Robust Regression Approach to Analyzing Fisheries Data

Y. Chen

Department of Zoology, University of Toronto, Toronto, ON M5S 1A1, Canada

D.A. Jackson[1]

Department of Zoology, University of Western Ontario, London, ON N6A 5B7, Canada

and J.E. Paloheimo

Department of Zoology, University of Toronto, Toronto, ON M5S 1A1, Canada

Chen, Y., D.A. Jackson, and J.E. Paloheimo. 1994. Robust regression approach to analyzing fisheries data. Can. J. Fish. Aquat. Sci. 51: 1420–1429.

Fisheries data often contain inaccuracies due to various errors. If such errors meet the Gauss–Markov conditions and the normality assumption, strong theoretical justification can be made for traditional least-squares (LS) estimates. However, these assumptions are not always met. Rather, it is more common that errors do not follow the Gauss–Markov and normality assumptions. Outliers may arise due to heterogenous variabilities. This results in a biased regression analysis. The sensitivity of the LS regression analysis to atypical values in the dependent and/or independent variables makes it difficult to identify outliers in a residual analysis. A robust regression method, least median squares (LMS), is insensitive to atypical values in the dependent and/or independent variables in a regression analysis. Thus, outliers that have significantly different variances from the rest of the data can be identified in a residual analysis. Using simulated and field data, we explore the application of LMS in the analysis of fisheries data. A two-step procedure is suggested in analyzing fisheries data.

Les données sur les pêches recèlent souvent des inexactitudes attribuables à différentes erreurs. Lorsque ces erreurs sont conformes aux conditions associées au théorème Gauss–Markov et à l'application de la normalité, on peut évoquer d'excellentes justifications théoriques pour le recours aux estimations par la méthode classique des moindres carrés (MC). Mais ce n'est pas toujours le cas. En fait, c'est plus souvent le contraire. Il peut y avoir des valeurs extrêmes aberrantes, attribuables à des variabilités hétérogènes. Celles-ci donnent lieu à des analyses de régression biaisées. La sensibilité de l'analyse (par régression des moindres carrés) aux valeurs atypiques prises par les variables dépendantes ou encore indépendantes, complique la tâche d'identifier les observations extrêmes aberrantes dans une analyse des résidus. Une méthode de régression robuste, la méthode par les moindres carrés médians (MCM), est insensible à ces valeurs atypiques. Ainsi peut-on identifier les valeurs extrêmes aberrantes auxquelles sont associées des variances significativement différentes de celles du reste des données, dans le cadre d'une analyse des résidus. En appliquant des données simulées et obtenues sur le terrain, nous explorons l'application de la méthode des MCM à l'analyse des données des pêches. Nous proposons une procédure à deux étapes pour analyser les données de ces dernières.

Received April 30, 1993
Accepted December 23, 1993
(JB908)

Reçu le 30 avril 1993
Accepté le 23 décembre 1993

Problems with the Least Squares Method

Investigation of relationships between variables related to fish and the environment that fish inhabit often is one of the most important objectives for fisheries studies. Such an investigation is usually quantified by a linear regression equation written as

$$(1) \quad Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_K X_{K,i} + \epsilon_i, \; i = 1, \ldots, N$$

where Y, the dependent variable, is some measure associated with fish, X's, the independent variables, are usually environmental variables, K is the number of independent variables, N is the sample size, ϵ is the error term, and β's are the parameters to be estimated. Because of tradition and ease of computation, the least-squares (LS) method has been adopted generally to estimate β's. To derive the unbiased LS estimates of β's and theoretically justified variance estimates, the Gauss–Markov conditions must be met, i.e., ϵ is independent and identically distributed with a constant variance. It is assumed also that the independent variable(s) is (are) free of error. The assumption of normality is required for hypothesis tests or the construction of confidence intervals on the estimated parameters.

Fisheries and environmental data are typically contaminated by nonnormal errors with heterogenous variances. These errors may result from various sources: errors in measurement and recording, variation from unusual events or abnormal environmental conditions, or samples containing measurements coming from different populations. These error sources

[1]Present address: Department of Zoology, University of Toronto, Toronto, ON M5S 1A1, Canada.

can affect not only the dependent variable but also the independent variables, thereby creating outliers (or influential data points) defined as observations that are aberrant or unusually different from the rest of the observations (Rousseeuw and Leroy 1987). The existence of outliers causes many fisheries data to fail meeting the Gauss–Markov conditions and normality assumption. In some cases, this problem can be remedied by an appropriate transformation of the data (e.g., log-transformation, Kimura 1989; Jolicoeur 1991), by using generalized least-squares methods such as weighted least squares (Sen and Srivastava 1990) and iterative reweighted least squares (e.g., biweight, Holland and Welsch 1977), or by fitting the data to more complex models such as a nonlinear model. However, it is predictable that those approaches may not always function in improving the quality of data in regression analysis with respect to the normal error and constant variances. A linear regression model may be required when studying the fish environment data for various reasons, e.g., studying biological theories and developed models. Moreover, if some data points do have atypical values relative to the majority of data in a regression analysis, perhaps they should not be included in the regression analysis. In the case of simple regression analysis, it is easy to identify outliers by examining the bivariate plot. However, such a visual inspection may be quite subjective, and it is impossible to do so if there are more than two independent variables. Regression diagnostics techniques have been applied to identify the outliers in the multiple LS regression analysis. However, it is difficult to do such an analysis if there are multiple outliers in a data set with a large number of observations. As well, it may be difficult to identify these outliers from a residual plot, since they do not always show up in such a plot (outliers usually pull the regression line towards themselves, Sen and Srivastava 1990). In practice, outliers in fisheries data often go unnoticed because much fisheries data are processed by computers without careful a priori inspection and screening. This is particularly true with data sets having a large number of observations. The result from a regression analysis of such data may be incorrect due to the existence of outliers.

Robust Regression Analysis

Robust regression (RR) techniques have been developed that are less sensitive to outliers in the data compared with the LS. Because RR estimation is not sensitive to outliers in the data, the outliers are usually far away from the fitted regression line, and their large residuals can be detected easily in a residual analysis. The following RR techniques have been well developed for the regression analysis: (1) least absolute value (LAV, Edgeworth 1887), (2) *M* estimates (Huber 1973), (3) least median of squares (LMS, Rousseeuw 1984), (4) *S* estimator (Rousseeuw and Yohai 1984), and (5) least trimmed squares (LTS, Rousseeuw 1984). These methods are all robust with respect to outliers in the dependent variable. However, the first two methods are not robust with outlying independent variables (Rousseeuw and Leroy 1987).

A parameter, the breakdown point, is used often to describe the ability of an estimation method in identifying the unbiased estimates for data having outliers. It is defined as the smallest fraction of contamination that can cause the estimator to take on biased values far away from the true estimates (Rousseeuw and Leroy 1987). It is obvious that the largest value that can be expected for the breakdown point is 50%. For a larger amount of contamination, it is impossible to distinguish the "good" and the "bad" parts of the data.

The following estimation methods are used in our study: LS, Ricker's geometric mean (GM), LAV, LMS, and the LMS-based reweighted least squares (RLS). The GM method (or reduced major axis) is well known in fisheries studies (Ricker 1975). It is recommended that it be used when the independent variable is subject to errors (McArdle 1988). Instead of minimizing the residual sum of squares like LS, the methods of LAV and LMS are

$$\underset{\hat{\beta}}{\text{minimize}} \sum_{i=1}^{N} |\epsilon_i|$$

and

$$\underset{\hat{\beta}}{\text{minimize}} \underset{i=1}{\overset{N}{\text{med}}} \epsilon_i^2,$$

respectively, where ϵ_i are the residuals calculated as the difference between the observed Y and estimated Y from the fitted regression line.

The estimation algorithm for LAV has been well documented in many regression books, but the estimation algorithm for LMS is relatively unknown. Unlike the traditional regression methods, it is perhaps impossible to write down a straightforward formula for the LMS estimator. The algorithm used by Rousseeuw and Leroy (1987) for the LMS analysis is similar to the bootstrap (Efron 1979), except for the following aspects: (1) sampling size equals the number of parameters ($K + 1$, K is number of the independent variables in a regression model as defined in equation (1)) in the regression model instead of equalling total sample size (N) as with the bootstrap and (2) sampling in LMS is not random; rather, includes all possible combinations of subsample of size ($K + 1$) from N data points. For equation (1), the LMS estimation algarithm can be described as follows: (a) repeatedly draw subsamples of "$K + 1$" different observations, (b) determine the regression surface (β_J) through the "$K + 1$" points for subsample J, (c) calculate $M = \text{med}(Y - X\beta_J)^2$ with respect to the whole data set, and (d) find the β_J which has the smallest M value for all possible subsamples of size $K + 1$, and it is the LMS-estimated β.

Rousseeuw and Leroy (1987) proposed procedures to identify outliers based on the analysis of residuals for LMS. They proposed (a) calculate $S_0 = 1.4826(1+5/(N-K-1))\sqrt{[\min(\text{med}\ \epsilon_i^2(\beta))]}$, (b) compute the residuals, R_i, based on the LMS-estimated regression equation, and (c) determine a weight w_i for the ith observation; if $|R_i/S_0| > 2.5$, data point i is an outlier (for more detailed information, refer to Rousseeuw and Leroy (1987)). They also suggested using a reweighted least-squares regression to estimate the variation for the parameters which is comparable with the variation estimated in an LS analysis. The RLS is an LS estimator with a weight of 0 for outliers and 1 for normal data points (i.e., outliers are excluded from the RLS analysis). The estimation algorithm for LMS was detailed by Rousseeuw and Leroy (1987) and for LAV by Bloomfield and Steiger (1983).

Simulation Study

Various simulation studies have been done using LMS and LAV (Rousseeuw and Leroy 1987). However, most of

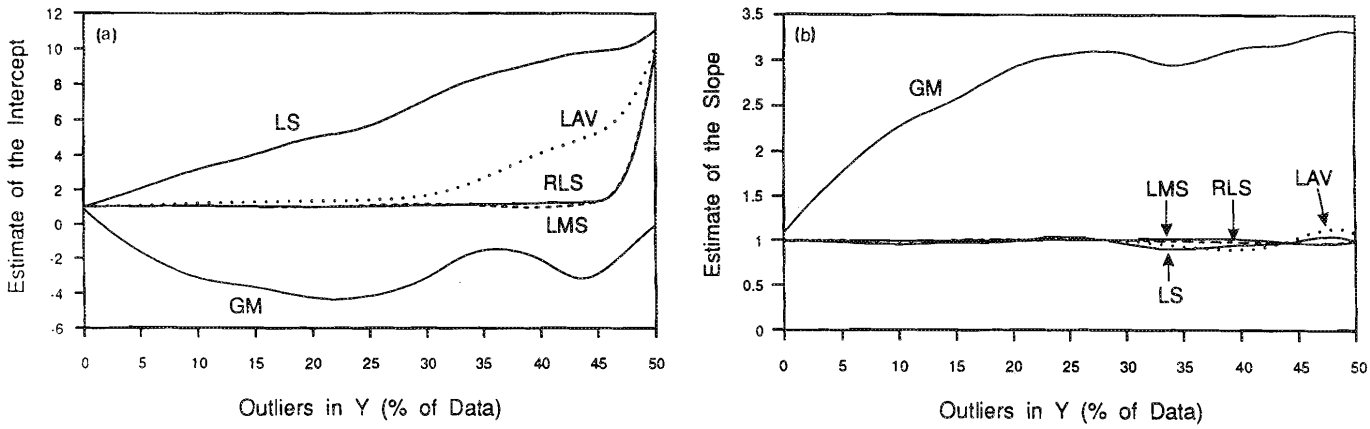

FIG. 1. Estimation of the parameters when there are outliers in Y: (a) intercept; (b) slope.

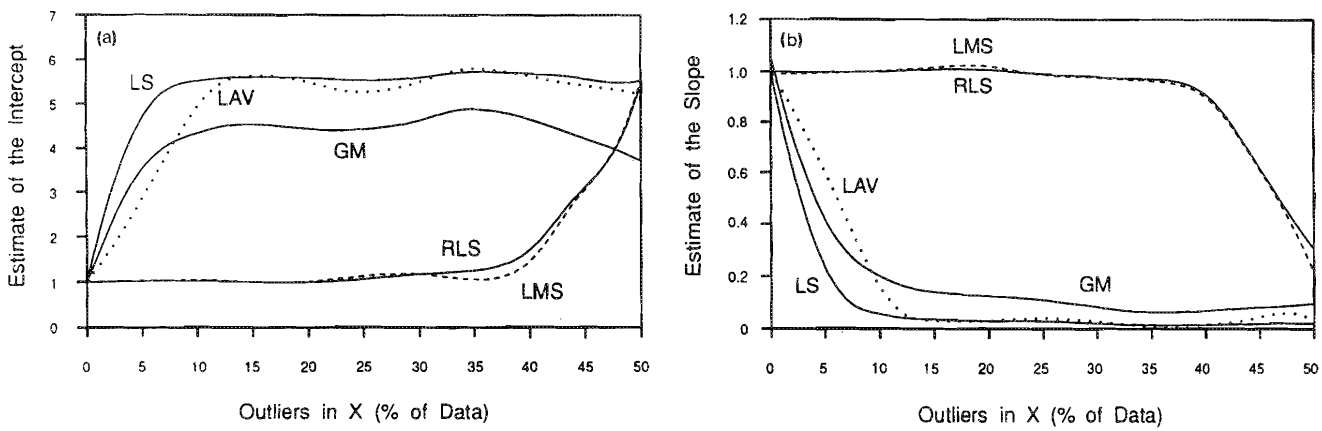

FIG. 2. Estimation of the parameters when there are outliers in X: (a) intercept; (b) slope.

these studies did not examine the effect of outliers on the estimation of each parameter in cases containing different error structures. In our study, a simple linear model was used for the simulation. The X's were randomly generated from a uniform distribution of numbers between 0 and 10. The "true" Y's were calculated from the X's using the simple linear model with both intercept (b_0) and slope (b_1) being 1. The "observed" Y's and/or X's were derived by adding the following error structures to the "true" Y's and/or X's. Four scenarios were considered in this study: (1) normal situation: ϵ in equation (1) having distribution of $N(0,1)$; (2) outliers in the dependent variable: a certain percentage of the data in Y is contaminated by an error term $N(20,2)$, but the rest of the data are the same as those in the normal situation; (3) outliers in the independent variable: a certain percentage of the data in X is contaminated by an error term $N(40,2)$, but the rest of the data are the same as those in the normal situation (the same observations were affected as in (2)), and (4) outliers in both Y and X: a certain percentage of the data is contaminated by the error terms in (2) and (3), respectively, but the rest of the data are the same as those in the normal situation.

The first scenario is usually assumed in the traditional LS regression analysis: errors in the Y variable are likely due to measurement errors or random environmental variation. The second and fourth scenarios can be observed in practice due to abnormal measurement errors for a small proportion of data points which may result from the failure or inappropriate use of measurement instruments, inexperienced workers (e.g., a new technician), errors in recording, or errors in entering data into computers. These two scenarios may also arise due to some unusual events such as abnormal environmental variables (e.g., very high or low temperature) or to unkowingly included individuals from a different population. The third scenario may be due to a limited ability to measure the variable accurately or errors in recording and typing data.

A sample size of 20 was simulated for each data set. The percentage of the data that was contaminated by an error term was chosen to increase from 0 to 50% by increments of 5%. One hundred simulations were performed for each of the four situations at each chosen percentage level of the contaminated data. The mean estimated values of b_0 and b_1 were plotted separately against the percentage of the data that had been contaminated by outliers. Because the true values of b_0 and b_1 are known, such plots show the differences in estimating the true values of the parameters among estimation methods.

Application of Robust Regression Methods to Field Fisheries Data

We applied the RR methods in the analysis of the following sets of fisheries data: (1) number of fish species and lake surface area (Barbour and Brown 1974), (2) fish sustained yields (SY) and lake thermal variables (Christie and Regier 1988), (3) fish SY and lake morphometric variables (THV = thermal habitat volume, THA = thermal habitat area; Christie and Regier 1988), (4) multiple regression analysis of fish SY versus THV or THA and total dissolved

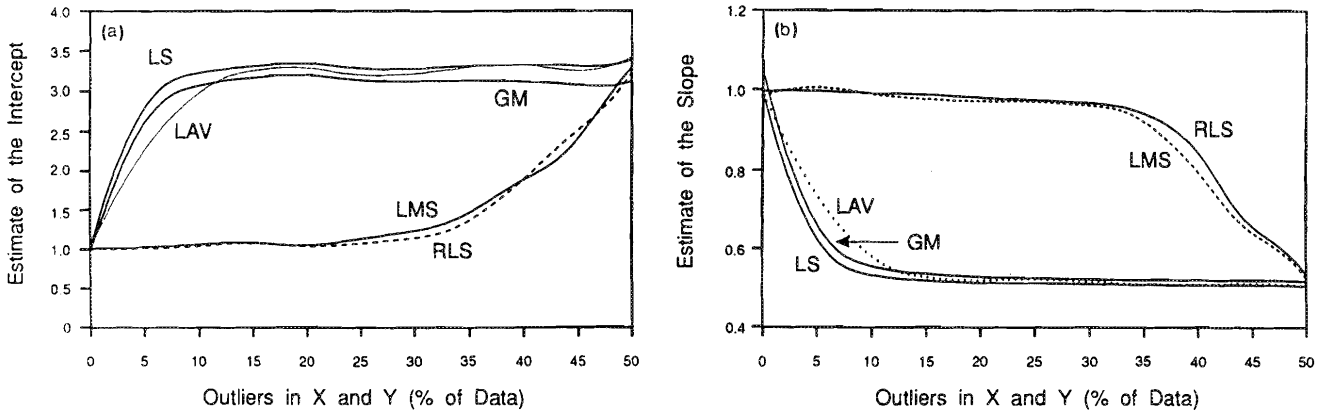

FIG. 3. Estimation of the parameters when there are outliers in both X and Y: (a) intercept; (b) slope.

solids (TDS; Christie and Regier 1988), and (5) fish size at age and lake pH (Ryan and Harvey 1980).

These data sets were all fitted to the linear regression equation by the original authors using the LS method. In our study, they were fitted to the same model using the LMS and RLS methods. The defined outliers in the LMS analysis were examined further as a means of explaining why they differed significantly from the rest of the data. If there were five or more outliers, a linear function was fitted to them using LMS and RLS (the minimum sample size for LMS analysis was five, as defined by Rousseeuw and Leroy (1987)). If the resultant RLS line was significant, differences in slope between the estimated line for the outliers versus the line for the remainder of the data were tested. If the differences were not significant at $p = 0.05$, the differences in intercept between two lines were examined.

Results

Simulation Study

Simulation results are summarized in Fig. 1 when there were outliers in Y. The LAV, LMS, and RLS appear to have smaller estimation biases than the LS and GM methods. The results are shown in Fig. 2 when only X contained outliers. It appears that the LMS and RLS methods yielded smaller estimation biases than the other three methods. When both X and Y had outliers, the LMS and RLS methods again yielded smaller estimation biases than the other three methods (Fig. 3).

Analysis of Field Fisheries Data

(1) Number of fish species and lake surface area

The LS-estimated linear regression equation was

$$\ln(\text{number of fish species}) = 2.34(0.199) + 0.143(0.0266)\ln(\text{lake area}),\ r = 0.55,\ N = 70$$

where the number in parentheses is the estimated standard error for the parameter, r is the correlation coefficient, and N is the sample size used in the estimation. The LMS-estimated equation was

$$\ln(\text{number of fish species}) = 2.649 + 0.019\ln(\text{lake area}),\ r = 0.45,\ N = 70.$$

The following 12 lakes were defined as outliers: Chad (3, numbers show lake position on Fig. 4), Malawi (7), Mweru (8), Tanganyika (13), Victoria (14), Zirahuen (31), Balkhash (37), Black Sea (39), Erie (54), Huron (55), Michigan (58), and Ontario (59) (Fig. 4). The RLS-estimated regression equation was

$$\ln(\text{number of fish species}) = 2.67(0.152) + 0.068(0.0222)\ln(\text{lake area}),\ r = 0.38,\ N = 58.$$

An LMS regression was used to fit a simple linear equation to 12 outliers. Four of 12 lakes were defined as outliers again: Malawi (7), Tanganyika (13), Zirahuen (31), and Balkhash (37). The resultant RLS equation was estimated as

$$\ln(\text{number of fish species}) = 3.32(0.177) + 0.136(0.0403)\ln(\text{lake area}),\ r = 0.78,\ N = 8.$$

There was a significant difference in slopes between the two RLS-estimated equations ($p = 0.04$).

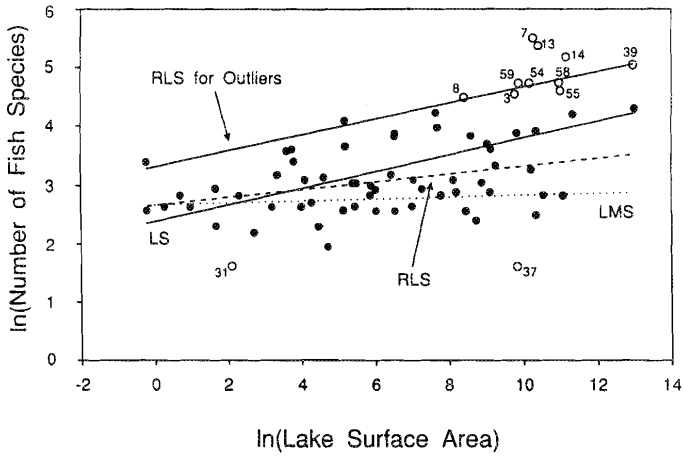

FIG. 4. Regression analysis of the number of fish species per lake to lake surface area. Lakes are numbered following Barbour and Brown (1974).

(2) Fish sustained yields and lake thermal variables

For lake trout (*Salvelinus namaycush*), Lake of the Woods (12) and Lake Erie (17) were identified as outliers in the LMS analysis of SY with THA and with THV. Both lakes had smaller $\log_{10}(SY)$ of lake trout than predicted for their values of $\log_{10}(THV)$ and $\log_{10}(THA)$ (hereafter, $\log_{10}$ is denoted as log). The RLS-estimated slope was smaller and the intercept was greater than those of LS regression analysis (Fig. 5a and 5b).

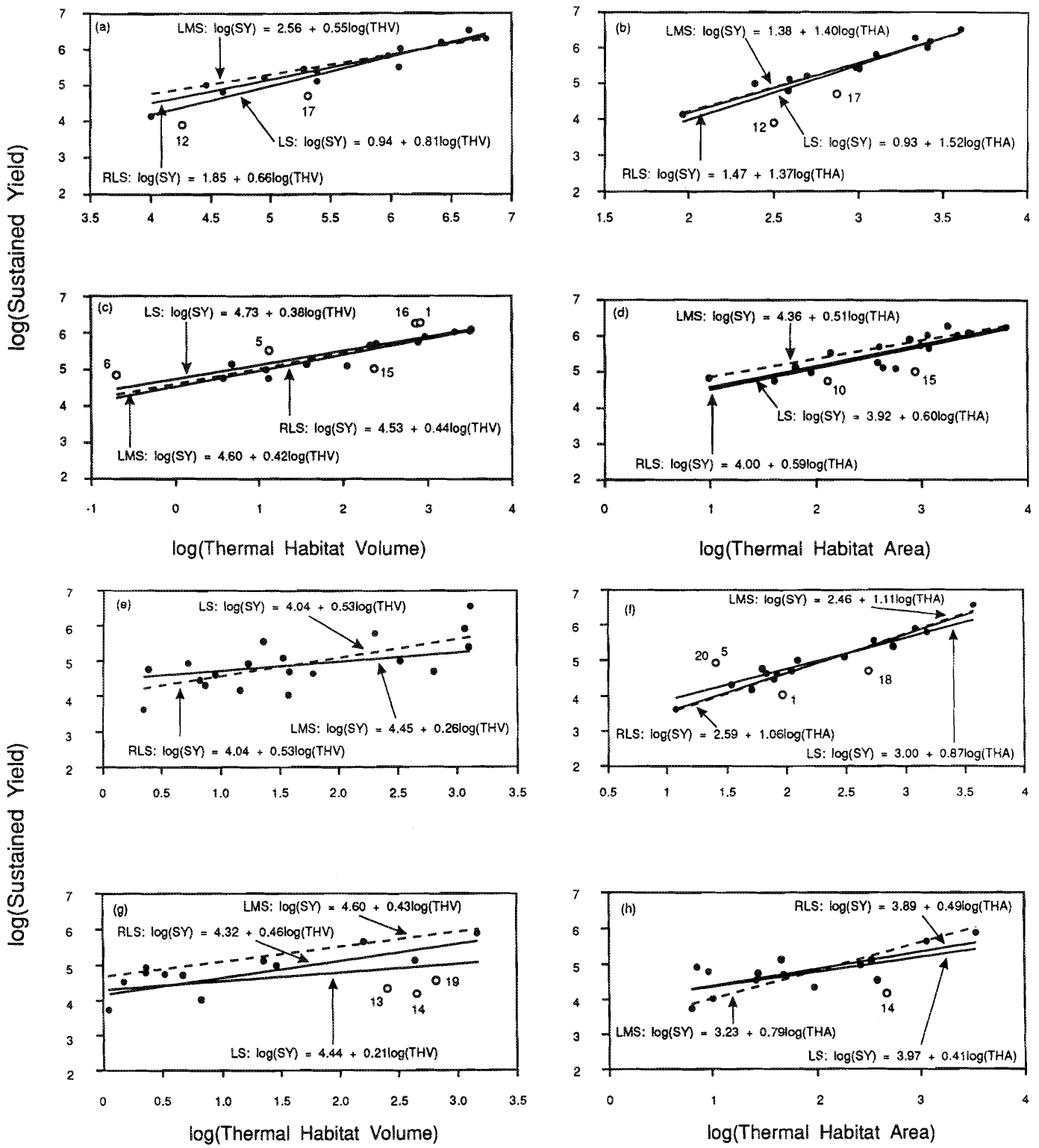

FIG. 5. Plots of ln-transformed total SY ($kg \cdot yr^{-1}$, ln = $\log_e$) versus ln-transformed THV ($hm^3 \cdot 10\ d^{-1}$) and THA ($ha \cdot 10\ d^{-1}$) for four fish species (data from Christie and Regier 1988): (a and b) lake trout; (c and d) lake whitefish; (e and f) walleye; (g and h) northern pike.

For lake whitefish (*Coregonus clupeaformis*), the following lakes were defined as outliers in the LMS analysis of log(SY) versus log(THV): Great Slave Lake (1), Big Peter Pond (5), Little Peter Pond (6), Georgian Bay (15), and North Channel (16) (Fig. 5c). The second RR analysis for these five lakes provided an RLS equation of log(SY) = 5.21 + 0.45log(THV), $r = 0.68$, $N = 4$, with Georgian Bay (15) remaining an outlier. There was no significant difference in the slope ($p = 0.10$), but the intercept of the above equation was significantly larger than that of the first RLS-estimated equation ($p < 0.001$). Georgian Bay had a significantly smaller log(SY) for its log(THV) compared with

TABLE 1. Coefficient of determination for the regressions of species SY ($\log_{10}$-transformed values, kg·yr^{-1}) on several independent variables ($\log_{10}$-transformed values). NS indicates that the slope of the regression is not significantly different from 0 ($p > 0.05$). Numbers in parentheses are the sample sizes used in the regression analysis.

Species	Estimation method	Independent variable: THA (ha·10 d^{-1})	THV (hm^3·10 d^{-1})	Area (ha)	Volume (hm^3)
Lake trout	LS	0.81 (15)	0.86 (15)	0.59 (15)	0.71 (15)
	RLS	0.95 (13)	0.88 (13)	0.90 (12)	0.83 (12)
Lake whitefish	LS	0.66 (19)	0.69 (19)	0.75 (19)	0.76 (19)
	RLS	0.75 (17)	0.89 (14)	0.75 (19)	0.76 (19)
Walleye	LS	0.72 (19)	0.47 (19)	0.15 (19) NS	0.07 (19) NS
	RLS	0.95 (15)	0.47 (19)	0.10 (18) NS	0.05 (18) NS
Northern pike	LS	0.38 (15)	0.17 (15) NS	0.03 (15) NS	0.01 (15) NS
	RLS	0.54 (14)	0.62 (12)	0.03 (15) NS	0.01 (15) NS

other lakes (Fig. 5c). Two outliers were identified in the LMS analysis of log(SY) versus log(THA): Lac Ile-à-la-Crosse (10) and Georgian Bay (15). Both lakes had smaller log(SY) for their log(THA) than other lakes (Fig. 5d).

For walleye (*Stizostedion vitreum*), no outliers were identified in the LMS analysis of log(SY) versus log(THV) (Fig. 5e). However, the following four lakes were identified as outliers in the LMS analysis of log(SY) versus log(THA): Great Slave Lake (1), Big Peter Pond (5), Lake Ontario (18), and Lake Nipigon (20). The RLS-estimated intercept and slope in the log(SY)–log(THA) regression were greater and smaller, respectively, than those from the LS (Fig. 5f).

For northern pike (*Esox lucius*), Lake Superior (13), Lake Huron (14), and Lake Michigan (19) were identified as outliers in the LMS analysis of log(SY) versus log(THV) (Fig. 5g). These three lakes had significantly smaller log(SY) for their log(THV) than other lakes. The RLS-estimated slope was twice as large as that of LS, but the RLS intercept was almost the same as that of LS (Fig. 5g). In the LMS analysis of log(SY) versus log(THA), Lake Huron (14) was defined as an outlier and had a smaller log(SY) for its log(THA) compared with other lakes. The RLS-estimated intercept and slope were smaller and larger, respectively, than those from the LS (Fig. 5h).

For all four fish species, the RLS-estimated coefficient of determination (r^2) was greater than the LS-estimated r^2 for those regression analyses in which there were outliers identified in the LMS analysis (Table 1).

(3) Fish sustained yields and lake morphometric variables

For lake trout, Amisk Lake (2), Lake of the Woods (12), and Lake Erie (17) were defined as outliers in the LMS analyses of log(SY) versus log(AREA) and log(SY) versus log(VOL). All three lakes had significantly smaller log(SY) for their values of log(AREA) and log(VOL) than other lakes in the LMS analysis (Fig. 6a and 6b). There were no outliers identified in the LMS analysis of log(SY) versus log(AREA) and log(VOL) for lake whitefish (Fig. 6c and 6d). For walleye, Lake Erie (17) was defined as an outlier in the LMS analysis of log(SY) and either log(AREA) or log(VOL) (Fig. 6e and 6f). There were no outliers identified for northern pike (Fig. 6g and 6h). For both lake trout and walleye, the differences in the slope and intercept between LS and RLS were the same as those in the analysis of SY versus THV and THA (Fig. 5a, 5b, 6a, and 6b for lake trout; Fig. 5e, 5f, 6e, and 6f for walleye). The RLS-estimated r^2 was larger than the LS r^2 for lake trout (Table 1). For walleye, the RLS r^2 was smaller than the LS r^2, but neither of the RLS- and LS-estimated equations was significant (Table 1).

(4) Multiple regression analysis of fish SY versus THV or THA and TDS

Variables selected as the regressors in this study were the same as those in Christie and Regier (1988). For all four fish species, there was at least one lake defined as an outlier (Table 2). The RLS-estimated parameters differed from those of LS. For lake trout, the LS-estimated parameter associated with log(TDS) was significantly different from 0, but the RLS-estimated value did not differ significantly from 0. For northern pike, the LS-estimated parameter associated with log(TDS) did not differ significantly from 0, but the RLS estimate differed significantly from 0. The RLS r^2 was greater than the LS r^2 for all four fish species.

(5) Fish size at age and lake pH

The size of yellow perch (*Perca flavescens*) at age 1 was regressed against lake pH values for 25 lakes from the La Cloche Mountain region of Ontario (Ryan and Harvey 1980). Two lakes with the lowest pH values were identified as outliers. The LS-estimated r^2 was larger than that of RLS by 40% (Fig. 7). All the estimated parameters were significantly different from 0. The negative slopes of LS and RLS show that the size of yellow perch at age 1 was greater in more acid lakes.

Discussion

The regression parameters are biased when X is subject to measurement error. The bias is often negligible if this error is small relative to the errors in the Y variable. If the error rate of the X variable is more than a third of that on the Y variable, intercept and slope are over- and underestimated, respectively, and the GM method is suggested to replace LS to reduce the bias (McArdle 1988). However, the validity of the GM method has been questioned recently (Jolicoeur 1990; Kimura 1992). In the simulation study, no normal

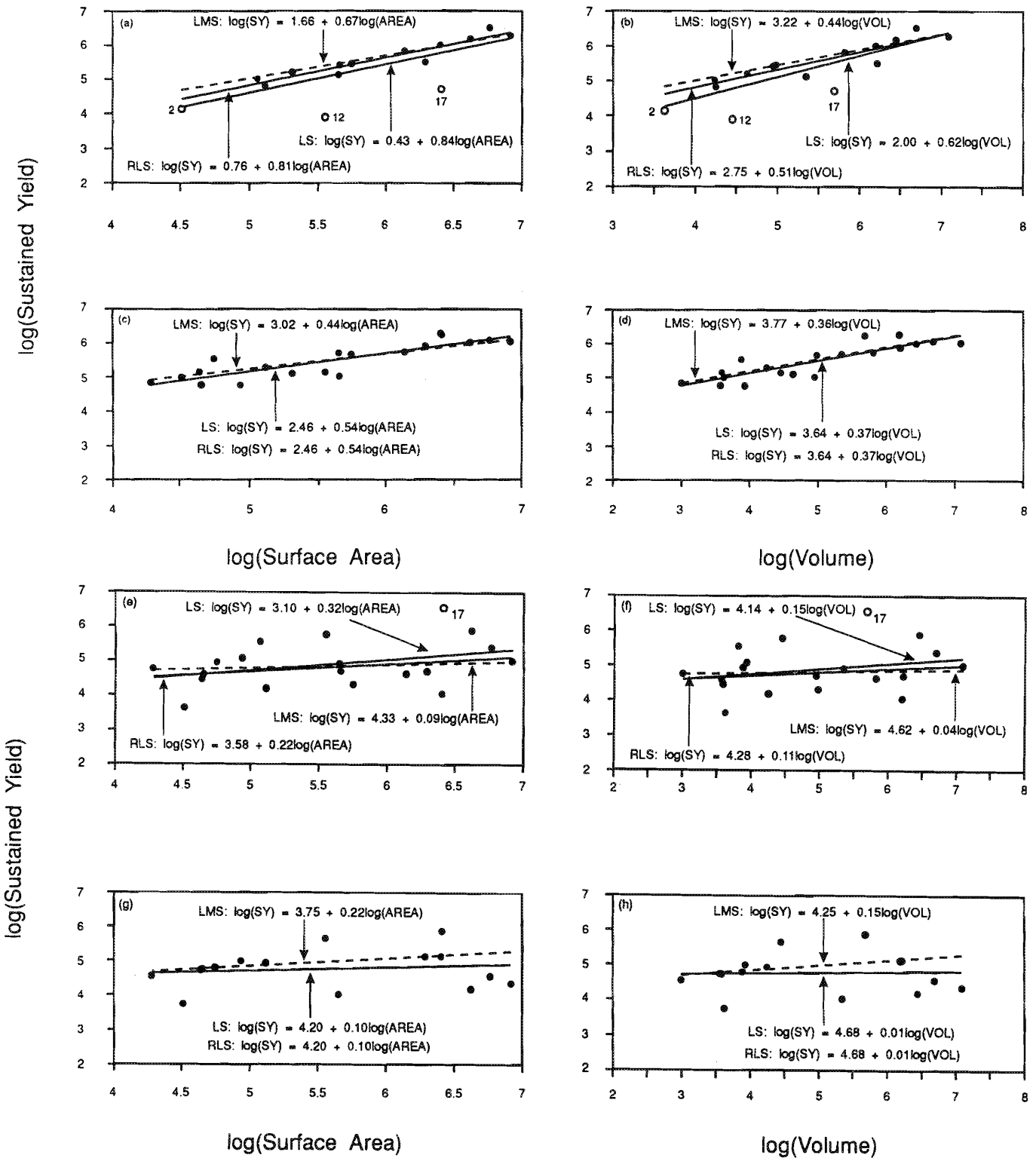

FIG. 6. Plots of ln-transformed total SY ($kg \cdot yr^{-1}$, ln = $\log_e$) versus ln-transformed lake area (AREA, ha) and volume (VOL, hm^3) (data from Christie and Regier 1988): (a and b) lake trout; (c and d) lake whitefish; (e and f) walleye; (g and h) northern pike.

error was assumed for the X variable; instead, a proportion of the X data were exposed to a large error distributed as $N(40,2)$. The GM estimates were closer to the true values than LS estimates when X contained outliers (Fig. 2 and 3), implying that GM was less sensitive to atypical errors in X than LS. However, the GM estimates still differed considerably from the true values. Our simulations show that atypical values in X and/or Y resulted in large biases in LS and GM estimation, indicating that the breakdown point is 0% for LS and GM. This means that inaccurate estimates of slope and intercept occur when any outliers are present. For robust estimation methods, LAV estimated the true values of parameters when only Y contained outliers and the number of outliers represented less than 35% of the data. However, LAV

TABLE 2. Multiple LS, LMS, and RLS regression analyses for species total SY ($kg \cdot yr^{-1}$) with the thermal habitat measures and the concentration of TDS ($mg \cdot L^{-1}$). The underlined parameter estimates are not significantly different from 0.

Lake trout	
LS: log(SY) = 4.05 + 0.87log(THV) − 0.69log(TDS),	$N = 15$, $r^2 = 0.90$
LMS: log(SY) = 3.24 + 0.80log(THV) − 0.42log(TDS),	$N = 15$, $r^2 = 0.93$
RLS: log(SY) = 3.27 + 0.77log(THV) − <u>0.42</u>log(TDS),	$N = 13$, $r^2 = 0.98$
Outliers: Lake of the Woods (7) and Lake Erie (12)	
Lake whitefish	
LS: log(SY) = 3.22 + 0.40log(THV) − 0.74log(TDS),	$N = 19$, $r^2 = 0.77$
LMS: log(SY) = 3.29 + 0.48log(THV) − 0.57log(TDS),	$N = 19$, $r^2 = 0.92$
RLS: log(SY) = 3.18 + 0.45log(THV) − 0.68log(TDS),	$N = 18$, $r^2 = 0.79$
Outlier: Little Peter Pond (6)	
Walleye	
LS: log(SY) = 3.15 + 0.88log(THA) − 0.08log(TDS),	$N = 19$, $r^2 = 0.72$
LMS: log(SY) = 3.03 + 1.21log(THA) − 0.36log(TDS),	$N = 19$, $r^2 = 0.84$
RLS: log(SY) = 2.84 + 1.09log(THA) − 0.18log(TDS),	$N = 16$, $r^2 = 0.92$
Outliers: Big Peter Pond (5), Lake Ontario (18), and Lake Nipigon (20)	
Northern pike	
LS: log(SY) = <u>1.25</u> + 0.43log(THA) + <u>1.30</u>log(TDS),	$N = 15$, $r^2 = 0.49$
LMS: log(SY) = 1.86 + 0.74log(THA) + 0.73log(TDS),	$N = 15$, $r^2 = 0.38$
RLS: log(SY) = <u>1.28</u> + 0.51log(THA) + 1.25log(TDS),	$N = 14$, $r^2 = 0.65$
Outlier: Lake Huron (14)	

was sensitive to outliers in X, and a few outliers in X resulted in great departures of the LAV estimates from the true values (Fig. 2 and 3). This implies that the breakdown point for LAV is similar to LS and GM. Differing from these three methods, LMS and RLS were not sensitive to outliers in X and/or Y. The estimates derived from these two methods were almost the same as the true values when the number of outliers represented less than 35% of the data, implying that the breakdown point is 35% for LMS and RLS with respect to the simulation data. These results are similar to those of Rousseeuw and Leroy (1987) who reported a breakdown point of 50%. However, the breakdown point of 35% defined in our study suggests that 50% is an overestimate for the defined simulation data. Because of the robustness of LMS and RLS with respect to outliers in X and/or Y, we suggest that researchers consider them when analyzing fisheries data.

Twelve lakes were defined as outliers in the first LMS analysis of fish species richness and lake area. Among these lakes, five were from Africa, five from North America, and two from the former USSR. All these lakes, except Zirahuen (31, Mexico) and Balkhash (37, Kazakhastan), were defined as outliers due to their exceptionally large number of fish species for their sizes (Fig. 4). The second LMS analysis indicated that four of these 12 lakes were outliers. Among these four outliers, Lakes Malawi (7) and Tanganyika (13) of Africa have exceptionally large numbers of fish species for their size. The exceptional age of these lakes relative to glaciated lakes of the Northern Hemisphere has also contributed to the large numbers of species present. However, Lake Zirahuen (31) and Balkhash (37) have exceptionally low numbers of fish species for their size. This may be due to their unique morphometry and geographic location: Zirahuen Lake (31) is located at a fairly high elevation on the faunistically depauperate Mesa Central of Mexico, and Lake Balkhash (37) lies in an interior basin of Asia and is partly saline and extremely shallow (Barbour and Brown 1974). The second RLS equation was estimated based on eight

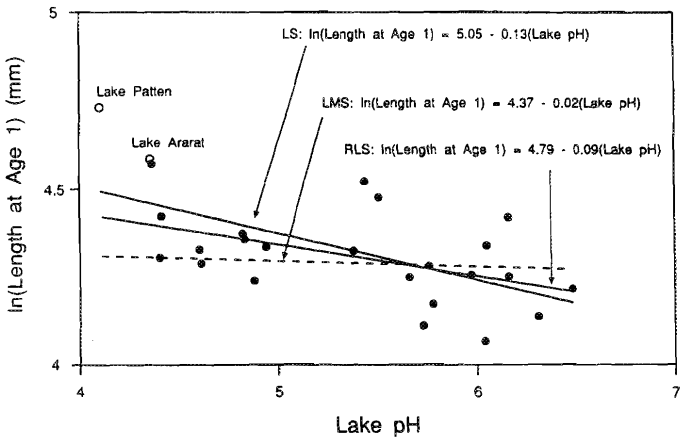

FIG. 7. Plot of ln-transformed length at age 1 of yellow perch against lake pH values (data from Ryan and Harvey 1980).

lakes, four from North America, three from Africa, and one from the former USSR. These lakes are among the largest lakes in each of the regions in this study (Barbour and Brown 1974). The regression slope derived from these lakes is twice as large as that derived from the rest of the lakes (Fig. 4), implying that these large lakes have increasingly greater numbers of fish species for their sizes relative to smaller lakes included in the first RLS analysis. This may result from the fact that there is a large number of lacustrine fish (i.e., cichlid "species flocks") in these three African lakes, and the four North American lakes are connected to each other (they are part of the Laurentian Great Lakes).

In the LMS analysis of log(SY) against log-transformed lake thermal or lake morphometric variables, the lakes identified as outliers are those having unique lake physical attributes. For lake trout, two lakes identified as outliers in all LMS analyses are Lake of the Woods (12) and Lake Erie (17) (Fig. 5a, 5b, 6a, and 6b). Both lakes are shallow and have high air temperatures relative to many of the other lakes. Such an environment is certainly not favourable for the

growth of lake trout which is a cold-water species and usually inhabits deep lakes. This may be the reason that SY's of lake trout in these two lakes were relatively lower than from other lakes. For lake whitefish, two levels of the log(SY)–log(THV) relationship can be established (Fig. 5c). The lakes in the high level of this relationship have either large mean depth and low average annual temperature (Great Slave Lake (1) and Big Peter Pond (5)) or much smaller THV than other lakes in the analysis, e.g., Little Peter Pond (6). For lake whitefish, which is a cold-water species, such an environment favours their growth, and hence higher productivity. However, we were unable to identify the reasons that the North Channel population had a high log(SY) for its log(THV). Georgian Bay was identified as an outlier because its log(SY) was significantly smaller for its log(THV) and log(THA) relative to other lakes in the LMS analysis (Fig. 5c and 5d). This may result from the fact that the lake whitefish fishery in Georgian Bay failed due to heavy commercial fishing (Cucin and Regier 1966). Without knowing the origin of the North Channel and Georgian Bay fish and whether they represent distinct stocks remaining in their own waters, it is conceivable that both locations are better considered a single population. No outliers were identified in the LMS analysis of log(SY) of lake whitefish against the lake morphometric variables (Fig. 6c and 6d), implying that all lakes included in the analysis had a similar log-transformed SY–morphometry relationship.

Walleye is a cool-water fish species and inhabits the shallow waters of these large northern lakes during most of the year (Scott and Crossman 1973). This may explain why Lake Erie (17) has significantly higher log(SY) for its log(area) and log(volume) (Fig. 6e and 6f) and Great Slave Lake (1) has significantly lower log(SY) for its log(THA) (Fig. 5f) relative to other lakes in the LMS analysis. However, we are unable to suggest reasons for other outliers. Three lakes identified as outliers in the LMS analysis of log(SY) against log(THV) or log(THA) have the three largest surface areas in this study (Fig. 5e and 5f). The reasons for this relationship remain unclear.

For both walleye and northern pike, the LS and RLS regression analysis between log(SY) and lake morphometric variables was not significant, implying that lake morphometric variables were not significant predictors of the SY of these two fish species. However, for two other species, lake morphometric variables tended to explain the majority of variance of log(SY) (Table 1). For all four species, lake thermal variables explained most of the variance of log(SY) in the regression analysis, and hence they were good predictors of lake SY of these four fish species. The RLS analyses tended to increase the values of r^2 greatly if there were outliers identified in the LMS analysis.

It is difficult to identify outliers in multiple regression analysis. However, as has been shown in the analysis of log(SY) versus THV or THA and TDS for the four fish species, the process of identifying outliers is rather straightforward using the LMS analysis. Similar explanations as those given in the simple linear analysis can be conferred on the lakes defined as outliers in the LMS multiple regression analysis. For lake trout, the LS-estimated parameter of log(TDS) differed significantly from 0. However, after excluding two outliers (lakes 12 and 17), the RLS parameter associated with log(TDS) was not significantly different from 0, indicating that TDS are not a significant factor in explaining the variance of log(SY). For northern pike, after excluding the outlier Lake Huron, the RLS-estimated parameter of log(TDS) became significant in contrast with the nonsignificant LS parameter.

Two lakes with the lowest pH values were defined as outliers in the LMS analysis of yellow perch size at age 1 versus lake pH (Fig. 7). Both lakes have larger positive residuals than other lakes in the LMS analysis, indicating that yellow perch inhabiting these two acidified lakes tended to be relatively larger at a young age compared with yellow perch from other lakes in the study. Yellow perch are among the most acid tolerant of the 30 species of fish found in the La Cloche lakes (Ryan and Harvey 1980). The loss of other fish species and increased mortality of yellow perch due to lake acidity likely reduced both interspecific and intraspecific competition for invertebrate food for young yellow perch (Ryan and Harvey 1980). In lakes with pH values of 4.35 (Lake Ararat) and 4.10 (Lake Patten), such competition may be significantly lower than in other lakes with higher pH values. The significance level of the RLS-estimated regression equation is much lower than that of LS ($p = 0.019$ for RLS versus $p < 0.0001$ for LS). The variation in log(size at age 1) explained by pH was much lower using RLS than using LS (24% of RLS versus 38% of LS). This indicates that pH may not be as important as reported in the original paper in explaining the variance of yellow perch length at age 1.

For biological reasons, a heterogeneous population of individuals can be regarded as mixtures of more homogeneous subpopulations (Lwin and Martin 1989). It is unreasonable to define a single model for such a heterogeneous population. Multiple functions should be used in defining such a population (e.g., the analysis of fish species richness and lake surface area in this study). Explanations of outliers in fisheries data may have significance in practice. There may be a misconception that the regression analysis for the data deleting outliers will be better than that for all the data in explaining the variance of the dependent variable. This is not true (e.g., when influential data points are defined as outliers in the LMS analysis, see Rousseeuw and Leroy 1987) as shown from the comparison of the RLS- and LS-estimated r^2 values for walleye in this study (Table 1). The RLS-estimated parameters may not necessarily have a more extreme probability value (i.e., significance level) than those estimated using the LS method. An example can be found in the multiple RLS analysis for lake trout (e.g., the LS-estimated parameter associated with log(TDS) differs significantly from 0, but the RLS-estimated parameter does not differ significantly from 0 at $p = 0.05$; Table 2).

In conclusion, we suggest using the two-step estimation procedure to analyze fisheries data: (1) applying the LMS method to identify outliers and (2) using the RLS to estimate parameters. The LMS-defined outliers should be studied separately with respect to the background information in the study (e.g., environmental conditions and fish biological characteristics). Such a study may identify the reason for outliers. In the case where there is a certain number of outliers, i.e., five or more, the two-step estimation procedures should be applied to these outliers. Thus, these data are modelled by multiple regression equations. When there are no outliers in data, there is no difference in parameter estimation using the LS- and LMS-based RLS methods. LMS cannot justify why a data point is an outlier unless biological and data collecting background information is available.

Acknowledgements

This study was supported financially by an NSERC operating grant to J.E.P. and NSERC postgraduate scholarship to Y.C., and an NSERC postdoctoral fellowship to D.A.J. We thank Dr. K. Somers for providing the computer program of LMS for this study. Constructive comments from Drs. W. Warren and P. Jolicoeur and an anonymous referee are greatly appreciated.

References

BARBOUR, C.D., AND J.H. BROWN. 1974. Fish species diversity in lakes. Am. Nat. 108: 473–489.

BLOOMFIELD, P., AND W.L. STEIGER. 1983. Least absolute deviations: theory, application, and algorithms. Birkhauser, Boston, Mass.

CHRISTIE, G.C., AND H.A. REGIER. 1988. Measures of optimal thermal habitat and their relationship to yields for four commercial fish species. Can. J. Fish. Aquat. Sci. 45: 301–314.

CUCIN, D., AND H.A. REGIER. 1966. Dynamics and exploitation of lake whitefish in southern Georgian Bay. J. Fish. Res. Board Can. 23: 221–274.

EDGEWORTH, F.Y. 1887. On observations relating to several quantities. Hermathena 6: 279–285.

EFRON, B. 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7: 1–26.

HOLLAND, P.H., AND R.E. WELSCH. 1977. Robust regression using iteratively reweighted least-squares. Commun. Stat.: Theory Methods 6: 813–827.

HUBER, P.J. 1973. Robust regression: asymptotics, conjectures, and Monte Carlo. Am. Stat. 1: 799–821.

JOLICOEUR, P. 1990. Bivariate allometry: interval estimation of the slopes of the ordinary and standardized major axes and structural relationship. J. Theor. Biol. 144: 275–285.

JOLICOEUR, P. 1991. Introduction à la biométrie. Éditions Masson, Paris, France.

KIMURA, D.K. 1989. Variability, tuning, and simulation for the Doubleday–Deriso catch-at-age model. Can. J. Fish. Aquat. Sci. 46: 941–949.

KIMURA, D.K. 1992. Symmetry and scale dependence in functional relationship regression. Syst. Biol. 41: 233–241.

LWIN, L., AND P.J. MARTIN. 1989. Probits of mixtures. Biometrics 45: 721–732.

MCARDLE, B.H. 1988. The structural relationship: regression in biology. Can. J. Zool. 66: 2329–2339.

RICKER, W.E. 1975. Computation and interpretation of biological statistics of fish populations. Bull. Fish. Res. Board Can. 191.

ROUSSEEUW, P.J. 1984. Least median of squares regression. J. Am. Stat. Assoc. 79: 871–880.

ROUSSEEUW, P.J., AND A.M. LEROY. 1987. Robust regression and outlier detection. John Wiley & Sons, Inc. New York, N.Y.

ROUSSEEUW, P.J., AND V. YOHAI. 1984. Robust regression by means of *s*-estimators, p. 256–272. *In* Robust and nonlinear time series analysis. J. Franke, W. Hardle, and R.D. Martin [ed.] Lecture notes in statistics No. 26. Springer Verlag, New York, N.Y.

RYAN, P.M., AND H.H. HARVEY. 1980. Factors accounting for variation in the growth of rock bass (*Ambloplites rupestris*) and yellow perch (*Perca flavescens*) in the acidifying LaClothe Mountain lakes of Ontario, Canada. Verh. Int. Ver. Limnol. 21: 1233–1237.

SCOTT, W.E., AND E.J. CROSSMAN. 1973. Freshwater fishes of Canada. Bull. Fish. Res. Board Can. 184.

SEN, A.K., AND M.S. SRIVASTAVA. 1990. Regression analysis: theory, methods and applications. Springer-Verlag, New York, N.Y.