Reprinted from

Canadian Journal of Fisheries and Aquatic Sciences

Réimpression du

Journal canadien des sciences halieutiques et aquatiques

Weighted multicollinearity in logistic regression: diagnostics and biased estimation techniques with an example from lake acidification

B. D. Marx and E. P. Smith

Volume 47 • Number 6 • 1990

Pages 1128–1135

Canada

Fisheries and Oceans Pêches et Océans

Printed in Canada by The Runge Press Limited
Imprimé au Canada par The Runge Press Limited

Weighted Multicollinearity in Logistic Regression: Diagnostics and Biased Estimation Techniques with an Example from Lake Acidification

Brian D. Marx[1]

Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803-5606 USA

and Eric P. Smith

Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA

Marx, B. D., and E. P. Smith. 1990. Weighted multicollinearity in logistic regression: diagnostics and biased estimation techniques with an example from lake acidification. Can. J. Fish. Aquat. Sci. 47: 1128–1135.

An historical data set from the Adirondack region of New York is revisited to study the relationship between water chemistry variables associated with acid precipitation and the presence/absence of brook trout (*Salvelinus fontinalis*) and lake trout (*Salvelinus namaycush*). For the trout species data sets, water chemistry variables associated with acid precipitation, for example pH and alkalinity, are highly correlated. Regression models to assess their effects on the probability of the presence of fish species are therefore affected by multicollinearity. Because the appropriate regressions are logistic, correction techniques based on least squares do not work. Maximum likelihood parameter estimation is highly unstable for the trout presence/absence data. Developments in weighted multicollinearity diagnostics are used to evaluate maximum likelihood logistic regression parameter estimates. Further, an application of biased parameter estimation is presented as an option to the traditional maximum likelihood logistic regression. Biased estimation methods, like ridge, principal component, or Stein estimation can substantially reduce the variance of the parameter estimates and prediction variance for certain future observations. In many cases, only a slight modification to the converged maximum likelihood estimator is necessary.

Un ensemble de données chronologiques pour la région des Adirondacks de l'État de New York est réexaminé pour étudier la relation entre les variables de la chimie de l'eau associées aux pluies acides et la présence/l'absence de l'omble de fontaine (*Salvelinus fontinalis*) et du touladi (*Salvelinus namaycush*). Dans le cas des ensembles de données sur les espèces de truite, les paramètres de la chimie de l'eau associés aux précipitations acides, comme le pH et l'alcalinité, sont fortement corrélés. La multicolinéarité influe donc sur les modèles de régression pour l'évaluation des effets sur la probabilité de la présence d'espèces de poisson. Les techniques de correction basées sur la méthode des moindres carrés ne fonctionnent pas car les régressions appropriées sont logistiques. L'estimation du paramètre de maximum de vraisemblance est très instable dans le cas des données sur la présence/l'absence de truite. Les diagnostics de multicolinéarité pondérée sont développés pour évaluer les estimations du paramètre de maximum de vraisemblance de la régression logistique. De plus, une application de l'estimation d'un paramètre biaisé est présentée à titre de solution de rechange à la régression logistique conventionnelle du maximum de vraisemblance. Les méthodes d'estimation biaisée, comme l'estimation par crête, par composante principale ou de Stein, peuvent énormément réduire la variance des estimations du paramètre et des prédictions dans le cas de certaines observations futures. Dans de nombreux cas, il suffit d'apporter une légère modification à l'estimateur convergent du maximum de vraisemblance.

Received July 5, 1989
Accepted January 12, 1990
(JA216)

Reçu le 5 juillet 1989
Accepté le 12 janvier 1990

There has recently been international concern and effort in estimating the probability of presence/absence of various fish species from models involving water chemistry variables responsible for lake acidification (Christensen et al. 1988; Reckhow et al. 1987; Baker 1984; Magnuson et al. 1984; Howells 1983; Haines 1981). Estimating and interpretating how the probability of presence depends on attributes of water chemistry which can have a great impact on efforts to prevent the loss of ecologically and economically important species.

The analysis of data on the presence or absence of fish species is different from ordinary regression analysis for two reasons. First, the response variable (presence or absense) is a binary variable which is one of two values and hence is not normal. The approach typically used to model this type of data is logistic regression. In logistic regression, the probability of presence is modelled as a logistic response function which depends on the water chemistry variables and other explanatory variables. Estimation of the parameters for the model is done via an extension of the least squares method called iterative reweighted least squares.

The second difference involved multicollinearities. The water chemistry variables associated with acid precipitation, such as pH and alkalinity, are often highly correlated. pH measures the concentration of hydrogen ion in the water while alkalinity is a measure of acid neutralizing capacity of the water. Alkalinity is dependent on pH but also reflects buffering capacity of the

[1]Author to whom correspondences should be addressed.

water system. Thus there is a strong but not perfect relationship between these variables. Linear or near linear dependencies between the explanatory variables are called multicollinearities and can adversely affect the results of regression analysis. Management strategies which attempt to control one physical–chemical variable based on a regression analysis may be misled as to the effect of controlling that variable. Either the joint effets of variables need to be considered or the effect of a single variable needs to be better estimated in the presence of other variables. While there is much information on the effects, diagnosis, and adjustment of multicollinearity in multiple regression analysis, little work has been done for the logistic regression problem.

Recent advances in the statistical literature (Schaefer 1986; Marx and Smith 1989) offer improved ways to analyze data resulting from studies concerned with the presence/absence of fish species due to changes in water chemistry. It is the objective of this paper to describe weighted multicollinearity diagnostics for logistic regression. These diagnostics can assist in identifying variables to be considered for potential deletion from the model. Further, biased estimation techniques are presented as an option to traditional maximum likelihood logistic regression. Ridge logistic, principal component logistic, and Stein logistic models will be formulated and interpreted for brook trout (*Salvelinus fontinalis*) and lake trout (*Salvelinus namaycush*) from the Adirondack region of New York. These biased estimation techniques can be particularly useful when theoretical models of interest have severe, or even moderate, multicollinearity problems. The techniques are applied to the data of Reckhow et al. (1987). Reckhow et al. described the problems of multicollinearity and presented alternatives to ordinary logistic regression. Their methods however, were based on the techniques used for multiple regression analysis and did not take into account the special structure of the data. The methods presented here lead to improved results.

Brief Overview of Logistic Regression

Hosmer and Lemeshow (1989) provides an excellent overview of logistic regression. Logistic regression is commonly used to model the probability of a dichotomous outcome when given explanatory variables of interest. Consider response data that is binary in nature, such as presence/absence of trout, and suppose there are N observations and p explanatory variables. Define the $N \times (p + 1)$ matrix, $X = (1\ x_1 \ldots x_p)$. The x_j are continuous explanatory variables. Denote x_i', of dimension $1 \times (p + 1)$, as a row vector of X. Let β be the $(p + 1) \times 1$ vector of unknown regression parameters. The response variable is the probability of presence which depends on the explanatory variables. This probability that a fish species is found in lake i is defined as the conditional probability, $\pi_i = P(Y_i = 1 | x_i')$. The response probability is then modelled as a sigmoidal curve which depends on the explanatory variables measured in the lake. The logit or log odds ratio model uses the cumulative logistic density which ensures the predicted probability is between 0 and 1. Other cumulative densities can be used to form an appropriate model, but are not considered in this paper. Given independent Bernoulli response data $Y_i (i = 1,2, \ldots, N)$, let Y be the $N \times 1$ binary response vector consisting of ones and zeros. The logistic regression model can be formulated,

$$(1) \quad \text{logit}(\pi_i) = \ln\{\pi_i (1 - \pi_i)^{-1}\} = x_i' \beta$$

$$(2) \quad \text{or} \quad \pi_i = \{1 + \exp(-x_i' \beta)\}^{-1}.$$

Because the distribution of Y is neither reasonably continuous nor symmetric, ordinary least squares is not an appropriate estimation technique and iterative maximum likelihood is commonly used to estimate the unknown regression parameters. The iterative maximum likelihood scheme for logistic regression can be expressed as

$$(3) \quad \hat{\beta}_t = \hat{\beta}_{t-1} + (X' \hat{V}_{t-1} X)^{-1} X' (y - \hat{\pi}_{t-1}),$$

where t denotes the iteration step, $\hat{V} = \text{diag}\{\hat{\pi}_i (1 - \hat{\pi}_i)\}$, and $\hat{\pi}$ is the vector of predicted probabilities. Note that the procedure is a weighted procedure, and that instead of using $X'X$, one uses $X'VX$. The diagonal matrix V contains variances of the estimated Y values. The matrix $\Phi = X'VX$ is called the information matrix. Denote $\hat{\Phi} = X' \hat{V} X$ as estimated information. An ill-conditioned information matrix can result in undesirable asymptotic properties of the logistic regression, such as large variances associated with parameter estimates and certain prediction regions. Analogous to what is done in multile regression, the effects of multicollinearity can be assessed by decomposing the information matrix into orthogonal components. Let G be the orthogonal matrix such that $G' \hat{\Phi} G = \Lambda$, where $\Lambda = \text{diag}\{\lambda_j\}$ and λ_j are the eigenvalues of $\hat{\Phi}$. Define $Z = XG$ and z_u' as a row of Z. As sample size gets large, the variance of the coefficients, $\text{Var}(\beta) \cong \hat{\Phi}^{-1}$. If $\hat{\Phi}$ has a small λ_j, then some examples of undesirable properties of maximum likelihood estimation include:

(a) extreme sensitivity of parameter estimates to small perturbations in explanatory variables.
(b) The sum of the variances of the coefficients tends to infinity.
(c) For certain future observation data vectors x_u', the prediction variance is approximately
$\text{Var}(\hat{\pi}_u) \cong \{\pi_u (1 - \pi_u)\}^2 \Sigma_{j=0}^{p} z_{u,j}^2 \lambda_j^{-1} \to \infty$.
(d) The test:
$H_0 : \beta = \beta_c$
$H_1 : \beta = \beta_f$,
may have low power, where c and f denote the current and full model, respectively (c is typically a subset of f).

The problems above resemble those of least squares multicollinearity. Multicollinearity results when one explanatory variable can be nearly expressed as linear combination of the remaining explanatory variables. Redundancies in the regressors yields multicollinearity. With multicollinearity in a standard multiple regression, the least squares estimation technique is not exposed to a regressor variable data structure that it truly needs to produce clear estimates of rate of change on the response variable (Myers 1990). Traditionally, with least squares estimation in standard multiple regression, spectral decomposition of the correlation matrix of explanatory variables has been used as a diagnostic tool to determine the ill-effects of multicollinearity on parameter estimation and prediction. Researchers have been unjustifiably utilizing these least squares multicollinearity diagnostics for models not based on least squares, logistic regression is one example. Many of the assumptions of least squares linear regression are not met. Certainly a response which is a 0 or a 1 fails to meet normality of error terms. Furthermore, Bernoulli variances are heteroscedastic.Theoretical investigation (Marx and Smith 1989) suggested that logistic regression diagnostics should be oriented toward the spectral decomposition of $X' \hat{V} X$, where $\hat{V} = \text{diag}\{\hat{\pi}_i (1 - \hat{\pi}_i\}$ is a diagonal matrix of Bernoulli variances. Multicollinearity diagnostics can be formulated among the weighted explanatory variables, i.e. the columns of $\hat{S} = \hat{V}^{1/2} X$.

These diagnostics often yield more pertinent information for selection of a candidate model, in that they can be used as a variable selection tool.

Weighted Multicollinearity Diagnostics for Logistic Regression

A matrix $\hat{\Phi}$ is ill-conditioned if it has nearly linearly dependent columns of $\hat{S} = \hat{V}^{1/2}X$. For X with full rank, notice that the strength of the linear dependence among the columns of X does not directly affect the condition of $\hat{\Phi}$ unless the diagonal matrix $\hat{V}^{1/2} \cong kI$. In such a special case, not only are the binary Y_i independent, but they further have nearly homogeneous variance. Diagnostics should be developed accordingly. In developing suitable diagnostics, scaling of the information matrix is preferred in order to have a standard for comparison (Belsley et al. 1980). A natural scaling method which gives the columns of $\hat{S}$ unit length is given in equation (4). Let

$$(4) \quad \hat{S}^*(i,j) = (\hat{S}_{ij} - \bar{S}_j) \left[\sum_{i=1}^{N} (\hat{S}_{ij} - \bar{S}_j)^2 \right]^{-1/2}.$$

Define the weighted correlation matrix $\hat{\Phi}^* = \hat{S}^{*\prime}\hat{S}^*$. *The* (i,j)th entry of $\hat{\Phi}^*$ yields the correlation between the ith and jth column of $\hat{S}$.

In assessing the degree of ill-conditioning of a general matrix B, Hartree (1952) pointed out the importance of a relative measure of ill-conditioning called a condition number. Belsley et al. (1980) argued that if a small eigenvalue is used as a multicollinearity diagnostic, then there is a natural tendency to compare small with the wrong standard, namely zero. Hence, the ratios of functions of eigenvalues, rather than the eigenvalues alone, have been useful diagnostics. Condition numbers are defined for the purpose of this paper as the ratio of the largest to the jth smallest eigenvalue of Φ^*,

$$(5) \quad \Psi_j = (\lambda^*_{\max}/\lambda^*_j)^{1/2},$$

where the λ^*_j were the ordered eigenvalues of $\hat{\Phi}^*$. Large values of Ψ_j (≥ 30) indicate ill-conditioning.

Weighted Variance Inflation

For least squares estimation, variance inflation factors (VIFs) indicate the inflation of variance above the ideal, i.e. the correlation matrix of explanatory variables being the identity matrix. Constructing such a diagnostic for logistic regression is a bit more complicated. For one, the covariance matrix for $\hat{\beta}$ is not a scalar multiple of $X'X$. Perhaps the most obvious construction of weighted variance inflation factors (WVIFs) for logistic regression is to consider the diagonal elements of $\hat{\Phi}^{*-1}$. Under ideal conditions, i.e. when the columns of $\hat{S}$ are orthogonal, $\hat{\Phi}^* = \hat{\Phi}^{*-1} = I$ and all $\Psi_j = 1$. Marx and Smith (1989) quantified a measure of inflation of asymptotic variance due to the nonorthogonality among the columns of $\hat{S}$. Define

$$(6) \quad \text{WVIF}_j = j\text{th diagonal element of } \hat{\Phi}^{*-1},$$

as a weighted variance inflation factor. Observe that there is WVIF associated with the intercept since the centering and scaling of $\hat{S}$ does not involve a constant column of ones, but rather $\hat{V}^{1/2}1$. Notice for the special case of $V = I$ and regression through the origin, WVIF reduces to a least squares VIF. A WVIF gives the inflation of variance above the standard of one under ideal conditions.

Weighted Variance Proportion

Once the WVIFs are calculated, it also can be informative to determine to what extent the proportion of variance of each coefficient is attributed to each near dependency in the information matrix. Weighted variance proportion decomposition requires the spectral decomposition of $\hat{\Phi}^*$. Let M be the orthogonal matrix such that $M'\hat{\Phi}^*M = \Lambda^*$, where $\Lambda^* = \text{diag}\{\lambda^*_j$ and λ^*_j are the eigenvalues of $\hat{\Phi}^*$. Using the asymptotic correlation form covariance matrix for the converged maximum likelihood parameter estimates, define

$$(7) \quad C_{jj} = \sum_{u=0}^{p} m^2_{ju}/\lambda^*_u.$$

A small eigenvalue has influence, to some degree, on all variances. The weighted proportion of variance for the jth estimated coefficient, attributed to uth eigenvalue of the sum in equation (7), can be expressed as

$$(8) \quad \rho_{uj} = \frac{m^2_{ju}/\lambda^*_u}{C_{jj}}.$$

Weighted variance proportion decomposition

Ordered Eivengalue	Proportion of Var $(\hat{\beta}_0)$	Var $(\hat{\beta}_1)$	...	Var $(\hat{\beta}_p)$
λ_0	ρ_{00}	ρ_{01}	...	ρ_{0p}
λ_1	ρ_{10}	ρ_{11}	...	ρ_{1p}
.	.	.		.
.	.	.		.
.	.	.		.
λ_p	ρ_{p0}	ρ_{p1}	...	ρ_{pp}

A matrix of proportions can be constructed. Hence, a small eigenvalue (relative to the maximum eigenvalue) responsible for at least two large proportions suggests that weighted multicollinearity is damaging desirable properties of the logistic regression. Action should be taken accordingly. If a large WVIF is present, then the researcher should consider deletion of one of the variables responsible for a large ρ_{uj} (associated with a small eigenvalue) or an alternative estimation technique, for example, a biased logistic estimation technique (Schaefer 1986). If specific theoretical models are not of interest, then perhaps wary variable deletion based on the above diagnostics can satisfy the researcher's needs for a predictive model. However, in instances where a coefficient estimate is of primary concern, variable deletion may not be reasonable.

To illustrate the effects of the ill-conditioned information matrix, we consider the models in Table 1 which were presented in Reckhow et al. (1987). Using brook trout presence/absence data from Adirondack lakes, Table 1 provides a model uses explanatory variables: lake pH, calcium content (log transformed), and the interaction between these two variables. Table 1 also provides a model for presence/absence of lake trout. Explanatory variables used are lake pH, alkalinity (log transformed), and the interaction term. Reckhow et al. (1987) provide plots justifying the importance and interpretation of the interaction terms for both species of trout. Table 1 provides a deviance goodness-of-fit measure. The deviance measures how well the model fits the data and is defined as $D = \Sigma_{i=1}^{N} d_i^2$, where $d_i^2 = -2\ln(1 - \hat{\pi}_i)$ for $y = 0$ and $d_i^2 = -2\ln(\hat{\pi}_i)$ for $y = 1$.

TABLE 1. Maximum likelihood estimates, standard errors, and deviance for Reckhow et al. (1987) trout data.

	$\hat{\beta}$	SE($\hat{\beta}$)	D
Brook trout ($N=46$)			
Intercept	−49.8460	26.4290	43.2936
pH	9.1906	3.9094	
In(Ca)	8.0101	5.9051	
pH*In(Ca)	−1.4885	0.8249	
Lake trout ($N=32$)			
Intercept	−233.4000	127.6000	34.4160
pH	32.3834	16.6454	
In(Alk)[a]	45.6460	26.0854	
pH*In(Alk)	−6.3188	3.4195	

[a]Alkalinity shifted + 150 to create all positive values.

TABLE 2. Logistic weighted muticollinearity diagnostics for brook trout data ($N=46$).

Correlation matrix of weighted regressors

	Intercept	pH	In(Ca)	pH*In(Ca)
Intercept	1.0000	0.8084	0.8794	0.6326
pH	0.8084	1.0000	0.9615	0.9554
In(Ca)	0.8794	0.9615	1.0000	0.9138
pH*In(Ca)	0.6326	0.9554	0.9138	1.0000

Weighted variance inflation factors

Intercept	235.5
pH	292.8
In(Ca)	384.0
pH*In(Ca)	564.5

Variance proportion decomposition

E-value	Ψ_j	Intercept	pH	In(Ca)	pH*In(Ca)
3.5846	1.0000	2.5E-04	2.6E-04	2.0E-04	1.2E-04
0.3857	3.0484	0.0067	3.0E-04	1.6E-05	0.0016
0.0290	11.1146	0.0018	0.0593	0.0419	9.9E-04
6.9E-04	71.8325	0.9912	0.9401	0.9578	0.9973

Asymptotic arguments suggest that D has a limiting (with sample size) χ^2_{N-p-1} distribution.

Although the fit of these above models seems reasonable, there are some disturbing problems. The magnitude and standard errors of the coefficients for these models are inflated relative to models containing only one variable. For example, when the presence of lake trout is modelled in terms of pH alone, the estimated coefficient is 0.923 (SE = 0.457), while the three variable model results in a coefficient of 32.38 (SE = 16.65). Further evidence of problems are given by the weighted multicollinearity diagnostics for the brook trout data (Table 2) and the lake trout data (Table 3). Notice the large positive correlations among the weighted regressors yielding extremely large weighted variance inflation factors, especially in the lake trout data. For both data sets, we find a large condition number (≥30) with the smallest eigenvalue of the information matrix. The decomposition matrix has a last row of values all nearly one indicating a redundancy among the explanatory variables. Certainly an option is to drop the interaction term from both models. Even though this may be a reasonable approach, the coefficient of interaction may be of interest. As an option to variable deletion, we consider a variety of biased estimation techniques.

TABLE 3. Logistic weighted muticollinearity diagnostics for lake trout data ($N=32$).

Correlation matrix of weighted regressors

	Intercept	pH	In(Ca)	pH*In(Alk)
Intercept	1.0000	0.8441	0.9423	0.7406
pH	0.8441	1.0000	0.9525	0.9781
In(Alk)	0.9423	0.9525	1.0000	0.9125
pH*In(Alk)	0.7406	0.9781	0.9125	1.0000

Weighted variance inflation factors

Intercept	3180.3
pH	3972.7
In(Ca)	4544.9
pH*In(ca)	6580.2

Variance proportion decomposition

E-value	Ψ_j	Intercept	pH	In(Alk	pH*In(Ca)
3.6885	1.0000	1.9E-05	1.8E-05	1.6E-05	1.0E-05
0.2920	3.5541	5.8E-04	7.0E-05	2.1E-05	1.8E-04
0.0194	13.7845	9.4E-04	0.00569	0.00518	3.6E-04
5.5E-05	259.3000	0.99850	0.99420	0.99480	0.9994

Biased Logistic Regression Estimators

Using Taylor series arguments, it can be shown that the maximum likelihood parameter estimates are asymptotically unbiased. In making certain adjustments to maximum likelihood, asymptotically biased parameter estimates can be constructed. Ridge, principal component, and Stein asymptotically biased estimators are presented in this paper. There is much controversy regarding centering and scaling of explanatory variables (i.e. adjusting each variable to mean zero and unit length). In presenting ridge and principal component estimators below, centering and scaling is done. This is consistent with Myers (1990) and Schaefer (1979, 1986). The notation becomes a bit more complicated. However, there is a connection between the centered and scaled estimated coefficients and the ones in the natural units. Let $X_p=(x_1 \ldots x_p)$. Define the (i,j)th element of X_p^*

$$(9)\quad X_p^*(i,j)=q_j^{-1}(X_{ij}-\bar{X}_j),$$

where $q_j = \{\Sigma_{i=1}^N (X_{ij}-\bar{X}_j)^2\}^{1/2}$. Notice that $X_p^{*\prime}X_p^*$ is the $p\times p$ correlation matrix for the explanatory variables. Augment the matrix X_p^* with the constant column of ones and denote $X^*=(1\ X_p^*)$.

In using X^* as the data matrix, let $\tilde{\beta}$ be the converged maximum likelihood estimate for unknown β^* and $\tilde{V}$ be the corresponding estimated diagonal matrix of Bernoulli variances. Denote $\tilde{\Phi}=X^{*\prime}\tilde{V}X^*$. F is the matrix of eigenvectors and the γ_j are the corresponding of eigenvalues of $\tilde{\Phi}$.

Certainly similar derivations for biased estimators can be made for uncentered and unscaled explanatory variables. In addition, if by design the explanatory variables are measured in the same units, then further standardization may not be needed. It is not this paper's objective to discuss alternate standardizations.

Ridge Logistic Regression Estimators

Schaefer (1979, 1986) suggested

$$(10)\quad \tilde{\beta}_R(k)=(X^{*\prime}\tilde{V}X^*+kI)^{-1}X^{*\prime}\tilde{V}X^*\tilde{\beta}$$
$$= \tilde{\Phi}_k^{-1}\tilde{\Phi}\tilde{\beta}$$

where $k>0$ is called the shrinkage or ridge parameter. Note that this is a simple adjustment to the maximum likelihood estimator. For $k=0$, $\tilde{\beta}_R$ is a maximum likelihood estimator of β^*. Upon choice of k, uncentering and unscaling the estimated coefficients to the natural units can be done. Define

$$(11)\quad \hat{\beta}_{R,j}=q_j^{-1}\tilde{\beta}_{R,j} \text{ for } j=1, 2, \ldots p$$

$$(12)\quad \text{and } \hat{\beta}_{R,0}=\tilde{\beta}_{R,0} - \sum_{j=1}^{p} q_j^{-1}\bar{X}_j\tilde{\beta}_{R,j}.$$

The asymptotic covariance matrix for $\tilde{\beta}(k)$ is $\tilde{\Phi}_k^{-1}\tilde{\Phi}\tilde{\Phi}_k^{-1}$. The asymptotic standard errors associated with the uncentered and unscaled ridge estimates can be expressed as

$$(13)\quad \text{SE}(\hat{\beta}_{R,j})=q_j^{-1}\text{SE}(\tilde{\beta}_{R,j}) \text{ for } j=1, 2, \ldots p$$

$$(14)\quad \text{SE}(\hat{\beta}_{R,0})=\{\text{Var}(\tilde{\beta}_{R,0}) + \sum_{j=1}^{p} (q_j^{-1}\bar{X}_j)^2\text{Var}(\tilde{\beta}_{R,j})$$

$$+2\sum_{i<j}\sum_{j\neq 0} q_j^{-1}q_j^{-1}\bar{X}_i\bar{X}_j\text{Cov}(\tilde{\beta}_{R,i},\tilde{\beta}_{R,j})$$

$$-2\sum_{j=1}^{p} q_j^{-1}\bar{X}_j\text{Cov}(\tilde{\beta}_{R,0},\tilde{\beta}_{R,j})\}^{1/2}.$$

The choice of k is subjective, however, Schaefer recommended a harmonic mean method

$$(15)\quad k=(p+1)/(\tilde{\beta}'\,\tilde{\beta}),$$

which is quite conservative (relatively small) under extreme ill-conditioning of $\hat{\Phi}$. Marx (1988) has also suggested a ridge trace approach (ridge coefficient estimates versus k) and other approaches for the generalized linear model. Only the ridge trace are presented in the trout examples.

TABLE 4. Ridge trace parameter estimates, deviance, and associated standard errors as a function of shrinkage for brook trout data ($N=46$).

k	$\hat{\beta}_{R,0}$	$\hat{\beta}_{R,1}$	$\hat{\beta}_{R,2}$	$\hat{\beta}_{R,3}$	D
0.0000	−49.846	9.19056	8.0101	−1.4885	43.2936
0.0005	−18.787	4.59572	1.3115	−0.5099	45.0160
0.0010	−11.965	3.53317	−0.1081	−0.2927	46.0002
0.0015	−9.019	3.03896	−0.6867	−0.1973	46.5424
0.0020	−7.402	2.74195	−0.9791	−0.1439	46.8974
0.0025	−6.395	2.53725	−1.1418	−0.1098	47.1588
0.0030	−5.719	2.38370	−1.2356	−0.0861	47.3674
0.0035	−5.239	2.26180	−1.2893	−0.0689	47.5436
0.0040	−4.887	2.16107	−1.3178	−0.0557	47.6985
0.0045	−4.620	2.07535	−1.3300	−0.0453	47.8387
0.0050	−4.414	2.00079	−1.3311	−0.0370	47.9682

k	$\text{SE}(\hat{\beta}_{R,0})$	$(\hat{\beta}_{R,1})$	$\text{SE}(\hat{\beta}_{R,2})$	$\text{SE}(\hat{\beta}_{R,3})$
0.0000	26.4290	3.9094	5.9051	0.8249
0.0005	9.9936	1.6018	2.5119	0.3007
0.0010	6.5147	1.1474	1.8643	0.1843
0.0015	5.0815	0.9668	1.6094	0.1332
0.0020	4.3309	0.8706	1.4720	0.1046
0.0025	3.8820	0.8093	1.3825	0.0864
0.0030	3.5888	0.7653	1.3165	0.0738
0.0035	3.3848	0.7310	1.2639	0.0645
0.0040	3.2347	0.7028	1.2196	0.0575
0.0045	3.1204	0.6787	1.1810	0.0520
0.0050	3.0304	0.6574	1.1465	0.0476

One-Step Principal Component Logistic Regression Estimators

Schaefer (1986) developed a principal component estimator for logistic regression of the form

$$(16)\quad \tilde{\beta}_{pc}^{s}=(X^{*\prime}\tilde{V}\,X^*)^+\,(X^{*\prime}\tilde{V}\,X^*)\,\tilde{\beta}$$
$$=\tilde{\Phi}^+\,\tilde{\Phi}\,\tilde{\beta},$$

where

$$\tilde{\Phi}^+=\sum_{j=0}^{p-r}\gamma_j^{-1}f_jf_j'.$$

Note that $r=p+1-s$ is the number of components deleted. The γ_j are the eigenvalues of $\tilde{\Phi}$ (usually in descending order) and f_j are the corresponding eigenvectors. A conversion of $\tilde{\beta}_{pc}^{s}$ to $\hat{\beta}_{pc}^{s}$, in the natural units, can be made using equations similar to equations (11)–(14).

Iterative Principal Component Logistic Regression Estimators

Marx and Smith (1990) developed an alternate iterative principal component estimator which behaves very closely to Schaefer's one step adjustment to maximum likelihood above in equation (16) in the framework of logistic regression. The iterative method considers fitting the model to the principal components, i.e.

$$(17)\quad \text{logit}(\pi^*)=Z^*\,\alpha^*,$$

where $Z^*=X^*F$ and $\alpha^*=F'\beta^*$. The iterative scheme is employed in a specified subset α_s^* and $\alpha_r^*=0$. The iterative equation becomes

$$(18)\quad \tilde{\alpha}_{s,t}^{pc}=\tilde{\alpha}_{s,t-1}^{pc}+\Gamma_{s,t-1}^{-1}Z^{*\prime}(y-\tilde{\pi}^{pc}),$$

where t denotes the iterative step until convergence and Γ_s is the diagonal matrix of subset eigenvalues of $\tilde{\Phi}$, which are not deleted. Upon convergence of $\tilde{\alpha}_s^{pc}$, an estimate of β^* is given by

$$(19)\quad \tilde{\beta}_s^{pc}=F\,\tilde{\alpha}_s^{pc}.$$

Again, a conversion can be made to the natural units, using an approach similar to that given in equations (11)–(12), to yield $\hat{\beta}_s^{pc}$. Asymptotically, the standard errors for the one-step and the iterative principal component techniques are the same. Noting that the estimated asymptotic covariance matrix for $\hat{\beta}_s^{pc}$ and $\tilde{\beta}_{pc}^{s}$ is $F_s\Gamma_s^{-1}F_s'$, an equation similar to equations (13)–(14) can be used to obtain the uncentered and unscaled standard errors.

Stein Logistic Regression Estimators

Schaefer (1986) suggested an extension of the Stein (1960) estimator for logistic regression. Consider shrinking the maximum likelihood estimate as follows

$$(20)\quad \hat{\beta}_s=c\,\hat{\beta},$$

where $0<c<1$. The motivation of Stein estimation is to shrink both the estimated parameter vector, and associated standard errors, by a simple scaling technique. One choice of c, which minimizes the $E(L_1^2)=(c\hat{\beta}-\beta)'(c\hat{\beta}-\beta)$ criterion (with respect to c), is

$$(21)\quad c=(\hat{\beta}'\hat{\beta})/\{\hat{\beta}'\hat{\beta}+\text{trace}(\hat{\Phi}^{-1})\}.$$

The standard error of a Stein estimator is c times the standard error of the maximum likelihood estimator.

TABLE 5. Ridge trace parameter estimates, deviance, and associated standard errors as a function of shrinkage for lake trout data ($N=32$).

k	$\hat{\beta}_{R,0}$	$\hat{\beta}_{R,1}$	$\hat{\beta}_{R,2}$	$\hat{\beta}_{R,3}$	D
0.00000	−233.35	32.3834	45.6460	−6.3188	34.4160
0.00025	−27.74	5.5119	3.7373	−0.7934	37.5601
0.00050	−13.16	3.5814	0.7797	−0.3992	38.0683
0.00075	−7.89	2.8670	−0.2797	−0.2552	38.2677
0.00100	−5.20	2.4906	−0.8126	−0.1807	38.3757
0.00125	−3.59	2.2557	−1.1258	−0.1352	38.4447
0.00150	−2.53	2.0933	−1.3268	−0.1046	38.4936
0.00175	−1.79	1.9732	−1.4629	−0.0826	38.5308
0.00200	−1.26	1.8799	−1.5583	−0.0661	38.5606
0.00225	−0.86	1.8046	−1.6265	−0.0532	38.5855
0.00250	−0.55	1.7422	−1.6757	−0.0430	38.6070

k	SE($\hat{\beta}_{R,0}$)	($\hat{\beta}_{R,1}$)	SE($\hat{\beta}_{R,2}$)	SE($\hat{\beta}_{R,3}$)
0.00000	127.6000	16.6454	26.0852	3.4195
0.00025	18.8117	2.3948	4.3011	0.4566
0.00050	12.1473	1.4899	3.1345	0.2459
0.00075	10.1183	1.2046	2.8026	0.1693
0.00100	9.2191	1.0753	2.6537	0.1300
0.00125	8.7276	1.0036	2.5669	0.1062
0.00150	8.4176	0.9581	2.5069	0.0903
0.00175	8.2006	0.9261	2.4604	0.0791
0.00200	8.0361	0.9019	2.4216	0.0707
0.00225	7.9039	0.8825	2.3876	0.0643
0.00250	7.7927	0.8662	2.3568	0.0592

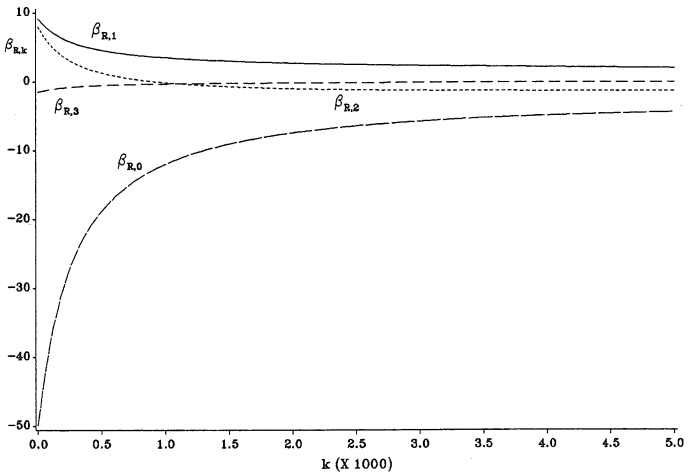

FIG. 1. Ridge trace plot for brook trout estimated regression coefficients as a function of shrinkage.

Application of Methodologies

The above methodologies were applied to the lake chemistry data sets for brook trout and lake trout. Tables 4 and 5 give ridge estimates of the parameters for the two models for different values of the ridge parameter (k). Also, standard errors are provided. Even with small k (0.001), the ridge method is effective in reducing the magnitude of the coefficients and their standard errors, but the sign of the coefficient for the second explanatory varaible, $\hat{\beta}_{R,2}$, differs from the univariate model which only includes the second explanatory variable. Ridge trace plots are given in Fig. 1 and 2 and demonstrate the instability of maximum likelihood estimates. These plots indicate rapid change in the coefficients until k is 0.001. Values near this would represent reasonable values for the ridge coefficient. The trout examples coefficients have been transformed to their

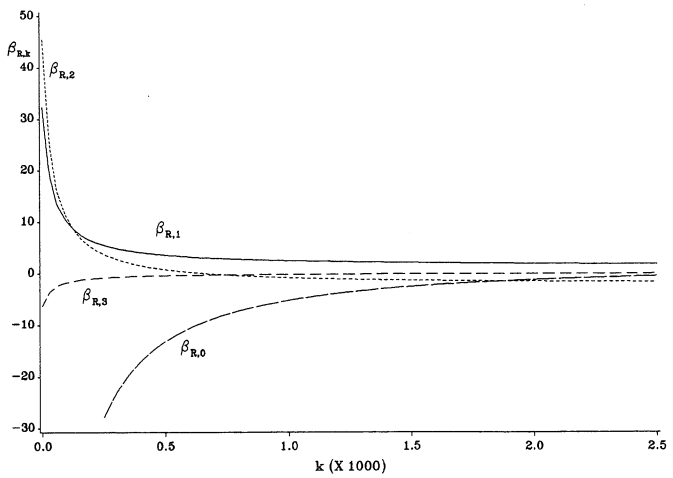

FIG. 2. Ridge trace plot for lake trout estimated regression coefficients as a function of shrinkage.

TABLE 6. Iterative principal component and Stein estimates and associated asymptotic standard errors for brook trout data ($N=46$).

	pc(-1)	pc(-2)	Stein($c=0.78$)
$\hat{\beta}_0$	−0.5382	−5.4530	−38.8111
$\hat{\beta}_1$	2.2315	0.2920	7.1560
$\hat{\beta}_2$	−2.9793	0.5514	6.2369
$\hat{\beta}_3$	0.0528	0.0380	−1.1590
SE($\hat{\beta}_0$)	2.9551	2.3125	20.5781
SE($\hat{\beta}_1$)	0.8189	0.1207	3.0440
SE($\hat{\beta}_2$)	1.4470	0.2278	4.5978
SE($\hat{\beta}_3$)	0.0175	0.0157	0.6423

TABLE 7. Iterative principal component and Stein estimates and associated asymptotic standard errors for lake trout data ($N=32$).

	pc(-1)	pc(-2)	Stein($c=0.77$)
$\hat{\beta}_0$	3.8311	−6.3505	−179.6000
$\hat{\beta}_1$	1.5139	0.2607	24.9259
$\hat{\beta}_2$	−2.9043	0.6635	35.1343
$\hat{\beta}_3$	0.0629	0.0355	−4.8637
SE($\hat{\beta}_0$)	8.2051	3.9684	98.1859
SE($\hat{\beta}_1$)	0.9215	0.1637	12.8122
SE($\hat{\beta}_2$)	2.5853	0.4164	20.0781
SE($\hat{\beta}_3$)	0.0302	0.0223	2.6320

natural units, i.e. Tables 4–5 and Fig. 1–2 provide $\hat{\beta}_R$ and SE($\hat{\beta}_R$).

The results of the iterative principal components analysis and Stein method are given in Tables 6 and 7. Tables 6–8 use the uncentered and unscaled $\hat{\beta}_s^{pc}$ and SE($\hat{\beta}_s^{pc}$). The estimates obtained by deleting one component are not satisfactory in terms of the signs of the coefficients. The second set of estimates, obtained by deleting two components, is more satisfactory. Note that the estimates obtained by Reckhow et al. (1987, their table 4) are quite close to these except for the coefficient for alkalinity. In contrast to Reckhow et al., we did not find close similarity between principal component estimates and the ridge estimates. In particular, the signs of the coefficients were not consistent between the two methods. Further, the ridge standard errors for the estimates still seem inflated. The addition of the interaction term is not supported. Thus this analysis would suggest that the model is not appropriate.

TABLE 8. Estimated rate of a change in probability of occurrence (at $\hat{\pi}_i = .5$) per unit change in explanatory variable ($\delta X_i = 1$) for various estimation techniques using trout data.

Brook trout $\frac{\delta\hat{\pi}_i}{\partial X_i}$	ML	Ridge (0.001)	Ridge (0.002)	pc (−1)	pc (−2)	Stein
pH	2.298	0.883	0.685	0.558	0.073	1.789
ln (Ca)	2.002	−0.270	−0.245	−0.745	0.138	1.559
pH * ln (Ca)	−0.372	−0.073	−0.036	0.013	0.009	−0.289
Lake trout $\frac{\delta\hat{\pi}_i}{\partial X_i}$						
pH	8.096	0.623	0.469	0.378	0.065	6.231
ln (Alkalinity)	11.411	−0.203	−0.389	−0.726	0.166	8.784
pH * ln (Alkalinity)	−1.579	−0.045	−0.016	0.016	0.009	−1.216

The Stein method does not seem to produce good results as the coefficients are too large relative to single variable models. We therefore cannot recommend it as a technique. Table 8 provides estimates of the rate of change in the probability of occurence of trout species for a given unit change in a specific explanatory variable (i.e. if x_i is changed by one unit, what is the resulting change in the probability of occurrence). It is interesting to note that except for maximum likelihood and Stein methods, the estimated changes are similar across species for a given method. However, there are considerable differences between the methods. Although we do not think that any of the methods is appropriate for this data, the method that incorporates all the variables and seems best is the principal components method, deleting two components associated with the smallest eigenvalues.

From a modelling point of view, notice that none of the approaches are totally satisfactory. Most of the coefficients are only marginally significant and some of the coefficients are not significant (at 0.05 significance level). Furthermore for the models given in Table 1, the maximum likelihood estimated standard errors are all deemed infinite (Hauck and Donner 1977) except for both intercepts and the pH coefficient in the brook trout model. Harrell (1986) deems the standard errors infinite, in the SAS software Proc Logist procedure, if the absolute value of a parameter estimate is greater than or equal to five divided by the range of the corresponding explanatory variable, and its standard error is greater than or equal to 15 divided by the range. The trout examples coefficients have such a limited range. In addition, the WVIFs and the condition numbers in Tables 2–3 indicate severe ill-conditioning of the information matrix, and hence an unstable inverse. The approaches given suggest that either variables need to be deleted or the model needs to be changed.

Discussion

The methods described above are useful for diagnosing possible problems in logistic regression. The diagnostics are quite effective at detecting multicollinearity and indicating which variables are involved. The logistic ridge, principal component, and Stein methods provide reasonable alternative parameter estimates to that of maximum likelihood, especially when weighted multicollinearities are detected and explanatory variables are not deleted. Evidence from applications to acid rain and other data sets indicates that the principal components and ridge methods generally give reasonable results. The Stein approach does not shrink the estimates enough and we cannot recommend it as a useful procedure. Choice of which method is better in any application depends on the purpose of the model. Models in which good parameter estimates are required should be assessed differently from models in which good prediction is required. With complex data, one need not expect a single model to be the best for all purposes.

The ridge, principal component, and Stein methods described here for logistic regression can substantially reduce the variance of the estimated coefficients and prediction variance for future observations outside the mainstream of weighted multicollinearity (i.e. future observations having levels of the explanatory variables not among the patterns of data when the columns of $\hat{S}$ are plotted). The principal components method presented here uses an approach for deletion of components based on the magnitude of the eigenvalues and assessment of fit of the reduced model. Other approaches are possible, i.e. prediction. The principal component approach given here differs from that of Reckhow et al. (1987). In their approach, a principal component analysis of $X_p' X_p$ was done, then logistic regression on leading components. This approach, while successful in many cases, ignores the underlying model and may lead to difficulties. In fitting models, assessment and adjustment for multicollinearity is only one aspect of the modelling process. Other concerns, such as influential values (Pregibon 1981) and assessment of prediction are also impotant and need to be addressed. Lastly, the biased estimation techniques described in this paper can be used in the broader framework of the generalized linear model as described in Marx and Smith (1990) and Marx (1988).

Acknowledgments

The authors would like to thank Dr. Geoffrey T. Evans and the referees for their careful and thought provoking review of this paper.

References

BAKER, J. P. 1984. Fish, p. 5–74 and 5–133. *In* A. Altshuller, and R. Linthurst, [ed.] The acidic deposition phenomenon and its effects. Critical assessment review papers, Vol. II. EPA-600/8-83-016BF. U.S. Environmental Protection Agency, Washington, DC.

BELSLEY, D. A., E. KUH, AND R. E. WELSCH. 1980. Regression diagnostics: influential data and sources of collinearity. John Wiley and Sons, New York, NY. 438 p.

CHRISTENSEN, S. W., J. E. BRECK, AND W. VAN WINKLE. 1988. Predicting acidification effects on fish populations, using laboratory data and field information. Environ. Toxicol. Chem. 7: 735–747.

HAINES, T. A. 1981. Acidic precipitation and its consequences for aquatic ecosystems: a review. Trans. Am. Fish. Soc. 110: 669–707.

HARRELL, F. E. 1986. The logist procedure. Sugi Supplemental Library User's Guide, Version 5 ed., Chap. 23. SAS Institute Inc. Cary, NC. 662 p.

HARTREE, D. R. 1952. Numerical analysis, p. 152–154. Oxford University Press, London.

HAUCK, W. W., AND A. DONNER. 1977. Wald's test as applied to hypothesis in logit analysis. J. Am. Stat. Assoc. Vol. 72 (360), 851–853.

HOSMER, D. W., AND S. LEMESHOW. 1989. Applied logistic regression. John Wiley and Sons Inc., New York, NY. 307 p.

HOWELLS, G. 1983. Acid waters – the effects of low pH and acid associated factors on fisheries. Adv. Appl. Biol. 9: 143–255.

MAGNUSON, J. J., J. P. BAKER, AND E. J. RAHEL. 1984. A critical assessment of effects of acidification on fisheries in North America. Philos. Trans. R. Soc. Lond. Ser. B 305: 501–516.

MARX, B. D. 1988. Ill-conditioned information matrices and the generalized linear model: an asymptotically biased estimation approach. Ph.D. dissertation. Virginia Polytechnic Institute and State University, USA.

MARX, B. D., AND E. P. SMITH. 1989. Weighted multicollinearity diagnostics for logistic regression. Proceedings of the 1989 SAS User's Group International.

1990. Principal component estimation for generalized linear regression. Biometrika. 77(1): 23–31.

MYERS, R. H. 1990. Classical and modern regression with applications, 2nd ed. Duxbury Press, Boston. 359 p.

PREGIBON, D. 1981. Logistic regression diagnostics. The Annals of Statistics 9(4): 705–724.

RECKHOW, K. H., R. W. BLACK, T. B. STOCKTON, JR., J. D. VOGT, AND J. G. WOOD. 1987. Empirical models of fish response to lake acidification. Can. J. Fish. Aquat. Sci. 44: 1432–1442.

SCHAEFER, R. L. 1979. Multicollinearity and logistic regression. Ph.D. dissertation, University of Michigan, USA.

1986. Alternative estimators in logistic regression when the data are collinear. J. Stat. Comp. Simul. 25: 79–91.

STEIN, C. M. 1960. Multiple regression. Contributions to probability and statistics. Stanford University Press, CA. p. 424–443.