COMMUN. STATIST.—THEORY METH., 22(7), 1933-1952 (1993)

Collinearity in Generalized Linear Regression

Emmanuel Lesaffre * Brian D. Marx †

KEYWORDS: Biased estimation; Condition index; Existence of ML estimates; Ill-conditioning; Maximum likelihood estimator; Collinearity.

Abstract

The problem of ill-conditioning in generalized linear regression is investigated. Besides collinearity among the explanatory variables, we define another type of ill-conditioning, namely *ML-collinearity*, which has similar detrimental effects on the covariance matrix, e.g. inflation of some of the estimated standard errors of the regression coefficients. For either situation there is collinearity among the columns of the matrix of the weighted variables. We present both methods to detect, as well as practical examples to illustrate, the difference between these two types of ill-conditioning. Also the applicability of alternative regression methods will be reviewed.

*Biostatistical Centre, Department of Epidemiology, Catholic University Leuven, 3000 Leuven, Belgium

†Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803-5606, U.S.A.

1 Introduction

In linear regression computational difficulties arise when the explanatory variables are collinear. A set of variables are said to be collinear if one or more variables in the set can be expressed exactly or nearly as a linear combination of the others in the set. If the problem is aimed at parameter estimation, then collinearity will also cause statistical difficulties. We define a fitted regression model to be ill-conditioned if the (estimated) information matrix is ill-conditioned. The detrimental effects of ill-conditioned information in generalized linear regression models (GLRs) [16] include inflated variances of the estimated coefficients and poor prediction in certain regions of the regressor space. Another effect can be a nonsignificant Wald statistic even when the regressors are highly predictive. This was first pointed out by Hauck and Donner [9] for the logistic model. More specifically, they showed that if $|\hat{\beta}_j| \to \infty$, then its Wald statistic converges to zero. Væth[21] explored conditions for this to happen for the other members of the family of generalized linear models. In GLRs, collinearity among the regressors can also, but may not be the only source of ill-conditioning problems. We demonstrate that ill-conditioning problems exist in GLR when the explanatory variables are not severely collinear.

In the logistic regression framework, a number of papers appeared on ill-conditioning problems, see e.g. [19,20,13]. Others discussed the problem in the framework of a GLR, see [12,14]. Belsley [4] treated the general nonlinear model. In all of the above referenced papers the authors point out that it is not the collinear relations among the explanatory variables matter, but rather of weighted explanatory variables. Yet, in all these mentioned GLR examples, the explanatory variables are highly correlated. Hence the previous research did not investigate the effects of other sources of ill-conditioning when collinearity is not severe. Mackinnon and Puterman [12] did notice that : ... *it is possible that the GLM might not be collinear at the maximum likelihood estimate of* β *while it is collinear at* β . Thus, they recognized the dependence of the ill-conditioning problems on the particular value of β. However, this was not further investigated in their paper. Related to the observation made by Mackinnon and Puterman, we illustrate that, in addition to collinearity, there is another situation where the (estimated) asymptotic variances of the regression coefficients are inflated.

As the vehicle to illustrate our viewpoint, the logistic regression model with a binary response will be taken in Section 2. In the presence of ill-conditioned information, the practitioner can and will resort to an alternative model, either by simply deleting certain regressors, which is often done

in medical applications, or by using alternative estimators, e.g. ridge regression, principal component regression, Stein estimators, etc.. We will indicate that the first method can lead to much less efficient estimators if the cause for ill-conditioning of the information matrix is not collinearity among the explanatory variables. In the third section our arguments will be extended to GLRs. It will be highlighted that there is a close connection between ill-conditioning and that of existence and uniqueness of maximum likelihood estimators. Methods to detect the two types of ill-conditioning problems will be treated in Section 4. The use of alternative regression models in the presence of an ill-conditioned information matrix will be investigated in the fifth section. Our proposals for a conditioning analysis will be applied to an example in Section 6.

2 Examples of collinearity in logistic regression

We first look at the analysis of two constructed data sets and then look at the pattern of the data. For the analysis we used the SAS PROC LOGISTIC [18].

Each of the two data sets (N=25) contain one 0-1 binary response variable and two explanatory variables, x and y. In Table 1, observe that models based on a single explanatory variable have a significant impact on the response. However, jointly their significance is lost when evaluated by the Wald test. This reminds us of problems associated with collinearity in standard multiple regression. Besides some changes in signs, we basically find the same result in Table 2. Thus when both explanatory variables are included, in either example, we find inflated estimated variances and apparently inflated regression coefficients. A possible remedy, often employed in medical applications is to drop one of the regressors. Notice, however, that this will reduce the correct classification rate from 84% to either 72% or 68% in the first data set, and from 92% to either 92% or 84% in the second data set (see Tables 1 and 2). Thus, there is some loss of efficiency observed in the first data set by deleting one of the regressors, however this loss could be ascribed to the size of the data set and the roughness of the classification rate as a performance measure. A more dramatic difference between the two data sets is observed when looking at the scatterplots of the two regressors together with the group information (Figures 1 and 2). In the first data set the regressors are uncorrelated (r=0.006), but more importantly the two groups are almost separated by a bisecting line. In the second data set, a

Table 1: Artificial data set 1: regression coefficients of regressors x and y in logistic regression model predicting the binary response variable with each explanatory variable separately (UNIV) and then predicting with both variables jointly (MULT).

data set 1					
type	model	coeff	SE(coeff)	P(Wald)	%corr
UNIV	x	-1.02	0.41	0.01	72%
UNIV	y	0.99	0.40	0.01	68%
MULT	x	-5.02	4.17	0.23	84%
	y	5.12	4.30	0.23	

Table 2: Artificial data set 2: regression coefficients of regressors x and y in logistic regression model predicting the binary response variable with each explanatory variable separately (UNIV) and then predicting with both variables jointly (MULT).

data set 2					
type	model	coeff	SE(coeff)	P(Wald)	%corr
UNIV	x	0.46	0.20	0.02	92%
UNIV	y	0.45	0.19	0.02	84%
MULT	x	2.75	2.90	0.34	92%
	y	-2.27	2.84	0.42	

high correlation exists among the regressors (r=0.999). For either of the two examples, SAS does not warn the user of the possible problems associated with ill-conditioned information.

In the first example, the data configuration approaches (quasi)-complete separation as described by Albert and Anderson [2], see also [17]. It is known

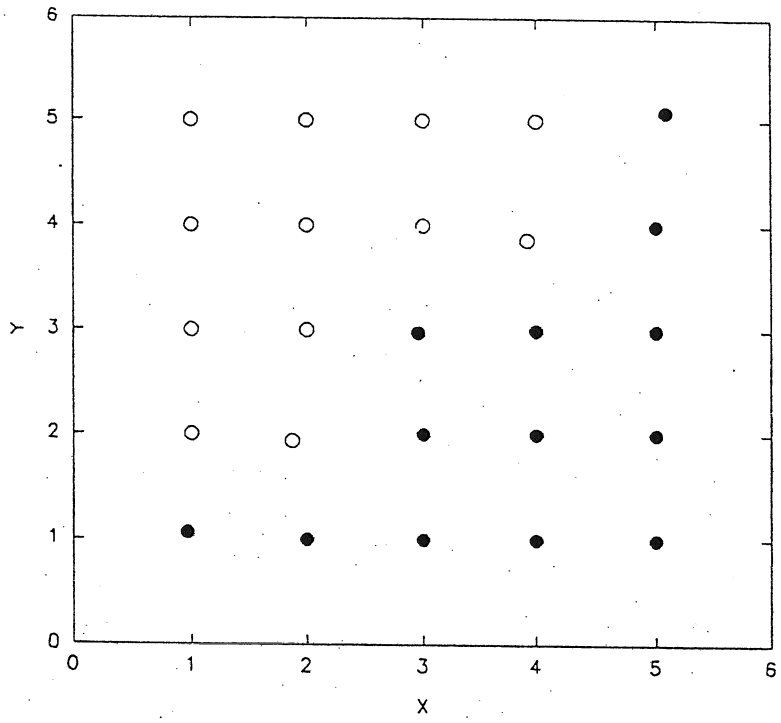

Figure 1: Artificial data set 1: scatterplot of regressors together with group information. Example of ML-collinearity.

that for exact (quasi-)complete separation the ML estimate of at least one of the regression coefficients is infinite. A completely different picture is seen in the second data set where collinearity among the regressors is the cause of concern. We present two completely different natures of ill-conditioned information.

We define the asymptotic covariance matrix of the estimators, $(X'VX)^{-1}$, where X is the design matrix, $V = \text{diag}(v_i)$ with $v_i = p(x_i)(1 - p(x_i))$ and $p(x_i)$ = probability of success in the logistic model for the ith observation with covariate vector x_i. The Fisher information matrix $W = X'VX$ is singular if $S = V^{\frac{1}{2}}X$ is less than full column rank, with $V^{\frac{1}{2}}$ the square root matrix of the positive semi-definite $n \times n$ matrix V. But for every set of finite true regression coefficients the matrix V is positive definite so that

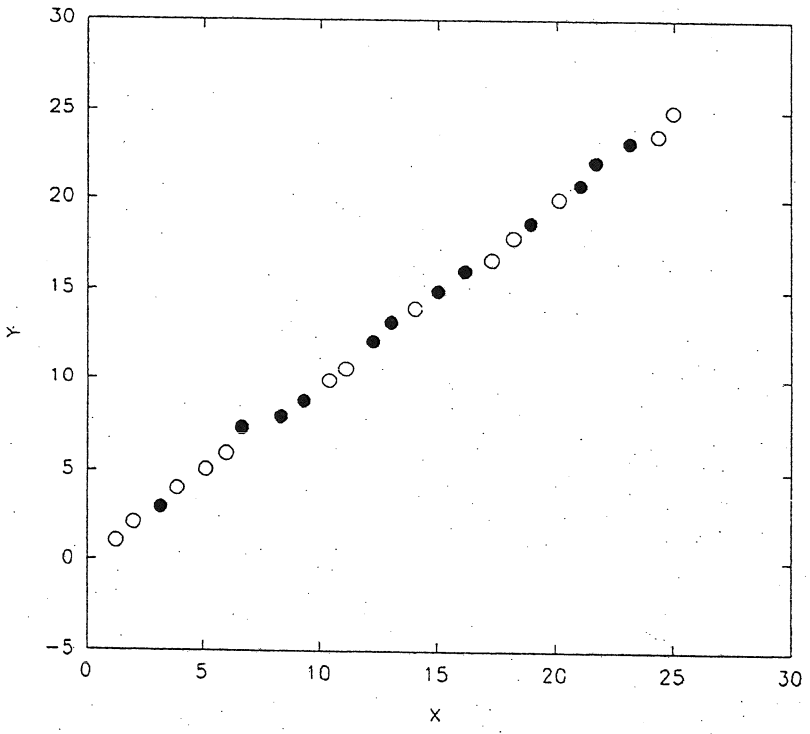

Figure 2: Artificial data set 2: scatterplot of regressors together with group information. Example of multicollinearity.

W can only be singular if X is not of full column rank. However, it is the estimated, and not the exact covariance matrix, which is reported as a result of a maximization program, given by $(X'\hat{V}X)^{-1}$, with $p(x_i)$ replaced by its MLE. It is known that if the data has a (quasi)-completely separated configuration, then at least one of the MLEs of the regression coefficient vector is infinite. Consequently, some of the diagonal elements of $\hat{V}$ become exactly equal to zero resulting in a singular $\hat{W} = X'\hat{V}X$ matrix.

It will next be shown for the logistic regression framework that exact collinearity among the regressors and nonexistence of the MLEs are the only causes for a singular estimated information matrix, $\hat{W}$. However, ill-conditioning of a nonsingular estimated information matrix occurs either when the regressors are strongly collinear or when the diagnostic groups ap-

proach (quasi)-complete separation. Both ill-conditioning situations result in inflated variances of the estimated regression coefficients and reflect strong collinearity in $\hat{S} = \hat{V}^{\frac{1}{2}}X$, yet are fundamentally different.

As in any regression model, logistic regression argues conditionally on the regressors. In experimental settings where the regressors are collinear, the true unknown variances will be inflated. On the other hand, while in experimental settings with near separation of the groups, the true unknown variance may or may not be inflated, it is the estimate of the true variance which suffers from inflation. This important and subtle distinction will be extended to the class of generalized linear regressions in the next section.

3 Collinearity in generalized linear regression.

3.1 Definitions and notation

Let $X = (x_0, x_1, x_2, \ldots, x_p)$ be a $N \times (p+1)$ matrix of regressors with the first column a vector of ones. Let Y be a $N \times 1$ random response vector for which each entry in Y follows the same distribution in the exponential family. It is assumed that each entry of the response vector, Y_i corresponds to a $1 \times (p+1)$ vector $x' = (x_0, \ldots, x_p)$. The probability density for Y_i is assumed to be

$$f(y; \theta, \phi) = exp[\{y\theta + c(\theta)\}/q(\phi) + d(y, \phi)], \tag{1}$$

where c, d, q are known functions. Let the nuisance parameter ϕ be constant for all Y_i.

Given a constant and a set of p explanatory variables, a generalized linear regression model utilizes the relationship,

$$g(\mu_i) = x_i'\beta = \eta_i, \tag{2}$$

satisfying (i) $\mu_i = E(Y_i)$; (ii) g is a link function with $g^{-1} = h$ and is twice differentiable in the interior of $[\mu_{min}, \mu_{max}] = I_\mu \subset \Re$; (iii) β is an unknown parameter vector and is an element of the parameter space, P; (iv) the estimation of β does not depend on having an estimate of ϕ.

The parameter space can be denoted as $P = \bigcap_i P_i$, where

$$P_i = \{\beta \in \Re^{p+1} \mid g^{-1}(x_i'\beta) \in I_\mu\}.$$

Define $int(P_i)$ as the interior of P_i and $cl(P_i)$ as the closure of P_i in $\bar{\Re}^{p+1}$. Since P_i is convex, we have $int(P) = \bigcap_i int(P_i)$ and $cl(P) = \bigcap_i cl(P_i)$. Thus β belongs to the boundary of P, i.e. $\beta \in \partial(cl(P))$, iff $\beta \in cl(P_i)$ for all i and there exists at least one i for which $\beta \notin int(P_i)$. In other words, β belongs to the boundary of P iff $\exists i$ such that $g^{-1}(x_i'\beta) = \mu_{min}$ or μ_{max}.

For Poisson and logistic regression models with the natural link function P_i equals the (p+1)-dimensional Euclidean space. Then, β belongs to the boundary of P if it is a vector "at infinity" corresponding with $\mu_{min} = 0$ or $\mu_{max} = \infty$ for Poisson models and to $\mu_{min} = 0$ or $\mu_{max} = 1$ for logistic regression models. If the identity link is used in Poisson models, then the $P_i = \{\beta \in \Re^{p+1} \mid 0 < x_i'\beta < \infty\}$, each are cones (without their origins). Therefore P is the intersection of n cones. The origin corresponds to $\{\beta \in \Re^{p+1} \mid x_i'\beta = 0\}$. This follows since if $\beta \in P_i$, then $\delta\beta \in P_i, \forall \delta > 0$.

Standard theory of generalized linear models can be found in McCullagh and Nelder [15] and Aitkin, Anderson, Francis and Hinde [1]. The scoring algorithm to estimate the unknown parameters is given by

$$\hat{\beta}_t = (X'\hat{V}_{t-1}X)^{-1}X'\hat{V}_{t-1}\hat{z}^*_{t-1}, \tag{3}$$

where $\hat{z}^*_{t-1}$ has elements $\hat{z}^*_i = \hat{\eta}_i + \hat{e}_i(\partial\eta_i/\partial\mu_i)$ evaluated in $\hat{\beta}_{t-1}$, as is also the residual $\hat{e}_i = y_i - \hat{\mu}_i$ and $\hat{V}_{t-1} = diag[\{h'(\hat{\eta}_i\}^2/var(Y_i)]_{t-1}$. The diagonal elements of $\hat{V}$ are equal to

$$g_n'(\hat{\mu}_i)/\{g'(\hat{\mu}_i)^2 q(\tilde{\phi})\},$$

where g is the link function used, g_n the natural link function, $q(\tilde{\phi})$ the estimate of $q(\phi)$. For $g \equiv g_n$, the $\hat{v}_i$ are inversely proportional to $g_n'(\hat{\mu}_i)$. The information matrix can be estimated by the actual or the expected Hessian matrix evaluated at the MLE. Here we take as estimate the expected Hessian matrix given by $\hat{W} = X'\hat{V}X$. The asymptotic covariance matrix of $\hat{\beta}$ is then estimated by $\hat{W}^{-1}$, if it exists.

3.2 Causes of ill-conditioned information

In the lemma below we investigate what can be concluded from the data if the information matrix $\hat{W}$ is singular.

Lemma 1 When $X'\hat{V}X$ is singular, then either X is not of full rank or $\hat{\beta} \in \partial(cl(P))$, or both.

Proof

If $X'\hat{V}X$ is singular and rank$(\hat{V}) = n$, then trivially X cannot be of full column rank. Thus when X is full column rank, the only other situation to

have a singular $X'\hat{V}X$ is when $\text{rank}(\hat{V}) < n$. Since $\hat{V}$ is diagonal, $\text{rank}(\hat{V}) < n$ implies that least one $\hat{v}_i = 0 = g_n'(\hat{\mu}_i)/\{g'(\hat{\mu}_i)^2 q(\tilde{\phi})\}$. However g is a monotone and twice differentiable function in the interior of I_μ. Thus $\hat{\mu}_i = \mu_{min}$ or μ_{max} for at least one i and $\hat{\beta} \in \partial(cl(P))$. □

Thus, whenever the estimated covariance matrix is singular, it is either because of exact linear dependence among the explanatory variables or because the MLE does not exist. In either of the two situations, there is an exact linear dependence in the constructed variables defined by the columns of $\hat{S} = \hat{V}^{\frac{1}{2}}X$. Further in the case when the MLE does not exist, $X'\hat{V}X = X^{*\prime}\hat{V}^*X^*$, where X^* and $\hat{V}^*$ are the submatrices corresponding to those elements in X and $\hat{V}$, respectively, for which $\hat{v}_i \neq 0$. Then, since $\hat{W}$ is singular, necessarily X^* cannot be of full rank, so that there are exact linear dependencies among the columns of X^*, i.e. restricting the sample space to those cases which do not have an extreme $\hat{\mu}_i$. The conditions for the converse to hold, i.e. if $\hat{\beta}$ belongs to the boundary of the parameter space then $\hat{W}$ is singular, depend on the chosen link function and the boundary $\partial(cl(P))$. We now illustrate that for the logistic regression model there is equivalence, but not for all GLRs.

Given full column rank X, then the MLE of β does not exist for the logistic regression model iff $\exists\beta_0$ such that $x_i'\beta_0 \geq 0$ for all observations of the first diagnostic class and $x_i'\beta_0 \leq 0$ for the remaining observations, i.e. if the groups lie opposite a particular hyperplane. In such a setting, $\hat{\beta}$ lies on the boundary of the parameter space implying that $\hat{\beta}$ has an infinite component. Thus if $\hat{\beta} \in \partial(cl(P))$, then $\hat{v}_i \neq 0$ if and only if $x_i'\hat{\beta} = 0$, since only the elements with $\hat{v}_i \neq 0$ contribute to $\hat{W}$. However, the regressors of these observations must satisfy a collinear relationship specified by the coefficients of the MLE. The converse is proved.

In the binomial model with a non-natural link function, e.g. probit and complementary log-log link, the converse can only hold true if $v_i = 0$ when $\beta \in \partial(cl(P)))$. However $\hat{\beta}$ can belong to the boundary of the parameter space, without singular $\hat{W}$, in the case of the power link function $\eta = (\mu^\lambda - 1)/\lambda$ when $\lambda = .5$. In such a case, $v_i = 1$ (up to the factor in ϕ) for $\mu = 0$ or 1. Of course this discussion relates to the expected Hessian matrix evaluated at the MLE and not to the actual Hessian matrix.

The case of a singular information matrix is treated above. However, the literature focuses on ill-conditioned nonsingular information. As presented in Section 2, the two data patterns presented in Figures 1 and 2 posed two different natures of ill-conditioning a nonsingular information matrix. In

GLRs, the structure of the data patterns to force $\hat{\beta}$ to the boundary of the parameter space may not have the clear visualization that is provided by separation in logistic regression. Nonetheless the dire consequences of ill-conditioned information are present. We can moreover conclude from above that there also exists two sources of ill-conditioning of information for all GLRs. The first arises when the explanatory variables are collinear. If the explanatory variables are not collinear, but at the MLE there is collinearity among the constructed variables $\hat{S}$, then we define this as *ML-collinearity*.

4 Detection of collinearity problems

The first priority is to detect whether or not the information matrix is ill-conditioned. If this is the case, the type of collinearity and the actual collinear relationship(s) need to be known. These steps will be treated below.

4.1 Detection of ill-conditioned information

Ill-conditioning of a nonsingular $\hat{W}$ can be flagged using its condition number. However, there is no agreement in the statistical literature of how to measure the severity of the ill-conditioning. Suppose the columns of X are standardized to unit length.

Let $\hat{\lambda}_0, \ldots, \hat{\lambda}_p$ be the eigenvalues of $\hat{W}$ in decreasing order. Define the information condition numbers $\kappa_{W_j} = (\hat{\lambda}_0/\hat{\lambda}_j)^{\frac{1}{2}}$, and the largest as $\kappa_W (\equiv \kappa_{W_p})$. Mackinnon and Puterman [12] suggested to use κ_W as a criterion to detect an ill-conditioned information matrix. Several geometrical and statistical properties can be attributed to κ_W. For instance, it can be shown that κ_W describes the worst relative precision with which linear combinations of $\hat{\beta}$ can be estimated. Belsley and Oldford [6] showed that this diagnostic can be used for loglikelihood conditioning. Belsley [4], Weissfeld and Sereika [22] suggested the condition number of a standardized W, i.e. of $\hat{W}_s = D\hat{W}D$, with $D^2 = diag(\hat{W})^{-1}$. Marx and Smith [13] proposed yet another condition number based on $\hat{\Phi}^* = \hat{S}^{*'}\hat{S}^*$, with $\hat{S}^*$ the centered and standardized (to unit length columns) version of the matrix $\hat{S} = \hat{V}^{\frac{1}{2}}X$. Yet only κ_{W_j} is structurally interpretable. Therefore, and in view of the distinction between the two types of collinearity problems, we propose to take κ_W as a diagnostic for detecting ill-conditioned information with the columns of X standardized to unit length.

4.2 Establishing the nature of the ill-conditioning problem

Whenever the information matrix is deemed ill-conditioned by κ_W, we recommend distinguishing between collinearity among the explanatory variables and ML-collinearity. It is useful to calculate the ratio $r_{WX} = \kappa_W/\kappa_X$, where κ_X is the condition number of X. The design matrix is not square so κ_X is defined as the condition number of $X'X$. If the ratio r_{WX} is high, e.g. more than 5, and $\kappa_W > 30$ there is ML-collinearity. If $\kappa_X > 30$ there is collinearity among the explanatory variables. Mackinnon and Puterman provided easy to determine bounds for κ_W [12]. They showed that the square of r_{WX} is bounded below by $\hat{v}_{min}/\hat{v}_{max}$ and above by $\hat{v}_{max}/\hat{v}_{min}$. However, these bounds are often too crude to help distinguish between the two types of collinearity.

A variance decomposition table, originally suggested by Belsley, Kuh and Welsch [5] for linear regression models, helps in identifying the collinear relationships among the regressors. Marx and Smith [13], Weissfeld and Sereika [22] proposed such a table for the estimated information matrix $\hat{W}$. We also suggest such a diagnostic table. However, it must be recognized that its value as a diagnostic is limited. Indeed, the table cannot always point to the correct collinear relationship among the explanatory variables, if one exists. Further, if there is ML-collinearity, then the table cannot provide the user with the collinear relationship among the columns of the submatrix X^*. Therefore we suggest the following strategy:

(1) standardize the columns of X, including the constant vector, to unit length;

(2) calculate the condition numbers κ_W, κ_X and their ratio;

(3) determine whether there is evidence of ill-conditioning. If $\kappa_X > 30$ there is collinearity in X, if $\kappa_W > 30$ and κ_X is not high, there is ML-collinearity. If both are large and the ratio is much more than 1 then both types of collinearity are present;

(4) calculate the variance decomposition table of $\hat{W}$, this shows the involved weighted regressors in the collinear relationships;

(5) if ML-collinearity exists, determine the cases with $\hat{v}_i > \varepsilon$, with ε an a priori determined small number. These data constitute the submatrix of X^* of X. The columns of this restricted data matrix provide information regarding the linear relationship leading to ML-collinearity, such as in logistic regression where such a relationship approximately defines the separating

hyperplane. To see this, in logistic regression we note that for quasi-complete separation the observations on the separating hyperplane have, at the MLE, a predicted probability away from 0 and 1 (see Albert and Anderson, [2]).

5 Alternative regression methods

The most simple "biased" regression procedure is variable selection. This is often done in medical studies where there are many regressors and the ill-conditioned information matrix is often interpreted as a redundancy of regressors. But, we have shown in Section 2 that variable deletion is not always the best method to alleviate ill-conditioning of the information matrix, especially with ML-collinearity, since it can lead to a dramatic loss in the model's discriminative ability. We further note that in the first data set the Wald statistic for each of the two regressors was not significant in the joint model, but the loglikelihood ratio statistic for y was highly significant when x was already in the model ($-2logL = 18.195$).

Several alternative regression procedures have been proposed for GLRs in the statistical literature. Basically they are straightforward extensions of methods developed in linear regression. Schaefer, Roi and Wolfe [19] first suggested the ridge logistic estimator. Other proposals are found in [20,12,8]. Marx and Smith [14] suggested an iterative principal component procedure. Consider further generalizing by utilizing a continuum of principal components for each observation, $Z_\delta = XM_\delta$, where the (i,j)th element of Z_δ is the score of the jth principal component for the ith observation. Define M_δ as the $(p+1)\times(p+1)$ matrix whose jth column is the jth eigenvector of the information matrix, $W_\delta = X'V^\delta X$. The parameter $0 \leq \delta \leq 1$ connects the information matrix ($\delta = 1$), on a continuum, to the correlation/covariance matrix ($\delta = 0$). Hence M_δ is an orthogonal matrix and $M_\delta' W_\delta M_\delta = diag(\lambda_{\delta i}) = \Lambda_\delta$, where $\lambda_{\delta i}$ are the corresponding eigenvalues of W_δ. We rewrite equation (2) as

$$\eta_{\delta i} = z_{\delta i}'\alpha_\delta, \tag{4}$$

where $z_{\delta i}'$ is the ith row of Z_δ and $\alpha_\delta = M_\delta'\beta$. Equation (4) provides the orthogonally transformed full principal component model. In constructing a principal component model to alleviate ill-conditioning, a choice of $\delta = 0$ can be fruitful in the presence of explanatory variable collinearity when ML-collinearity is not present. Whereas with ML-collinearity (regardless of collinearity in X), $\delta = 1$ is a better choice. As pointed out in [14], the full or reduced principal component estimators can be estimated by either iteratively or via a one-step adjustment to the ML estimator, if it exists.

Conditional on collinearity in X, the variances of the estimated parameters are inflated for all sampled situations. Instead, ML-collinearity is a sign that only for this sample, due to possibly the sparseness of the data set, the variance of the estimated parameters is biasedly estimated. Thus, since the biased regression methods rely heavily on the variance-covariance matrix of the estimated parameters, it still has to be investigated whether they are useful in the case of ML-collinearity. But, for the one-step methods a cautionary note can already be made if ML-collinearity is present. Indeed, due to the fact with ML-collinearity that some of the estimated $\hat{v}_i \approx 0$, but the corresponding v_i are nonzero, the ML one-step adjustment to the ML estimate can lead to inconsistent biased estimation. However, it is clear that both for collinearity in X, as well as for ML-collinearity, the alternative regression methods will shrink the length of the estimated parameter vector.

6 The analysis of a practical example

For the analysis of the example we used SAS PROC LOGISTIC together with SAS-IML programs. These programs are available from the authors upon request.

6.1 Illustration of collinearity

The cancer data from Lee [10] are taken to illustrate collinearity in logistic regression. A table with the original data is included in the Appendix. The continuous characteristics associated with cancer remission are: *Cell index* (CELL), *Smear index* (SMEAR), *Infil index*, *Li index* (LI), *Blast index* (BLAST) and *Temperature* (TEMP). The binary response is 1 if the patient experiences a complete cancer remission and 0 otherwise. There are 27 patients, 9 of which experienced a complete cancer remission. Forward selection and backward elimination procedures were performed for these data. Forward selection chooses regressors LI, TEMP, and CELL. This will be our first model of interest.

The regressors LI, TEMP, CELL are standardized to unit length. The parameter estimates (standard error, P-value) are of the Intercept, LI, TEMP, CELL : 3.99 (2.25, 0.08), -1.81 (0.83, 0.03), 1.22 (0.92, 0.18), -1.80 (1.45, 0.21), respectively. The condition number of X equals 190.78 indicating collinearity problems in X. The fitted model is almost double the condition number of X, with $\kappa_W = 329.95$. We note that this is close to the condition number of the standardized matrix $\hat{W}_s$ which is equal to 315.14. Mackinnon

Table 3: Variance decomposition table of Lee cancer remission data set applied to the original data.

Variance Decomposition Table Original Data					
Lee data with regressors LI, TEMP and CELL					
Eigen value	Cond Number	Variance Proportion			
		Interc	LI	TEMP	CELL
3.843	1.00	0.00	0.01	0.00	0.00
0.129	5.45	0.00	0.98	0.00	0.02
0.028	11.78	0.00	0.01	0.00	0.97
0.00011	190.78	1.00	0.01	1.00	0.01

Table 4: Variance decomposition table of Lee cancer remission data set applied to the information matrix obtained from logistic regression predicting cancer remission from LI, TEMP and CELL.

Variance Decomposition Table Information Matrix					
Lee data with regressors LI, TEMP and CELL					
Eigen value	Cond Number	Variance Proportion			
		Interc	LI	TEMP	CELL
0.576	1.00	0.00	0.00	0.00	0.00
0.015	6.17	0.00	0.45	0.00	0.01
0.00055	32.29	0.01	0.10	0.00	0.82
5.29E-6	329.95	1.00	0.44	1.00	0.18

and Puterman's bounds for κ_W are 5.21 and 6992, which are uninformative for this data. According to [12] the linear combination with the highest variance is $-0.68 - 0.02 * LI + 0.73 * TEMP - 0.03 * CELL$, while that with the lowest variance is $0.48 + 0.53 * LI + 0.48 * TEMP + 0.51 * CELL$. Thus, both the intercept and the coefficient for temperature are badly estimated. The variance decomposition table for the original data and for the information matrix are given in Table 3 and 4, respectively. These tables show a

Table 5: Variance decomposition table of Lee cancer remission data set applied to the original data but with an artificial variable, ARTVAR, added which causes almost quasi-complete separation.

Variance Decomposition Table Original Data					
Lee data with regressors LI, ARTVAR and CELL					
Eigen value	Cond Number	Variance Proportion			
		Interc	LI	ARTVAR	CELL
3.500	1.00	0.00	0.00	0.00	0.00
0.463	2.75	0.01	0.00	0.09	0.01
0.028	11.24	0.11	0.20	0.15	0.71
0.010	18.60	0.87	0.79	0.76	0.28

similar picture for the decomposition of the variance of the regression estimators into its basic components. There seems to be a collinear relationship between TEMP and the intercept. This only happens if the values of TEMP are relatively high with respect to their variability. Indeed, the average temperature (in Fahrenheit) is equal to 99.7 and the standard deviation equal to 1.5, further the minimal value is 98 and the maximal value is 103.8. Thus, we can conclude that one collinear relationship in the original regressors caused the numerical and statistical instability of some of the estimated regression coefficients and perhaps a variable selection or biased estimation technique should be implemented.

6.2 Illustration of ML-collinearity

To illustrate ML-collinearity, we have taken again the cancer remission data set of Lee but we removed TEMP and included an artificial variable ARTVAR. Almost *(quasi)-complete separation* between the two remission groups was created by the line $ARTVAR - 2 * LI + 1 = 0$, i.e. this line almost separates the two classes. For *complete separation*, i.e. if all data points of one class lie with respect to the above line opposite to all points of the other class, PROC LOGISTIC cannot provide unique maximum likelihood estimates. By construction, ARTVAR and LI are highly correlated (r=0.94) but this correlation is not high enough to cause ill-conditioning in the basic data, the condition number of X is equal to 18.60. The variance decomposition

Table 6: Variance decomposition table of Lee cancer remission data set applied to the information matrix obtained from logistic regression predicting cancer remission from LI, ARTVAR and CELL.

Variance Decomposition Table Information Matrix					
Lee data with regressors LI, ARTVAR and CELL					
Eigen value	Cond Number	Variance Proportion			
		Interc	LI	ARTVAR	CELL
0.134	1.00	0.00	0.00	0.00	0.00
0.010	3.59	0.00	0.45	0.00	0.01
0.000025	73.64	0.52	0.39	0.34	0.14
0.000013	100.75	0.48	0.61	0.66	0.86

table of X is shown in Table 5. A logistic model was estimated predicting remission from INTERCEPT, LI, ARTVAR and CELL. The parameter estimates (standard error, P-value) of the standardized estimates are : 13.00 (13.11, 0.32), 24.99 (15.59, 0.11), -32.30 (20.06, 0.11), -4.48 (7.08, 0.53), respectively. Thus none of the regressors are significantly predicting remission, still 85.2% of the cases are correctly classified. The condition number of the information matrix from this logistic model is much higher than for the basic data, $\kappa_W = 100.75$, which is more than 5 times higher than the condition number of X. The variance decomposition table of $\hat{W}$ is shown in Table 6. From this table apparently all regressors are involved in the collinearity corresponding to the condition number of 100.75. A linear regression analysis on the 9 cases with $v_i > 0.01$ shows that these points are all close to the two-dimensional plane $LI = 0.48 + 0.53 * ARTVAR$, this model has an R^2 of 0.99. Observe that this hyperplane almost coincides with the original chosen quasi-separating hyperplane.

In principle also here an alternative regression procedure can be applied to obtain more stable estimates. However, in this case it is not yet clear what the statistical properties are of such a procedure. Anyway, variable selection would be a bad choice.

7 Conclusions

In linear regression collinearity problems only involve the regressors. We have shown that in GLRs also the response and the choice of the model play a role

in the degree of ill-conditioning of the information matrix. This lead us to introduce the concept of ML-collinearity, where there is no collinearity among the original regressors but, due to the combination of response, regressors and the choice of the model, the information matrix becomes ill-conditioned at the MLE. This was not yet recognized in the literature and we have shown that it is of importance to distinguish between multicollinearity and ML-collinearity.

We have concentrated on the logistic regression model to exemplify our ideas. Of course any other member of the class of GLRs (with the natural link function) could have been chosen. Further, our discussion above also applies to other models, not especially GLRs but with an information matrix having the expression $X'\hat{V}X$. For example, for the conditional logistic regression model the estimated information matrix has this expression with $\hat{V}$ a block diagonal matrix [7]. The multiple group logistic regression model is another example [3,11].

When the information is not ill-conditioned, the estimation of parameters in GLRs often works well in practice. Yet, when ill-conditioning occurs, practitioners should be able to take the appropriate corrective action. The current versions of the commercial statistical programs like SAS, BMDP, etc. do not provide the user with sufficient statistical tools to adequately respond to problems associated with ill-conditioning of the information matrix. They do not contain any variance decomposition table of the information matrix nor any alternative GLR regression method. We hope that this paper will add to change this situation.

Appendix

The Lee cancer remission data set was taken from *SAS, SUGI* Supplementary Guide (1986). The study was conducted to associate patient's characteristics with cancer remission. Information was recorded on the following variables: $Y = 1$ if complete cancer remission, $= 0$ if incomplete cancer remission; X1, Cell index; X2, Smear index; X3, Infil index; X4, LI index; X5, Temperature.

OBS	Y	X1	X2	X3	X4	X5
1	1	.800	.830	.660	1.900	.996
2	1	.900	.360	.320	1.400	.992
3	0	.800	.880	.700	.800	.982
4	0	1.000	.870	.870	.700	.986

(continued)

OBS	Y	X1	X2	X3	X4	X5
5	1	.900	.750	.680	1.300	.980
6	0	1.000	.650	.650	.600	.982
7	1	.950	.970	.920	1.000	.992
8	0	.950	.870	.830	1.900	1.020
9	0	1.000	.450	.450	.800	.999
10	0	.950	.360	.340	.500	1.038
11	0	.850	.390	.330	.700	.988
12	0	.700	.760	.530	1.200	.982
13	0	.800	.460	.370	.400	1.006
14	0	.200	.390	.080	.800	.990
15	0	1.000	.900	.900	1.100	.990
16	1	1.000	.840	.840	1.900	1.020
17	0	.650	.420	.270	.500	1.014
18	0	1.000	.750	.750	1.000	1.004
19	0	.500	.440	.220	.600	.990
20	1	1.000	.630	.630	1.100	.986
21	0	1.000	.330	.330	.400	1.010
22	0	.900	.930	.840	.600	1.020
23	1	1.000	.580	.580	1.000	1.002
24	0	.950	.320	.300	1.600	.988
25	1	1.000	.600	.600	1.700	.990
26	1	1.000	.690	.690	.900	.986
27	0	1.000	.730	.730	.700	.986

References

[1] Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989), *"Statistical Modelling in GLIM."*, Oxford: Claredon Press.

[2] Albert, A. and Anderson, J.A. (1984), "On the existence of maximum likelihood estimates in logistic regression models.", *Biometrika*,71, 1, 1-10.

[3] Albert, A. and Lesaffre, E. (1986), "Multiple group logistic discrimination.", *Computers and Mathematics with Applications.*, 12, 209-224.

[4] Belsley, D.A. (1991), *"Conditioning diagnostics: collinearity and weak data in regression."* , New York: John Wiley and Sons.

[5] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980), *"Regression diagnostics: Influential Data and Sources of Collinearity."*, New York: John Wiley and Sons.

[6] Belsley, D.A. and Oldford,, R.W. (1986), "The general problem of ill conditioning and its role in statistical analysis.", *Computational Statistics and Data Analysis*, 4, 104-20.

[7] Davis, C.E., Hyde, J.E., Bangdiwale, S.I. and Nelson, J.J. (1985), "An example of dependencies among variables in a conditional logistic regression." in: *Modern Statistical Methods in Chronic Disease Epidemiology. Proceedings of the SIMS conference*, editors: S.H. Moolgavkar and R.L Prentice, New York: John Wiley and Sons.

[8] Duffy, D.E. and Santner, T.J. (1989), "On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models.", *Communications in Statistics-Theory and Methods*, 18(3), 959-80.

[9] Hauck, W.W., and Donner, A. (1977), "Wald's test as applied to hypotheses in logit analysis.", *Journal of the American Statistical Association*, 72, 360, 851-3.

[10] Lee, E.T. (1974), "A computer program for linear logistic regression analysis.", *Computer Programs in Biomedicine*, 80-92.

[11] Lesaffre, E. and Albert, A. (1989), "Partial separation in logistic discrimination.", *Journal of the Royal Statistical Society, Series B*, 51, 109-16.

[12] Mackinnon, M.J. and Puterman, M.L. (1989), "Collinearity in generalized linear models.", *Communications in Statistics-Theory and Methods*, 18(9), 3463-72.

[13] Marx, B.D. and Smith, E.P. (1990), "Weighted multicollinearity in logistic regression: diagnostics and biased estimation techniques with an example from lake acidification.", *Canadian Journal of Fisheries and Aquatic Sciences*, 47, 6, 1128-35.

[14] Marx, B.D. and Smith, E.P. (1990b), "Principal component estimation for generalized linear regression. ", *Biometrika*, 77, 1, 23-32.

[15] McCullagh, P. and Nelder, J.A. (1989), *"Generalized Linear Models."*, 2nd edition, London: Chapman and Hall.

[16] Nelder, J., Wedderburn, R.W.M. (1972), "Generalized linear models.", *Journal of the Royal Statistical Society*, A 135, 370-84.

[17] Santner, T.J. and Duffy, D.E. (1986), "A note on Albert and Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models.", *Biometrika*, 73, 755-58.

[18] SAS institute Inc. *SAS/IML^{TM}, User's Guide for Personal Computers, Version 6 Edition.*, Cary, NC: Sas Institute Inc, 1985, 243 pp

[19] Schaefer, R.L., Roi, L.D. and Wolfe, R.A. (1984), "A ridge logistic estimator.", *Communications in Statistics-Theory and Methods*, 13(1), 99-113.

[20] Schaefer, R.L. (1986), "Alternative estimators in logistic regression when the data are collinear.", *Journal of Statistical Computations and Simulations*, 25, 75-91.

[21] Væth, M. (1985), "On the use of Wald's test in exponential families.", *International Statistical Review*, 53, 2, 199-214.

[22] Weissfeld, L.A. and Sereika, S.M. (1991), "A multicollinearity diagnostic for generalized linear models.", *Communications in Statistics-Theory and Methods*, 20(4), 1183-98.

Received June 1992; Revised February 1993