35.1490	1.7041 1.7091	34.3416 34.2935	1.6614 1.6663
1000	35.2693 35.2027	1.7074 1.7133	35.1969 35.1303	1.7053 1.7101	34.3378 34.2636	1.6623 1.6669
True Values	35.256	1.713	35.256	1.713	35.256	1.713

Table 2.1: Average estimates of the intercept and slope of the 90th percentile line for stated sample sizes and methods of estimation for 1000 and 2500 simulations.

7. Simulation

The following simulation investigates the suggested methods: maximum likelihood based on a linear model approach, and two variations on a nonparametric approach. In the nonparametric regression approach we actually considered three variations, but report here on only two: one observation per interval and two observations per interval; three observations per interval is not reported because it was not competitive.

The data sets were generated according to the model **(2.4)** with several sets of values for the constants α_1 , α_2, γ_1 and γ_2 and for sample sizes of n = 200, 250, 300, 350, . . . , 1000. We used GAUSS, a computer language adapted for executing and

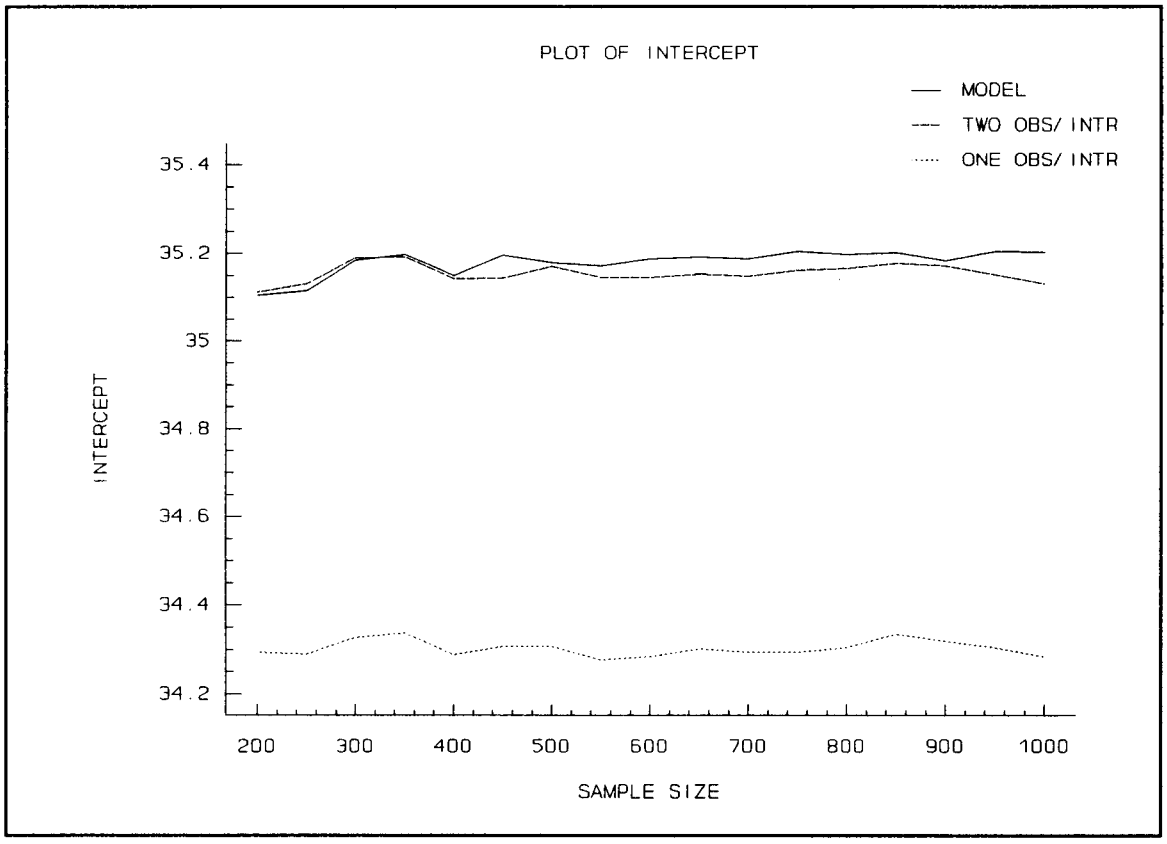

Figure 2.6: Average estimates of the intercept by three different approaches from 2500 simulations.

simulating statistical computation. Although we used several different sets of values for α_1, α_2, γ_1 and γ_2, we report only $\alpha_1 = 25.0$, $\alpha_2 = 8.0$, $\gamma_1 = 1.5$ and $\gamma_2 = 0.4$, because the basic results did not depend on the values used in the simulation; the patterns displayed in Tables 2.1 and 2.2 occurred for all sets of values of α_1, α_2, γ_1 and γ_2. We investigated five sizes of simulations for each sample size: 500, 1000, 1500, 2000 and 2500, but we report only 1000 and 2500 here. The X_i were taken as $X_i = (30 \times i)/n$ for $i = 1, 2, \ldots\ldots n$.

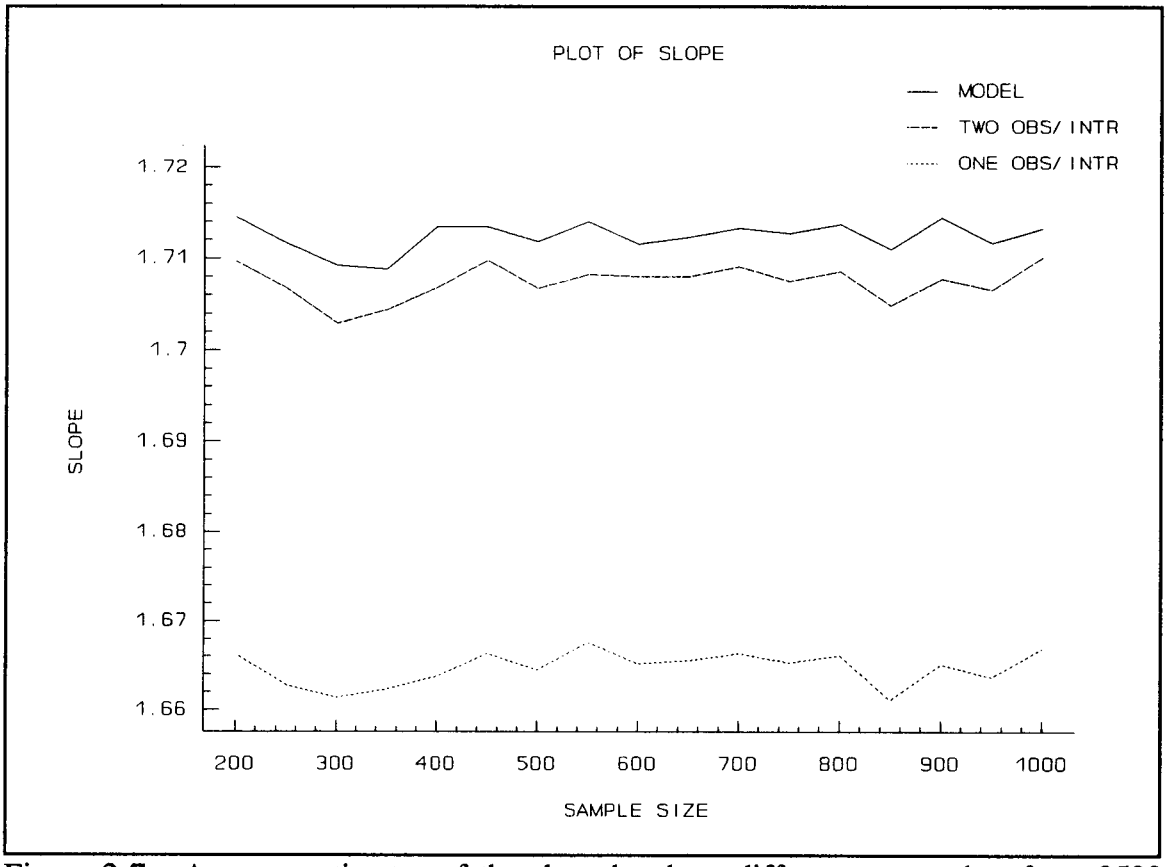

Figure 2.7: Average estimates of the slope by three different approaches from 2500 simulations.

Table 2.1 shows selected results for n = 200, 500, 700 and 1000. The upper value in each cell gives averages for the estimates of intercept and slope from 1000 simulations while the lower value gives averages for the estimates of intercept and slope from 2500 simulations. Table 2.2 gives the analogous standard errors for those estimates. These simulation results clearly show that the maximum likelihood approach gives good estimates of the parameters of the 90th percentile line, provided the data follow the assumed model (**2.4**). If the assumed model does not fit, the results would be worse; how bad would depend on the nature of the failure of the model. The

n	General Linear Model		Two Observations per Interval		One Observation per Interval	
	Intercept	Slope	Intercept	Slope	Intercept	Slope
200	1.9921 1.9542	0.1440 0.1435	2.4360 2.4355	0.1760 0.1762	2.4772 2.4350	0.1774 0.1751
500	1.2387 1.2448	0.0960 0.0937	1.5772 1.5282	0.1197 0.1136	1.5366 1.5301	0.1176 0.1136
700	1.0677 1.0737	0.0820 0.0788	1.3432 1.3350	0.1019 0.0972	1.3326 1.3213	0.1021 0.0965
1000	0.9024 0.8953	0.0674 0.0674	1.0961 1.1016	0.0811 0.0824	1.0935 1.0917	0.0806 0.0814

Table 2.2: Standard errors of intercept and slope of the 90th percentile line, for 1000 and 2500 simulations.

nonparametric method of selecting the largest two observations from sets of 10 observations with consecutive values of the predictor variable also gives consistent values with virtually no bias, where as selecting one observation from intervals of 5 gives biased results.

Figures 2.6 and 2.7 display results from which Tables 2.1 was extracted. The maximum likelihood approach and the nonparametric approach using two observations per interval display very similar average estimates, but the nonparametric approach using one observation per interval underestimates both the intercept and slope of the percentile regression line. Thus in terms of *biasedness* the former two methods clearly are superior to the latter.

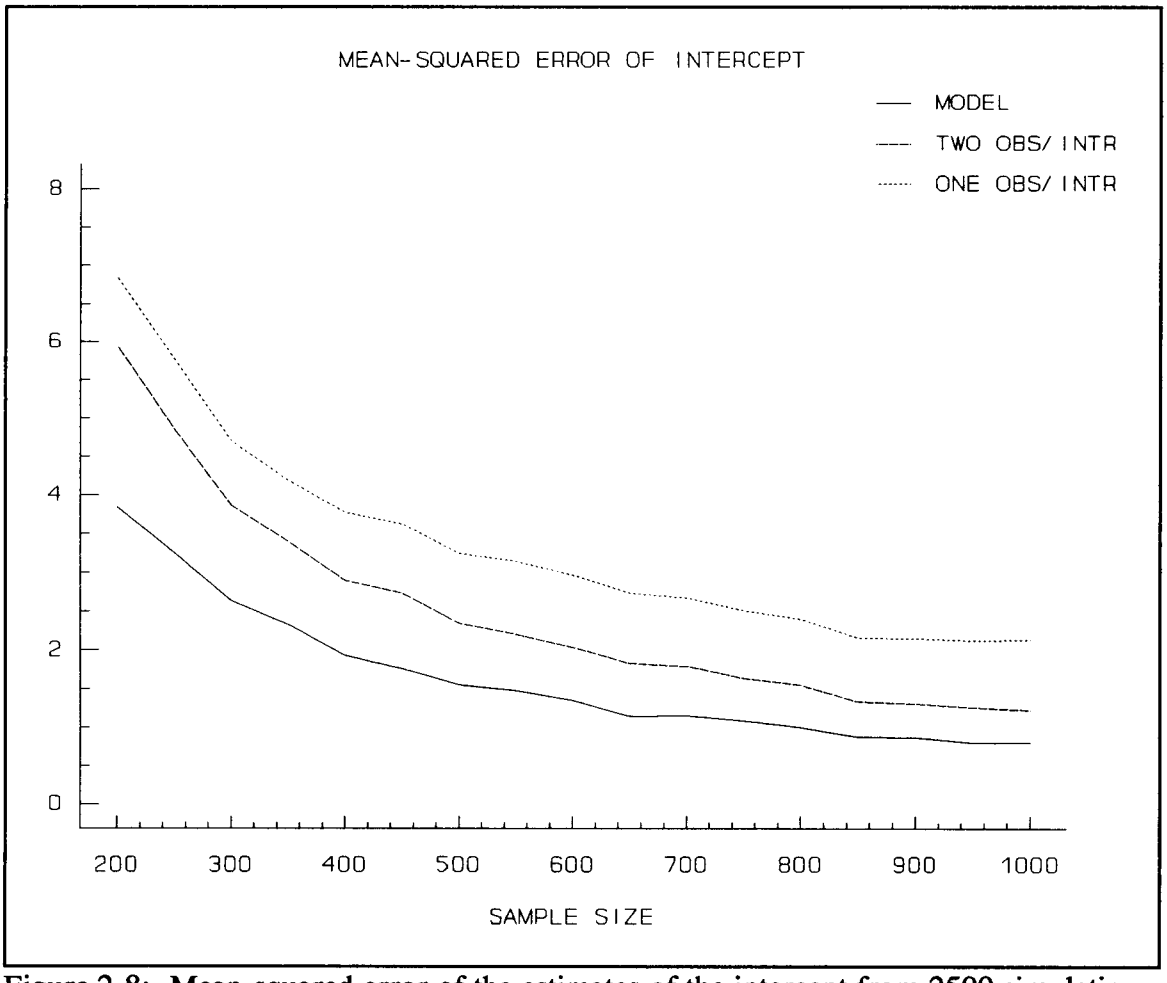

Figure 2.8: Mean-squared error of the estimates of the intercept from 2500 simulations.

We also investigated *relative efficiency* of the approaches discussed in previous paragraph. The two nonparametric methods have estimates with very similar standard errors, see Table 2.2, but both substantially exceed the standard errors of the maximum likelihood estimates. Although the nonparametric approach using two observations per interval and the maximum likelihood approaches have essentially no bias, the maximum likelihood approach clearly is the more efficient of the two.

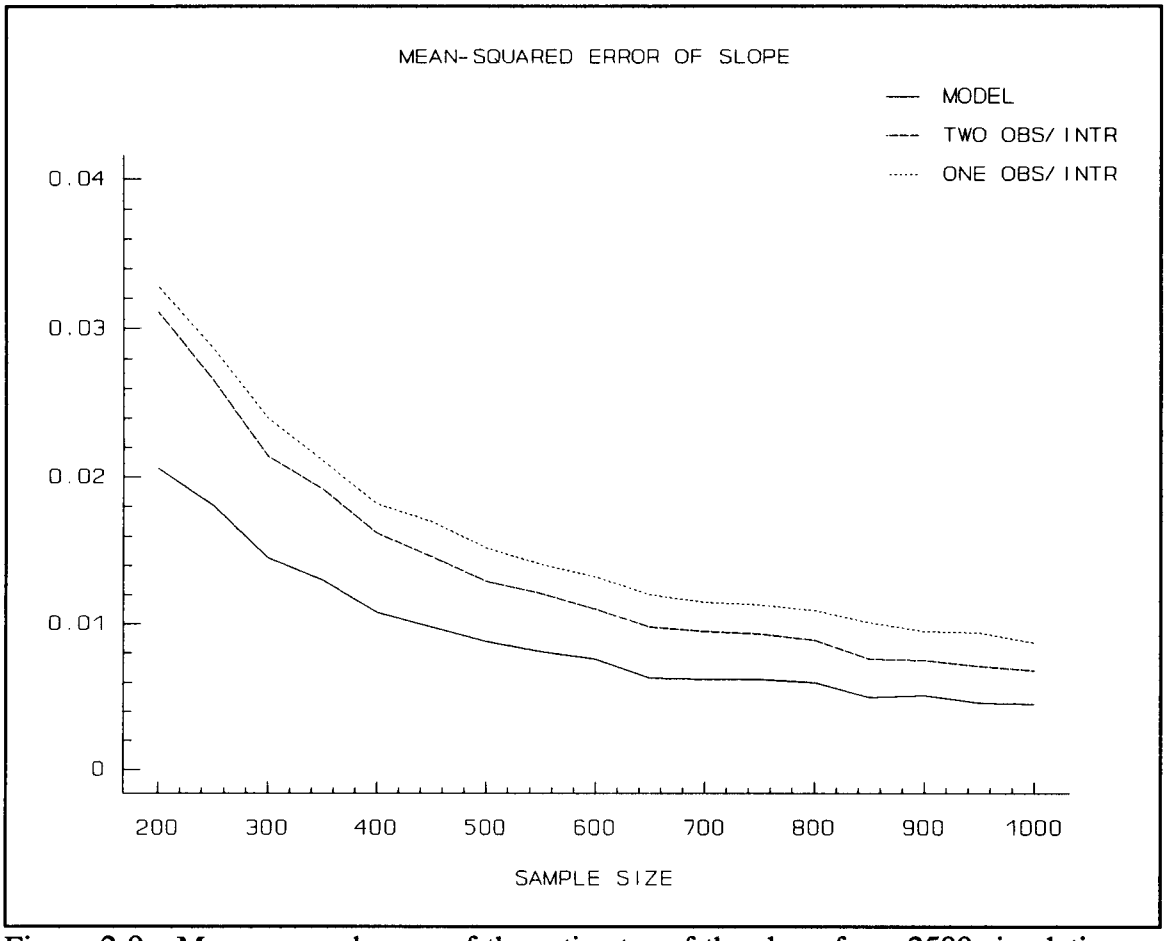

Figure 2.9: Mean-squared error of the estimates of the slope from 2500 simulations.

Figure 2.8 and 2.9 show the mean-squared errors of the estimates of the intercept and slope of the percentile regression line across the simulations. The mean-squared error for the maximum likelihood approach is the smallest, while one observation per interval has largest mean-squared error. We recommend the general linear model approach, provided the model on which it is based is reasonable.

Figure 2.10 and Tables 2.3 and 2.4 illustrate the kinds of situations under which the nonparametric approach is superior. This example has an essential feature: The data set consists of two almost distinct subsets. The upper one might correspond to relatively

n	General Linear Model		Two Observations per Interval	
	Intercept	Slope	Intercept	Slope
200	34.4724	1.6610	37.5651	1.7962
300	34.5082	1.6585	37.6265	1.7928
400	34.4901	1.6612	37.5710	1.7989
500	34.4652	1.6644	37.5155	1.8035
600	34.4969	1.6630	37.5491	1.8015
700	34.4960	1.6612	37.5116	1.8023
800	34.5099	1.6620	37.5511	1.8011
900	34.4698	1.6637	37.4932	1.8037
1000	34.5241	1.6616	37.5581	1.8021

Table 2.3: Average estimates of the intercept and slope of the 90th percentile line from mixture of two normal distributions.

uncontaminated sites while the other might reflect fairly contaminated sites, with no sites having intermediate gradations of contamination. The model used to generate this data was very similar to model (**2.4**); however the stochastic component came from a mixture of normals with different means, but the same variance, rather than from one normal. The "uncontaminated" subset of the data was simulated as a normally distributed random variable having a mean of $1.5(25 + 1.2X)$ while the "contaminated" subset had a mean of $0.875(25 + 1.2X)$; both subsets had a variance of $(4 + 0.2X)^2$. The first subset had probability of occurrence of 0.2 and while the second occurred with probability 0.8.

n	General Linear model		Two Observations per Interval	
	Intercept	Slope	Intercept	Slope
200	1.1307	0.0872	1.7943	0.1353
300	0.9364	0.0712	1.4450	0.1107
400	0.8049	0.0606	1.2812	0.0933
500	0.6988	0.0528	1.1019	0.0821
600	0.6542	0.0494	1.0258	0.0760
700	0.6130	0.0447	0.9631	0.0694
800	0.5874	0.0435	0.8832	0.0658
900	0.5209	0.0395	0.8081	0.0614
1000	0.5128	0.0391	0.7993	0.0600

Table 2.4: Standard error of the intercept and slope of the 90th percentile line from mixture of two normal distributions.

Given our definition of percentile regression, the 90% line should essentially be the mean line of the upper subset of the data, namely, 37.5 + 1.8*X*. Tables 2.3 and 2.4 present the same kind of information as Tables 2.1 and 2.2, but for this mixture simulation model. They show that the maximum likelihood estimate, using model (**2.4**) - admittedly an incorrect thing to do - somewhat underestimates the parameters while the nonparametric method gives estimates very close to the parameters. The maximum likelihood estimates still have smaller standard errors than the nonparametric method, but this has little relevance when it gives inconsistent estimates. This deficiency also is evidenced by the fact that about 16%, rather than 10% of the data points are above the

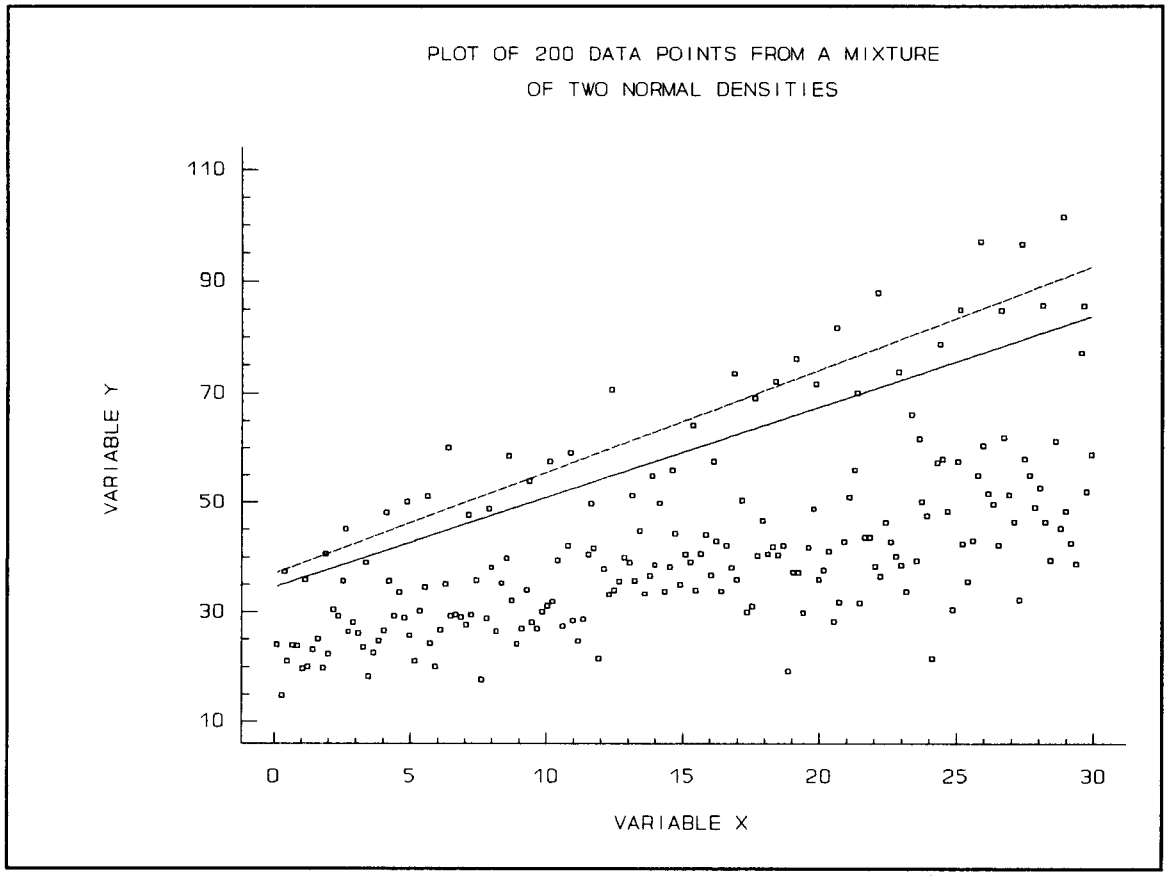

Figure 2.10: Data from a mixture of two normal densities with 90th percentile lines, maximum likelihood procedure (solid line) nonparametric (dashed line).

line estimated by maximum likelihood methods. We examined several cases to develop this example. The maximum likelihood estimates got progressively worse as the two subsets became more different, but the two subsets had to become almost distinct before the maximum likelihood estimates were very far off. This suggests that the maximum likelihood method may not be particularly sensitive to moderate failures of distributional assumptions.

Computationally, the maximum likelihood estimates can be obtained directly by using the Newton-Raphson methods (Section 3) and derivatives given in the appendix.

The profile likelihood method outlined in Section 4 gives the same estimates as the Newton-Raphson method because it is based on a reparametrization of the original model. On the other hand it is simpler to implement because it can be viewed as an adaptation of weighted regression, a relative familiar method. Another aspect of the simulation study showed that the profile likelihood method took about three times as long to produce estimates of the parameters of the percentile regression line as the direct approximation using the Newton-Raphson method. Unless very large or many data sets are involved, this difference in computing time probably can be ignored. Using a 486, 33 megahertz microcomputer, the Newton-Raphson method took 0.22 seconds on the Ohio data set, whereas the profile likelihood method took 0.77 seconds.

8. Conclusions

The results in sections 3, 4, 5 and 7 show that maximum likelihood estimation using a general linear model provides a suitable tool for estimating a percentile regression line, provided the model underlying it applies. The model assumes the response, like IBI, has an approximately normal distribution with heterogeneous variances. If the data are highly skewed so as to grossly violate the normality assumption, we need either a transformation of the data to a scale on which it is normally distributed or an alternative approach. A nonparametric approach to estimating the percentile regression line provides a needed alternative. It provides slightly less efficient estimates if the normal model holds, but performs well in the case of nonnormality. The Newton-Raphson method and

a profile likelihood method provide useful methods for calculating the maximum likelihood estimates.

ACKNOWLEDGMENTS

The use of iteratively reweighted least squares in the profile likelihood method resulted from discussions with David R. Thomas in the same department as the authors. The work of the senior author was supported by the Pakistan Participant Training Program funded by US AID. The work of the second author was supported in part by Cooperative Agreement CR 816721 between the Department of Statistics at Oregon State University and the Environmental Protection Agency.

REFERENCES

Angers, C. (1979). Simultaneous Estimation of Percentile Curves with Application to Salary Data. *Journal of the American Statistical Association*, 74, 621-625.

Apteck Systems (1992). *The GAUSS System Version 3.0.* Aptech Systems, Inc. 23804 S. E. Kent-Kangley Road, Maple Valley, Washington 98038.

Carroll, R. J., and D. Rupert, (1982). Robust Estimation in Heteroscedastic Linear Models. *The Annals of Statistics*, 10, 429-441.

Davidian, M., and R. J. Carroll, (1987). Variance Function Estimation. *Journal of the American Statistical Association,* 82, 1079-1091.

Draper, N. R., and H. Smith, (1981). *Applied Regression Analysis.* John Wiley & Sons, New York.

Fausch, K. D., J. R. Karr, and P. R. Yant, (1984). Regional application of an index of biotic integrity based on stream-fish communities. *Transactions of the American Fisheries Society*, 113, 39-55.

Griffiths, D., and M. Willcox, (1978). Percentile Regression: A Parametric Approach. *Journal of the American Statistical Association*, 73, 496-498.

Hogg, R. V. (1975). Estimates of Percentile Regression Lines Using Salary Data. *Journal of the American Statistical Association*, 70, 56-59.

Karr, J. R. (1981) Assessment of biotic integrity using fish communities. *Fisheries*, 6, 21-27.

Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant, and I. J. Schlosser, (1986). Assessing biological integrity in running waters, A method and its rational. *Illinois Natural History Survey Special Publication* 5.

Kennedy, W. J. Jr., and J. E. Gentle, (1980). *Statistical Computing*. Marcel Dekker, Inc. New York.

McCullagh, P., and J. A. Nelder, (1989). *Generalized Linear Models*, (2nd Ed.) London: Chapman and Hall.

Ohio Environmental Protection Agency (1987). Biological Criteria for the Protection of Aquatic Life: Volume II. *Users Manual for Biological Field assessment of Ohio Surface Waters*. Division of Water Quality Monitoring and Assessment, Surface Waters Section, Columbus, Ohio.

Yoder, C. O. (1991). The integrated Biosurvey as a tool for the evaluation of Aquatic life use attainment and impairment in Ohio surface waters. Biological criteria: Research and Regulation. *Proceeding of a National Conference*, U. S. EPA, Office of Water, Washington, D. C.

APPENDIX

APPENDIX

For the normal model used in section 2 the log likelihood function given as **(2.6)** is

$$L = -\sum_{i=1}^{n} \ln (\alpha_2 + \gamma_2 X_i) - .5 \sum_{i=1}^{n} \left(\frac{Y_i - \alpha_1 - \gamma_1 X_i}{\alpha_2 + \gamma_2 X_i} \right)^2$$

The first and second derivatives with respect to the four parameters are

$$\frac{\partial L}{\partial \alpha_1} = l_1 = \sum_{i=1}^{n} \left[\frac{Y_i - \alpha_1 - \gamma_1 X_i}{(\alpha_2 + \gamma_2 X_i)^2} \right]$$

$$\frac{\partial L}{\partial \gamma_1} = l_2 = \sum_{i=1}^{n} \left[\frac{X_i (Y_i - \alpha_1 - \gamma_1 X_i)}{(\alpha_2 + \gamma_2 X_i)^2} \right]$$

$$\frac{\partial L}{\partial \alpha_2} = l_3 = -\sum_{i=1}^{n} \left[\frac{1}{\alpha_2 + \gamma_2 X_i} \right] + \sum_{i=1}^{n} \left[\frac{(Y_i - \alpha_1 - \gamma_1 X_i)^2}{(\alpha_2 + \gamma_2 X_i)^3} \right]$$

$$\frac{\partial L}{\partial \gamma_2} = l_4 = -\sum_{i=1}^{n} \left[\frac{X_i}{\alpha_2 + \gamma_2 X_i} \right] + \sum_{i=1}^{n} \left[\frac{X_i(Y_i - \alpha_1 - \gamma_1 X_i)^2}{(\alpha_2 + \gamma_2 X_i)^3} \right]$$

$$\frac{\partial^2 L}{\partial \alpha_1^2} = [\ddot{l}_{11}] = -\sum_{i=1}^{n} \left[\frac{1}{(\alpha_2 + \gamma_2 X_i)^2} \right]$$

$$\frac{\partial^2 L}{\partial \alpha_1 \partial \gamma_1} = [\ddot{l}_{12}] - \sum_{i=1}^{n} \left[\frac{X_i}{(\alpha_2 + \gamma_2 X_i)^2} \right]$$

$$\frac{\partial^2 L}{\partial \alpha_1 \partial \alpha_2} = \left[\ddot{l}_{13}\right] = -2\sum_{i=1}^{n}\left[\frac{Y_i - \alpha_1 - \gamma_1 X_i}{(\alpha_2 + \gamma_2 X_i)^3}\right]$$

$$\frac{\partial^2 L}{\partial \alpha_1 \partial \gamma_2} = \left[\ddot{l}_{14}\right] = -2\sum_{i=1}^{n}\left[\frac{X_i(Y_i - \alpha_1 - \gamma_1 X_i)}{(\alpha_2 + \gamma_2 X_i)^3}\right]$$

$$\frac{\partial^2 L}{\partial \gamma_1^2} = \left[\ddot{l}_{22}\right] = -\sum_{i=1}^{n}\left[\frac{X_i^2}{(\alpha_2 + \gamma_2 X_i)^2}\right]$$

$$\frac{\partial^2 L}{\partial \gamma_1 \partial \alpha_2} = \left[\ddot{l}_{23}\right] = -2\sum_{i=1}^{n}\left[\frac{X_i\,(Y_i - \alpha_1 - \gamma_1 X_i}{(\alpha_2 + \gamma_2)^3}\right]$$

$$\frac{\partial^2 L}{\partial \gamma_1 \partial \gamma_2} = \left[\ddot{l}_{24}\right] = -2\sum_{i=1}^{n}\left[\frac{X_i^2\,(Y_i - \alpha_1 - \gamma_1 X_i)}{(\alpha_2 + \gamma_2 X_i)^3}\right]$$

$$\frac{\partial^2 L}{\partial \alpha_2^2} = \left[\ddot{l}_{33}\right] = \sum_{i=1}^{n}\left[\frac{1}{(\alpha_2 + \gamma_2 X_i)^2}\right] - 3\sum_{i=1}^{n}\left[\frac{(Y_i - \alpha_1 - \gamma_1 X_i)^2}{(\alpha_2 + \gamma_2 X_i)^4}\right]$$

$$\frac{\partial^2 L}{\partial \alpha_2 \partial \gamma_2} = \left[\ddot{l}_{34}\right] = \sum_{i=1}^{n}\left[\frac{X_i}{(\alpha_2 + \gamma_2 X_i)^2}\right] - 3\sum_{i=1}^{n}\left[\frac{X_i\,(Y_i - \alpha_1 - \gamma_1 X_i)^2}{(\alpha_2 + \gamma_2 X_i)^4}\right]$$

$$\frac{\partial^2 L}{\partial \gamma_2^2} = \left[\ddot{l}_{44}\right] = \sum_{i=1}^{n}\left[\frac{X_i^2}{(\alpha_2 + \gamma_2 X_i)^2}\right] - 3\sum_{i=1}^{n}\left[\frac{X_i^2\,(Y_i - \alpha_1 - \gamma_1 X_i)^2}{(\alpha_2 + \gamma_2 X_i)^4}\right]$$

The 4 by 1 vector $\mathbf{u}(\theta)$ of the first partial derivatives of the log likelihood function, known as the score vector, is

$$u(\theta) = \left[\dot{l}_1, \dot{l}_2, \dot{l}_3, \dot{l}_4\right]'$$

The observed information matrix is the 4 by 4 symmetric matrix obtained from the second partial derivatives of the log likelihood function, evaluated at the observed estimates of the four parameters:

$$G(\theta) = -\begin{bmatrix} \ddot{l}_{11} & \ddot{l}_{12} & \ddot{l}_{13} & \ddot{l}_{14} \\ . & \ddot{l}_{22} & \ddot{l}_{23} & \ddot{l}_{24} \\ . & . & \ddot{l}_{33} & \ddot{l}_{34} \\ . & . & . & \ddot{l}_{44} \end{bmatrix}$$

Suppose that θ_0 is the first guess at $\hat{\theta}$ then expanding $\mathbf{U}(\hat{\theta})$ as a taylor series about θ_0 is

$$u(\hat{\theta}) = u(\theta_0) + G(\theta_0)(\hat{\theta} - \theta_0)$$

The maximum likelihood estimators $\hat{\theta}$ ordinarily satisfy the equation $\mathbf{u}(\hat{\theta}) = 0$, yielding the approximation

$$\hat{\theta} = \theta_0 - G(\theta_0)^{-1} u(\theta_0)$$

In practice the above equation is used to define an iteration scheme for obtaining $\hat{\theta}$. On the first iteration the above equation produces a second approximation θ_1 to $\hat{\theta}$. This second approximation is then inserted in the right hand side of the above equation to produce a third approximation and so on until convergence occurs.

CHAPTER III

ADAPTING ROBUST REGRESSION TO PERCENTILE REGRESSION FOR ESTIMATING THE MAXIMUM SPECIES RICHNESS LINE.

by

Mohammad F. Qadir[1] and N. Scott Urquhart[2]

1. Abstract

A new method of estimating the maximum species richness line for the index of Biotic Integrity is proposed here. In the case of an approximately normal model with unequal variances which increase as the independent variable increases, a method of analysis is required that is insensitive to misspecification of the distribution and/or to possible outliers. We propose an adaptation of robust regression to estimate percentile regression lines. This robust method uses weighted least squares, where the weights are calculated from a beta function. It offers the user of the maximum species richness line a robust alternative to the methods proposed by Fausch et al. (1984) and Qadir and Urquhart (1993). We compare this approach with the two other approaches from Qadir and Urquhart (1993). Simulated data sets containing heteroscedasticity are used to compare the approaches; the basis of comparison is the mean-squared error. The maximum likelihood procedure based on a linear model dominates both nonparametric and robust procedures, when the

---

[1] Department of Statistics, University of Peshawar, N.W.F.P., Pakistan; formerly a graduate student in the Department of Statistics, Oregon State University, Corvallis, OR 97331-4606, USA.

[2] Department of Statistics, Oregon State University, Corvallis, OR 97331-4606, USA.

assumptions of the model are satisfactory. Otherwise the robust performs well; it needs to be explored further.

2. Introduction

The Index of Biotic Integrity (IBI) was proposed by Karr (1981), Fausch et al. (1984), Karr et al. (1986) as an index to the biological condition of streams. This index utilizes several attributes of fish populations in a water body. A sample from a site of interest is compared to a sample drawn in the same way at an excellent (undisturbed condition) and equivalent regional site, known as the reference site for that area. The number of species of a type, often called species richness, forms an integral part of the IBI, but this response generally increases with the size of the water body. Regional reference sites often have to be identified from the same data sets on which IBI will be evaluated. This leads to the need to identify a line or curve which gives the excellent species richness for sites as a function of stream size or similar features of the site. A line above which a specified percentage of the points, such as 5% or 10%, can serve this purpose.

Plotting total score at a site versus the water body size produces a fan shape of points whose upper bound forms an upper bound for the species richness. Such fan-shaped plots reflect the nature of sampling stream fish communities. Then a line known as maximum species richness line Fausch et al. (1984), Karr et al. (1986) with slope "fit by eye" forms the upper bound for about 90 or 95% of the sites. The sites whose score falls above the line are considered as "excellent" fish communities. Qadir and Urquhart (1993) proposed two methods for estimating the maximum species richness line. Their

methods, maximum likelihood procedure using a linear model and nonparametric procedure, are a first attempt at presenting a statistically related method for estimating the maximum species richness line.

A different method is proposed here; it is based on the weighted least squares estimation procedure. First an ordinary least squares model is fit to the data and the residuals are obtained from that fit. Using the standardized residuals, the weight function is determined for each observation and then the weights are used to find the weighted least squares estimate for the maximum species richness line. This method of estimation is insensitive to outliers and also can even be applied to outlier detection.

The paper is organized as follows: Section 3 gives an overview of the robust regression, discussing M-estimators, W-estimators, and the weighted least squares approach outlined above. This Section also contains two examples showing and comparing our procedure with some other robust estimation procedures. These two numerical examples show the accuracy of our weighted least squares estimator compared to various robust methods. In section 4 we adapt the robust method for estimating the maximum species richness line. Section 5 presents simulation studies of the weighted least squares procedure. The conclusions and summary are presented in section 6.

3. Development of an Alternative

Ordinary Least Square (OLS) has dominated the regression techniques for more than a century. OLS works well on a "nice" data set which satisfies the assumptions of constant variance and uncorrelated of observations; in the presence of normality, OLS

produces estimates with additional desirable properties. Robust regression has emerged as an alternative to OLS, Birkes and Dodge (1993). These procedures still seek estimates which produce lines or planes which, in a sense, pass through the "center" of the data, as does OLS. We are headed toward a further adaptation which passes a line through the data, but near an edge of the data cloud.

A general linear model can be represented by

$$y = X\beta + \epsilon \quad ,$$

where y is a vector of observed responses, X is a $n \times p$ matrix of p explanatory variables, β is a vector of p unknown parameters and ϵ is a vector of n random variables. The OLS estimates of β are

$$\hat{\beta} = (X'X)^{-1}X'y \quad . \tag{3.2}$$

If $\text{cov}(\epsilon) = \mathbf{V} \neq \sigma^2 \mathbf{I}$ and $\mathbf{V}^{-1} = \mathbf{W}$, where $\mathbf{W}$ is a diagonal matrix,

$$W = \begin{bmatrix} w_1 & 0 & & 0 \\ 0 & w_2 & & 0 \\ & & \cdot & \\ & & & \cdot \\ 0 & 0 & & w_n \end{bmatrix} \quad , \tag{3.3}$$

then

$$\hat{\beta} = (X'WX)^{-1}X'Wy \quad . \tag{3.4}$$

The above generalizes OLS to weighted regression, a point to which we will return.

When a data set follows the model being used for analysis, statistically valid outcomes ordinarily result. If, however, a data set contains an occasional observation not following that model, results founded on the model may be seriously compromised. Outliers in regression illustrate such a problem for an OLS analysis. Two approaches for dealing with this problem lie in regression diagnostics and robust regression. Diagnostics procedures attempt to identify the influential observations and possible outliers so they can be removed from the data and analysis conducted on the reduced data set to produce somewhat more reliable results. Such procedures may involve extensive computation and subjective judgements at several stages. It may be difficult to find all outliers if there are many. On the other hand robust regression procedures are designed to be insensitive to outliers see Huber (1981) and Hoaglin *et al* (1983).

Within the context of robust regression, the M-estimate for location using the function ρ and the sample $x_1, x_2, \ldots, x_n$ is the value of t that minimizes the objective function:

$$\sum_{i=1}^{n} \rho(x_i; t) \quad . \tag{3.5}$$

The most familiar M-estimate is the sample mean, the least squares estimate of location. For least squares estimation, ρ is the square of the residuals

$$\rho(x: t) = (x - t)^2 \quad . \tag{3.6}$$

More generally, the M-estimate for the slope in simple linear regression is the value of β which minimizes

$$\sum_{i=1}^{n} \rho(y_i - x_i\beta) \ , \tag{3.7}$$

where ρ is a convex function symmetric around zero. If ρ is continuous and differentiable, as we ordinarily expect, and we let

$$\psi(t) = \rho'(t) \ . \tag{3.8}$$

Often it is more convenient to estimate $\hat{\beta}$ by finding the value of β satisfying

$$\sum_{i=1}^{n} \psi(y_i - x_i\beta) = 0 \ . \tag{3.9}$$

From Ψ-function we can find the ω-function with the relation

$$\omega(t) = \frac{\psi(t)}{t} \ , \tag{3.10}$$

where the ω-function is a weight function which can be used in weighted regression as in Eq. **3.3** where the diagonal elements of W matrix are the corresponding weights on the response variable. Then β has weighted least squares estimates given by Eq. **3.4** which also is the maximum likelihood estimate when the residuals are normally distributed.

Now consider a new objective function

$$\rho(r) = \begin{cases} \dfrac{r^2}{96a^4}\,(3a^4 - 3a^2r^2 + r^4) & \text{if} \quad |r| \le a \\ \dfrac{a^2}{96} & \text{if} \quad |r| > a \end{cases} \tag{3.11}$$

where r is the OLS residual divided by the standard deviation of the observed residuals and a is the tuning constant. The above objective function has all the *nice* properties, namely,

a) $\rho(0) = 0$

b) $\rho(-r) = \rho(r)$

c) for $0 < r_1 < r_2 \Rightarrow \rho(r_1) \leq \rho(r_2)$

d) ρ is continuous

e) let $a = \sup \rho(r)$, then $0 < a < \infty$

f) if $\rho(r_1) < a$ and $0 < r_1 < r_2$, then $\Rightarrow \rho(r_1) < \rho(r_2)$

The derivative with respect to r of the above objective function is

$$\psi(r) = \begin{cases} \dfrac{r}{16a^4}(a+r)^2(a-r)^2 & \text{if} \quad |r| \leq a \\ 0 & \text{if} \quad |r| > a \ . \end{cases} \tag{3.12}$$

Like the Andrews M-estimator, Andrews (1974), and the biweight estimator, this Ψ function belongs to the class known as redescending Ψ functions because the Ψ function comes back to zero when the absolute value of the argument is greater than a specified positive number, in this case the tuning constant a.

The weight function corresponding to this Ψ function is

$$\omega(r) = \begin{cases} \dfrac{1}{16a^4}(a+r)^2(a-r)^2 & \text{if} \quad |r| \leq a \\ 0 & \text{if} \quad |r| > a \ . \end{cases} \tag{3.13}$$

If we let $r = 2at - a$ the above weight function can be rewritten as

$$\omega(t) = \begin{cases} t^2(1-t)^2 & \text{if} \quad 0 \le t \le 1 \\ 0 & \text{otherwise} \end{cases} . \tag{3.14}$$

This is a beta function with parameters $\alpha = 3$ and $\beta = 3$. Beta function becomes symmetric when $\alpha = \beta$. Using a beta function as a weight function actually down-weighs outliers and gives more weight to the central observations. Use of this sort of weight function is similar to the use of trimmed-means and winsorized-means, robust estimates of a mean which ignore or give smaller weight to the extreme observations.

The above motivates a more general weight function and objective function. Consider a general weight function of the form

$$\omega(t) = \begin{cases} t^{\alpha-1}(1-t)^{\beta-1} & \text{if} \quad 0 \le t \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

where the parameters α and β are any real numbers greater than 0.

When $\alpha = \beta = 1$, this function gives equal weight to all observations, thereby producing ordinary least square estimates from robust regression. Increasing α and β implies more trimming or winsorizing of extreme observations on both sides of the distribution. When α is substantially larger than β, this function places its weight on the observations above the mean. Suitable choices of a, α and β will produce estimates of percentile regression lines.

Although we are not trying to advance the weight function of Eq. **3.14** for use in robust regression, we examined its use there as a way to check on its default performance. We applied it to two well known data sets, getting results very similar to

those from other robust regression studies of those data sets. The sets were the number of phone calls from Belgium Rousseeuw and Leroy (1987), and the *stackloss* data presented by Brownlee (1965).

Example 3.1.

The first example is taken from Rousseeuw and Leroy (1987). This is a real data set with a few outliers present in the data. The data set is taken from Belgian Statistical Survey. The dependent variable y is the annual number of international Phone calls made from Belgium and the independent variable x is the year.

The plot of the data is shown in Figure 3.1 with least squares fit, least median of square (LMS) fit proposed by Rousseeuw (1984), and the fit by our method. From the plot it is clear that the observations from 1964 to 1969 are outliers; recording system contaminated the data. The least squares fit

$$\hat{y} = -26.01 + 0.504\,x$$

is highly influenced by the outliers, and thus fits neither the good nor bad data points well. The LMS fit is

$$\hat{y} = -5.61 + 0.115\,x\,;$$

this effectively ignores the outliers. The fit from our method of weighted least squares,

$$\hat{y} = -5.505 + 0.115\,x\,,$$

is similar to the LMS fit.

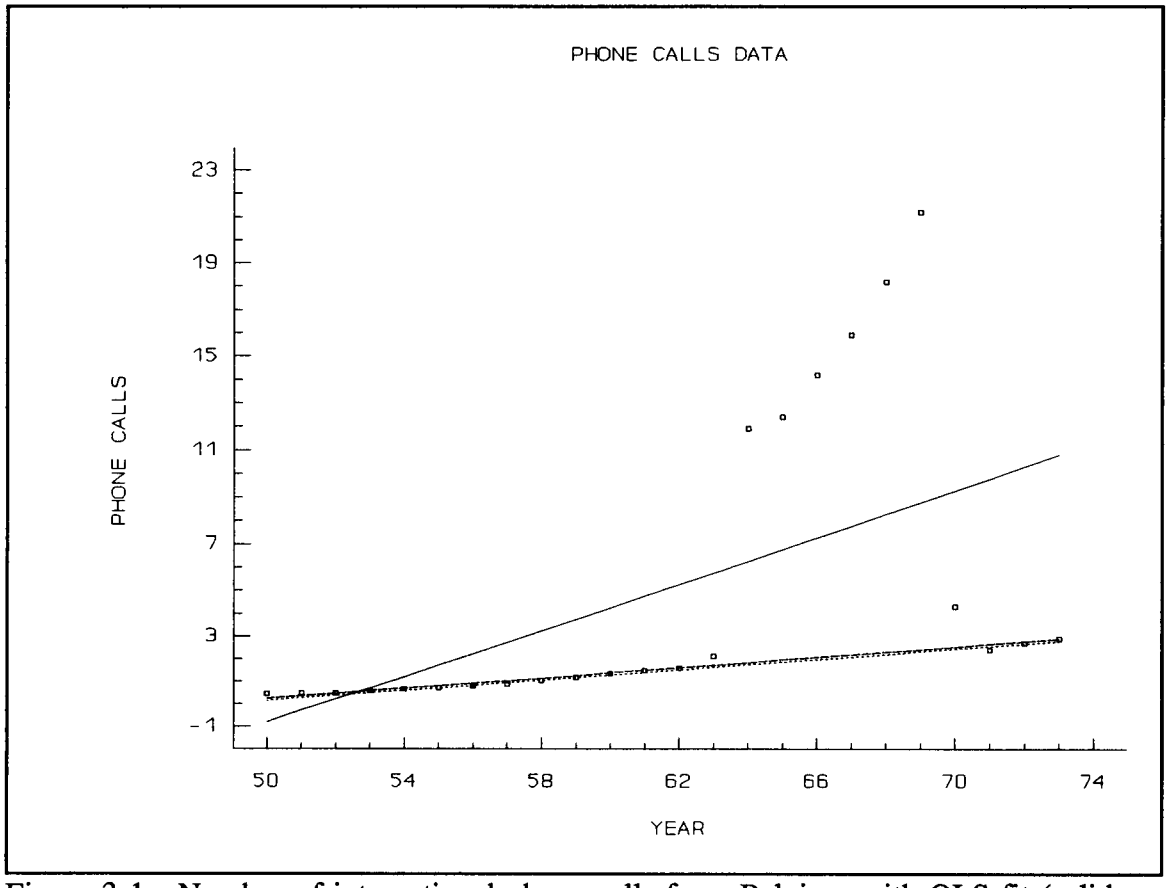

Figure 3.1: Number of international phone calls from Belgium with OLS fit (solid line) WLS fit (dashed line) and LMS fit (dotted line).

Example 3.2.

The second example applies multiple regression problem to the famous stackloss data set presented by Brownlee (1965). This real data set describes the operation of a plant for the oxidation of ammonia to nitric acid. This data set has 21 four-dimensional observations, the stackloss (y) has to be explained by the rate of operation (x_1), the cooling water inlet temperature (x_2), and the acid concentration (x_3). This data set has been examined by a number of statisticians (Daniel and Wood 1971, Andrews 1974, Andrews and Pregibon 1978, Cook 1979, Dempster and Gasko-Green 1981, Draper and

smith 1981, Atkinson 1982, Atkinson 1985, Carroll and Ruppert 1985, Li 1985, Rousseeuw and Leroy 1987, Birkes and Dodge 1993, and many others) in studies of robust regression. Atkinson (1985) gives a summary of analyses to the end of 1981.

Least squares fit for all the 21 observations is

$$\hat{y} = -39.920 + 0.716\,x_1 + 1.295\,x_2 - 0.152\,x_3 .$$

Most of the statisticians concluded that observations number 1, 3, 4, and 21 are outliers. The ordinary least squares fit for the remaining 17 observations is

$$\hat{y} = -37.652 + 0.798\,x_1 + 0.577\,x_2 - 0.067\,x_3 .$$

Andrews (1974) applied a robust M-estimator to the stackloss data. His Ψ function was

$$\psi(t) = \begin{cases} \sin\left(\dfrac{t}{1.5}\right) & \text{for} \quad |t| \le 1.5\pi \\ 0 & \text{for} \quad |t| > 1.5\pi \end{cases} \tag{3.16}$$

with scale parameter estimated by $\hat{\sigma}$ = median $\{ \mid y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}} \mid \}$. His robust fit for all 21 observations was

$$\hat{y} = -37.200 + 0.820\,x_1 + 0.520\,x_2 - 0.070\,x_3 .$$

This is quite close to the 17-point ordinary least squares fit, and automatically ignores the outliers. The fit for all 21 observations using the weighted least squares procedure and $\omega(t)$ from Eq. **3.14** with the tuning constant value for $a = 1.5$, gave

$$\hat{y} = -37.537 + 0.726\,x_1 + 0.923\,x_2 - 0.103\,x_3 .$$

The residuals from the OLS fit, the 17-point least squares fit, from Andrews robust fit, and from our weighted least squares fit are given in Table 3.1. The residuals from the

Observations	OLS Residuals	17-point Residuals	Robust Residuals	Weighted Least Square Residuals
1	3.235	6.218	5.970	5.671
2	-1.917	1.151	0.720	0.568
3	4.556	6.428	6.000	6.253
4	4.698	8.174	7.970	7.309
5	-1.712	-0.671	-0.990	-0.844
6	-3.007	-1.249	-1.510	-1.768
7	-2.389	-0.424	-0.610	-1.073
8	-1.389	0.576	0.390	-0.073
9	-3.144	-1.058	-1.230	-1.863
10	1.267	0.359	-0.120	1.033
11	2.636	0.962	0.510	-1.961
12	2.779	0.473	-0.040	-1.781
13	-1.429	-2.507	-2.980	-1.761
14	-0.051	-1.346	-1.730	-0.551
15	2.361	1.344	1.070	1.771
16	0.905	0.143	-0.140	0.462
17	-1.520	-0.372	-0.640	-0.904
18	-0.455	0.096	-0.150	-0.183
19	-0.598	0.586	0.400	-0.003
20	1.412	1.934	1.620	1.845
21	-7.238	-8.630	-9.230	-7.396

Table 3.1: Residuals from four different fits to the Stackloss data.

weighted least squares fit agree closely with the Andrews and 17-point least squares fit. The outliers can be easily recognized by looking at the residuals of the three robust fits. The residuals for observations 1, 3, 4, and 21 are higher in all the three robust fitted models as compare to the ordinary least squares fit. The sum of squares of residuals is smallest for the least squares fit for all 21 observations, next smallest for the weighted least squares fit and largest for the 17-point least squares and Andrews robust fits. The weighted least squares is equally efficient for detecting outliers as is the Andrews robust method.

The method of weighted least squares proposed here requires only modest computational resources and can be executed with standard statistical software because no iterative procedure is involved like in other M-estimators. This method requires calculating an ordinary least squares fit, finding weights from the residuals and applying weighted least squares technique to find the W-estimate for parameter.

4. Adaptation for the Problem

A beta distribution is symmetric about its mean when its two parameters are equal. In its standard form the beta distribution is

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1 \quad . \tag{3.17}$$

The mean, variance and mode of this distribution are

When $\alpha > \beta$ then the distribution is negatively skew and the left hand tail is longer than the right hand tail (the longer tail is directed towards x = 0), and as a weight function,

$$\begin{cases} \textit{Mean } X = \mu = \dfrac{\alpha}{\alpha+\beta} \\ \textit{Var } X = \sigma^2 = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ \textit{Mode } X = \tilde{x} = \dfrac{(\alpha-1)}{(\alpha+\beta-2)} \end{cases} \quad . \tag{3.18}$$

it gives more weight to observations above the mean than to ones below. Increasing the differences in α and β imply shifting the weight to the upper side of the distribution. By using the beta distribution as a weight function with different α and β we can find different percentile regressions from the data. This approach can be adapted for estimating of maximum species richness line proposed by Fausch *et al.* (1984), Karr *et al.* (1986). It gives an estimate for the maximum species richness line which is very close to the approach proposed by Qadir and Urquhart (1993).

Methodology 4.1.

Suppose $X_1, . X_2, \ldots\ldots, X_n$ denote response values on a random sample of units from a population with mean μ and variance σ^2, both unknown parameters. We also allow the possibility that there is some kind of contamination in the data away from the assumed model. We are interested in the maximum species richness line, that in the 90^{th} percentile regression line, for example. Obtain the residuals from the ordinary least squares fit; standardize the residuals dividing by their standard deviation; scale the standardized residuals as

$$t = \begin{cases} 0 & \text{if} \quad \dfrac{r_i + a}{2a} \le 0 \\ \dfrac{r_i + a}{2a} & \text{if} \quad 0 < \dfrac{r_i + a}{2a} < 1 \\ 1 & \text{if} \quad \dfrac{r_i + a}{2a} \ge 1 \end{cases} \tag{3.19}$$

where a is a tuning constant yet to be determined. Calculate the weight function according to Eq. **3.15**. Then use weighted least squares to estimate the maximum species richness line by using Eq. **3.4**. Values of α, β and tuning constant a should be used according to Table 3.2 to find various percentile lines. This method is discussed further in next section.

5. Simulation Results

Extensive simulation studies show that we can used different values of α, β and the tuning constant a to get various percentile lines. Table 3.2 gives the values for 60^{th}, 70^{th}, 80^{th} and 90^{th} percentile lines; if we switch the values of α and β for the same tuning constant a we can find 100(1 - P)% lines, for example to find 10 percentile line, we need to switch the values of α and β given for 90^{th} percentile line. Since our main concern is to estimate the maximum species richness line, we did not extend this table for general purposes. Clearly more research is needed in this direction for generalizing the table for other purposes.

The simulation study here is similar to the one used by Qadir and Urquhart (1993): The simulation size for each experiment was 1000. The data sets were

tuning constant	60 percentile		70 percentile		80 percentile		90 percentile	
	α	β	α	β	α	β	α	β
1.5	1.964	1.295	2.774	1.190	3.877	1.103	4.706	0.927
1.7	2.093	1.367	2.921	1.231	4.083	1.131	5.495	0.982
2.0	2.589	1.717	3.286	1.380	4.440	1.201	6.277	1.046
2.5	4.230	3.011	4.893	2.320	5.621	1.630	7.305	1.174
3.0	6.101	4.568	6.936	3.683	7.840	2.750	8.780	1.575

Table 3.2: Values of α and β for different values of tuning constant to find selected percentile regression lines.

generated according to the normal distribution with a mean of (25 + 1.2 X) and a variance of (8 + 0.4 X). The programming language GAUSS was used for simulation. Sample sizes of 50, 100, 150, ... 1000 were considered for the simulation purpose and X_i's were chosen as $X_i = (30 \times i)/n$ for $i = 1, 2, 3, \ldots\ldots n$. Table 3.3 displays the result for this simulation study.

Qadir and Urquhart (1993) compared two approaches for estimating maximum species richness line: maximum likelihood procedure based on a linear model, and a nonparametric procedure. Our robust weighted least squares approach gives results very similar to those of the maximum likelihood estimate and to the two observations per interval nonparametric approach. Table 3.3 shows selected results for n = 200, 300, ..., 1000 and for different values of the tuning constant. The values of α, β and tuning constant are selected from Table 3.2. These simulation results show that the weighted

tuning constant	1.5		2.0		2.5		3.0	
sample size	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
200	35.197	1.712	35.238	1.710	35.247	1.713	35.183	1.712
300	35.280	1.710	35.301	1.709	35.367	1.708	35.334	1.709
400	35.245	1.713	35.246	1.713	35.314	1.712	35.305	1.708
500	35.250	1.714	35.264	1.714	35.340	1.713	35.292	1.714
600	35.308	1.713	35.318	1.713	35.374	1.712	35.339	1.713
700	35.259	1.716	35.271	1.716	35.333	1.714	35.314	1.714
800	35.256	1.716	35.268	1.716	35.327	1.715	35.291	1.716
900	35.259	1.717	35.270	1.717	35.337	1.716	35.309	1.715
1000	35.258	1.717	35.276	1.717	35.351	1.715	35.290	1.717
true value	35.256	1.713	35.256	1.713	35.256	1.713	35.256	1.713

Table 3.3: Using values from Table 3.2, the estimates of intercept and slope for 90th percentile regression line obtained from 1000 simulations.

least squares procedure gives consistent values with virtually no bias. Figure 3.2 and Figure 3.3 display the mean-squared errors for the estimates of intercept and slope for the maximum species richness line using tuning constant 1.5, 2.0, 2.5, and 3.0. From these two graphs it is clear that there is no substantial difference among the four tuning constant.

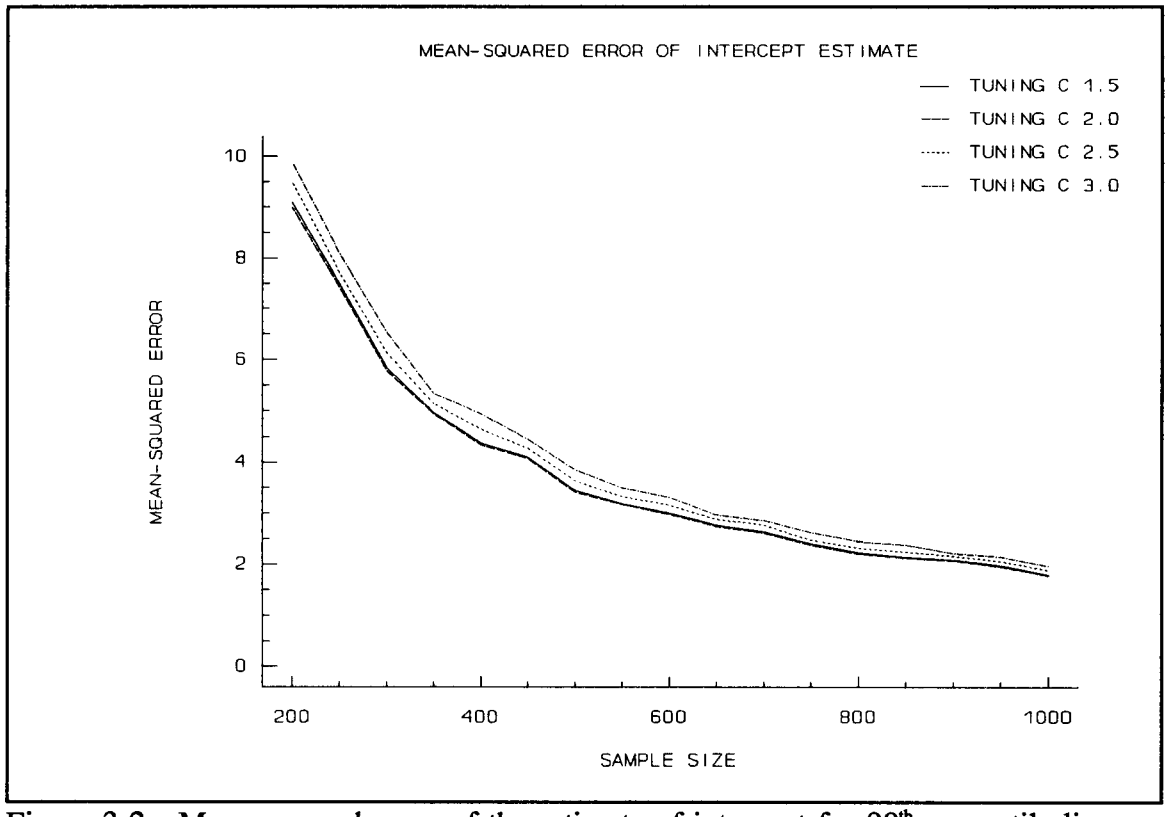

Figure 3.2: Mean-squared error of the estimate of intercept for 90th percentile line from 2500 simulations.

Example 5.1

Now consider the example of Ohio data (a real data set). The Ohio Environmental Protection Agency has an ongoing program of evaluating the biological condition of streams in that state. The resulting data is available in a public database described by Yoder (1991). Details of the field protocols and allied matters are documented in a User's Guide available from the Ohio EPA (1987). This data set was used by Qadir and Urquhart (1993) for illustrating estimation of the maximum species richness line. The data set has 245 observations; the response variable Y is species

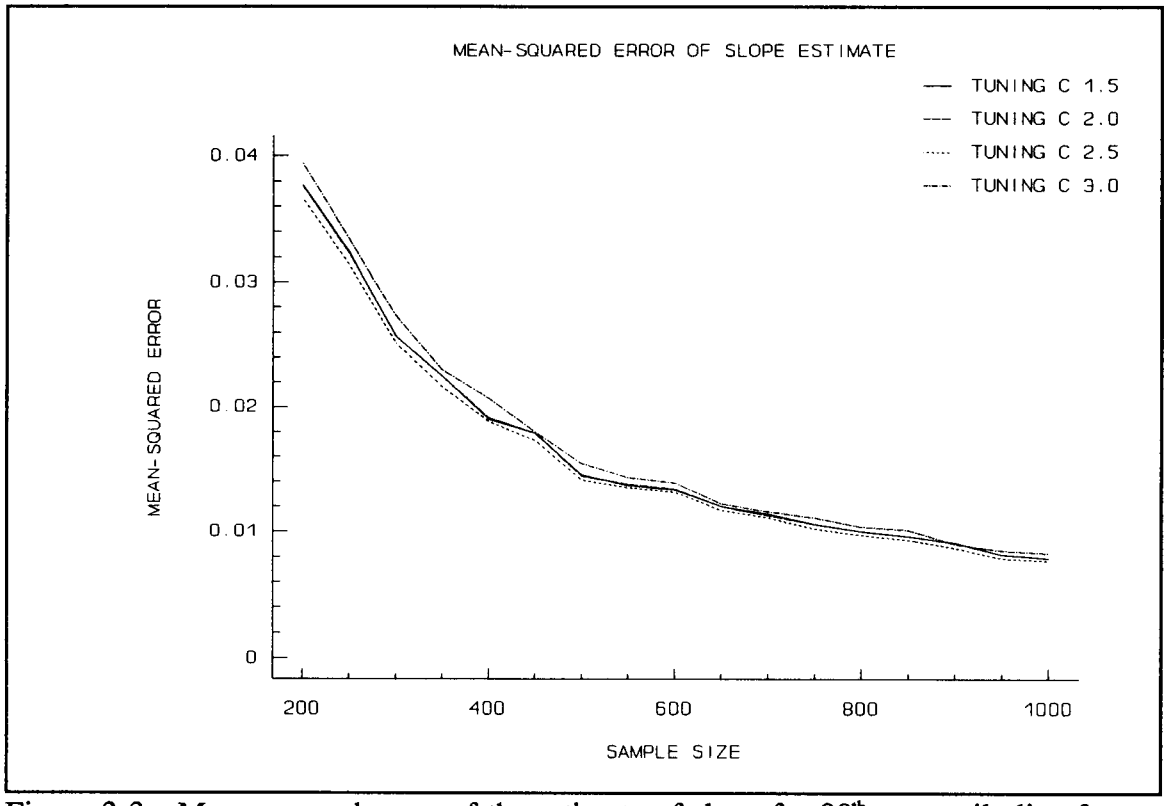

Figure 3.3: Mean-squared error of the estimate of slope for 90th percentile line from 2500 simulations.

richness, and the regressor variable X is the drainage area above the stream sampling point. After taking log transformation of the regressor variable X, the data appears to follow the assumed model.

The ordinary least squares fit to the data is

$$\hat{y} = 6.427 + 9.203 \log(X) .$$

Using Eq. **3.15** with $\alpha = 8.78$ and $\beta = 1.575$ and tuning constant $a = 3$ we got the weighted least squares fit

$$\hat{y}_{90} = 11.289 + 9.895 \log(X) ,$$

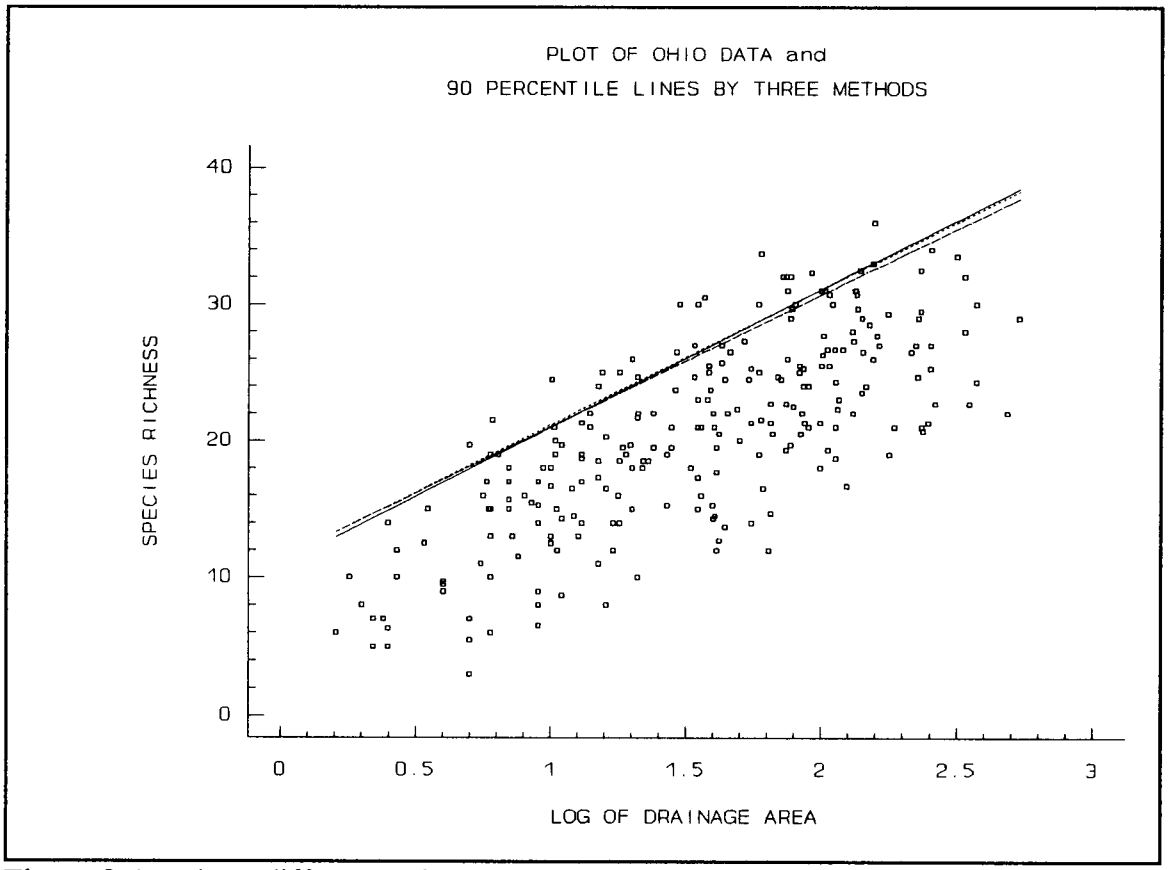

Figure 3.4: Three different estimates of maximum species richness line, linear model (solid line), nonparametric (dashed line) and weighted least squares (dotted line).

where $\hat{y}_{90}$ is the fitted value of the 90th percentile regression line, an estimate of maximum species richness line. This estimate of maximum species richness line is very close that obtained by Qadir and Urquhart (1993) using the maximum likelihood estimates based on a linear model with four unknown parameters:

$$\hat{y}_{90} = 10.869 + 10.119 \log(X) ,$$

and the nonparametric approach of two observations per interval:

$$\hat{y}_{90} = 11.347 + 9.675 \log(X) .$$

The robust weighted least squares fit is very closed to these fits shown in Figure 3.4. The advantage of the weighted least squares fit is that this is robust, and computationally simpler than the maximum likelihood approach based on a linear model; in fact it requires no iteration. Consequently it requires no starting values for an iterative procedure.

6. Conclusions

The simulation study shows that the method of weighted least squares proposed here can be applied successfully to a practical sized problem. Finding the weights from a beta function poses only a minor inconvenience; present personal computer speed makes it computationally feasible. However the approach advanced here is simple compared to maximum likelihood approach based on a linear model discussed by Qadir and Urquhart (1993). The procedure requires two steps which can use standard software: First, residuals are computed using simple linear regression, then weighted least squares is applied. This is computationally simpler and faster than the maximum likelihood estimation.

The selection of a tuning constant depends on the data set. If we suspect too many outliers in the data then we need to use a tuning constant of 1.5 or 1.7 or 2, but if we expect no potential outliers, then use the large tuning constant value of 2.5 or 3.

Clearly, further studies of these robust analysis are needed. The effect of parameters α and β used in the $\omega(t)$ function in these analyses needs to be more fully

explored. On the basis of these methods, it is hoped that biologist will find these statistical methods useful.

ACKNOWLEDGMENTS

The work of the senior author was supported by the Pakistan Participant Training Program funded by US AID. The work of the second author was supported in part by Cooperative Agreement CR 816721 between the Department of Statistics at Oregon State University and the Environmental Protection Agency.

REFERENCES

Andrews, D. F., (1974). A robust method for multiple linear regression, *Technometrics*, 16, 523-531.

Andrews, D. F., and D. Pregibon, (1978). Finding the outliers that matter, *Journal of Royal Statistical Society Series B*, 40, 85-93.

Apteck Systems (1992). *The GAUSS System Version 3.0.* Apteck Systems, Inc. 23804 S.E. Kent-Kangley Road, Maple Valley, Washington 98038.

Atkinson, A. C. (1985). *Plots, transformations and regression*, Oxford: Oxford university press.

Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables, *Journal of Royal Statistical Society Series B,* 44, 1-36.

Birkes, D., and Y. Dodge., (1993). *Alternative Methods of Regression*, John Wiley & Sons, New York.

Brownlee, K. A. (1965). *Statistical Theory and methodology in Science and Engineering*, 2nd ed., John Wiley & Sons, New York.

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

Cook, R. D. (1979). Influential observations in regression, *Journal of the American Statistical Association,* 74, 169-174.

Daniel, C., and F. S. Wood, (1971). *Fitting Equations to Data*, John Wiley & Sons, New York.

Dempster, A. P., and M. Gasko-Green, (1981). New tools for residual analysis, *The Annals of Statistics,* 9, 945-959.

Draper, N. R., and H. Smith, (1981). *Applied Regression Analysis*, John Wiley & Sons, New York.

Fausch, K. D., J. R. Karr, and P. R. Yant, (1984). Regional application of an index of biotic integrity based on stream-fish communities. *Transactions of the American Fisheries Society*, 113, 39-55.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey, (1983). *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York.

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

Karr, J. R. (1981). Assessment of biotic integrity using fish communities. *Fisheries*, 6, 21-27.

Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant, and I. J. Schlosser, (1986). Assessing biological integrity in running waters, A method and its rational. *Illinois Natural History Survey Special Publication* 5.

Li, G. (1985). *Robust regression, in Exploring Data Tables, Trends, and Shapes*, edited by D. Hoaglin, F. Mosteller, and J. Tukey, John Wiley & Sons, New York.

Ohio Environmental Protection Agency (1987). Biological Criteria for the Protection of Aquatic Life: Volume II. *Users Manual for Biological Field assessment of Ohio Surface Waters.* Division of Water Quality Monitoring and Assessment, Surface Water Section, Columbus, Ohio.

Qadir, M. F., and N. S. Urquhart, (1993). Percentile Regression: A Tool for Relating Species richness to System Size and Environmental Impact. Personal communication; manuscript to be submitted for publication.

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

Rousseeuw, P. J., and A. Leroy, (1987). *Robust regression and outlier Detection*, John Wiley & Sons, New York.

Yoder, C. O. (1991). The integrated Biosurvey as a tool for the evaluation of Aquatic life use attainment and impairment in Ohio Surface waters. Biological criteria: Research and Regulation. *Proceeding of a National Conference*, U. S. EPA, Office of Water, Washington D. C.

CHAPTER IV

SUMMARY AND CONCLUSIONS

The use of regression percentiles provides a natural approach for estimating a maximum species richness line. In the previous two chapters three different approaches were proposed: maximum likelihood using a normal model with hetroscedastic variances, a nonparametric approach, and an adaptation of robust regression. Simulation experiments were carried out to compare these approaches for estimating the maximum species richness line.

The maximum likelihood approach using an assumed model works well if model assumptions are satisfied, while the robust regression works well if there are outliers. The maximum likelihood approach can be applied to find any percentile regression; once the maximum likelihood estimates of the model are obtained any percentile line can be determined easily. Among the three methods compared in the simulation study the generalized linear model method is most statistically efficient. Simulation studies show that this method works reasonably well, even when its underlying model is violated, as for example when the data come from mixtures of normal distributions. If however the mean of uncontaminated normal data from one population differs substantially from the mean of other population, then the nonparametric method of two observations per interval performs better than the maximum likelihood estimate.

The iteratively reweighted least squares method merely uses a reparametrization of the original model, but it may have computational appeal for some users. It is simple

to use because it iterates on only one estimator. This method might be attractive to those people who are familiar with and use ordinary least squares regression. This method is slower than the four-dimensional maximization of the likelihood function, but for a sample size of n = 245 and using the GAUSS software on a 486 personal computer with 33 megahertz processor both computational methods take less than one second.

The nonparametric approach provides a simple method for estimating the maximum species richness line. Simulation studies show that a part of the data can be used safely to find the maximum species richness line. Using only 20% data can be very appealing for extremely large data set. This approach is very effective if the data come from a mixture of subpopulations which is likely in situations in which the IBI is used, i.e., where pollution and other human interference contaminate streams. The nonparametric method seems to be a reasonable choice for estimating maximum species richness line when line is constructed from primarily undisturbed data.

The robust method presented in chapter III can be used when we suspect outliers. The two examples discussed in section 3.2 show that this method can be used generally for detecting outliers in regression models. The weight function used for estimating the maximum species richness line depends on the two parameters of a beta function and a tuning constant. It is proposed that if a user suspects outliers then he/she should use a smaller tuning constant ($\approx$ 1.5). For data with no apparent outliers, the tuning constant of 3 works well.

The robust method discussed in the previous chapter is easy to use. The standardized ordinary least squares residuals can be used to find the weight function from

beta function using specified values of α, β and the tuning constant (Table 3.2). Once these weights are determined, then the weighted least squares procedure is applied to find an estimate for the maximum species richness line. This approach requires no iteration, a distinct advantage in finding estimates of maximum species richness line.

Based on these results and other simulation experiments not reported here, the following conclusions can be drawn:

- All three approaches are moderately effective for estimating maximum species richness line and can be applied to moderate sample sizes.
- The maximum likelihood approach using a linear model is clearly the most statistically efficient approach. This method should be used if the assumptions are not grossly violated. However the iterative procedure is tedious for nonstatisticians, because it uses a 4×1 score vector and a 4×4 hessian matrix, and requires initial values.
- The iteratively reweighted least squares computational method is based on a reparameterization of the linear model. It yields the same parameter estimates as the Newton-Raphson computation. Its advantage is computational simplicity because it requires iteration on only one parameter.
- The Robust regression (weighted least squares) approach automatically minimizes the effect of outliers if there are any. Thus this method is more useful in situations where outliers can frequently occur.

- The nonparametric method is the simplest method to use. This method is also useful for finding an initial estimate of a maximum species richness line. This is more robust to distributional assumptions than the maximum likelihood approach.

All three of these procedures are comparable in a sense, but in general, we found that maximum likelihood approach perform noticeably better than the others methods when the assumptions of its model are approximately satisfied. Design variations such as mixtures of subpopulations make it vulnerable, but in general, the maximum likelihood approach outperforms the nonparametric and robust methods of estimation.

BIBLIOGRAPHY

Andrews, D. F., (1974). A robust method for multiple linear regression, Technometrics, 16, 523-531.

Andrews, D. F., and D. Pregibon, (1978). Finding the outliers that matter, *Journal of Royal Statistical Society Series B,* 40, 85-93.

Angers, C. (1979). Simultaneous Estimation of Percentile Curves with Application to Salary Data. *Journal of the American Statistical Association,* 74, 621-625.

Apteck Systems (1992). The GAUSS System Version 3.0. *Apteck Systems, Inc.* 23804 S. E. Kent-Kangley Road, Maple Valley, Washington 98038.

Atkinson, A. C. (1985). *Plots, transformations and regression*, Oxford: Oxford university press.

Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables. *Journal of Royal Statistical Society Series B*, 44, 1-36.

Birkes, D., and Y. Dodge., (1993). *Alternative Methods of Regression*, John Wiley & Sons, New York.

Brownlee, K. A. (1965). *Statistical Theory and methodology in Science and Engineering*, 2nd ed., John Wiley & Sons, New York.

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

Carroll, R. J., and D. Ruppert, (1982). Robust Estimation in Heteroscedastic Linear Models. *The Annals of Statistics,* 10, 429-441.

Cook, R. D. (1979). Influential observations in regression. *Journal of the American Statistical Association,* 74, 169-174.

Daniel, C., and F. S. Wood, (1971). *Fitting Equations to Data,* John Wiley & Sons, New York.

Davidian, M., and R. J. Carroll, (1987). Variance Function Estimation. *Journal of the American Statistical Association*, 82, 1079-1091.

Dempster, A. P., and M. Gasko-Green, (1981). New tools for residual analysis. *The Annals of Statistics*, 9, 945-959.

Draper, N. R., and H. Smith, (1981). *Applied Regression Analysis*, John Wiley & Sons, New York.

Fausch, K. D., J. R. Karr, and P. R. Yant, (1984). Regional application of an index of biotic integrity based on stream-fish communities. *Transactions of the American Fisheries Society,* 113, 39-55.

Griffiths, D., and M. Willcox, (1978). Percentile Regression: A Parametric Approach. *Journal of the American Statistical Association,* 73, 496-498.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey, (1983). *Understanding Robust and Exploratory Data Analysis,* John Wiley & Sons, New York.

Hogg, R. V. (1975). Estimates of Percentile Regression Lines Using Salary Data. *Journal of the American Statistical Association,* 70, 56-59.

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

Karr, J. R. (1981). Assessment of biotic integrity using fish communities. *Fisheries*, 6, 21-27.

Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant, and I. J. Schlosser, (1986). Assessing biological integrity in running waters, A method and its rational. *Illinois Natural History Survey Special Publication* 5.

Kennedy, W. J. Jr., and J. E. Gentle, (1980). *Statistical Computing,* Marcel Dekker, Inc. New York.

Li, G. (1985). Robust regression, in *Exploring Data Tables, Trends, and Shapes,* edited by D. Hoaglin, F. Mosteller, and J. Tukey, John Wiley & Sons, New York.

McCullagh, P. and J. A. Nelder, (1989). *Generalized Linear Models,* (2nd Ed.) London: Chapman and Hall.

Ohio Environmental Protection Agency (1987). Biological Criteria for the Protection of Aquatic Life: Volume II. *Users Manual for Biological Field assessment of Ohio Surface Waters.* Division of Water Quality Monitoring and Assessment, Surface Waters Section, Columbus, Ohio.

Qadir, M. F., and N. S. Urquhart, (1993). Percentile Regression: A Tool for Relating Species Richness to System Size and Environmental Impact. Personal communication; manuscript to be submitted for publication.

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

Rousseeuw, P. J., and A. Leroy, (1987). *Robust Regression and Outlier Detection,* John Wiley & Sons, New York.

Yoder, C. O. (1991). The integrated Biosurvey as a tool for the evaluation of Aquatic life use attainment and impairment in Ohio surface waters. Biological criteria: Research and Regulation. *Proceeding of a National Conference*, U. S. EPA, Office of Water, Washington, D. C.