90
83

Ill-Conditioned Information Matrices and the Generalized Linear Model: An Asymptotically Biased Estimation Approach

by

Brian D. Marx

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Statistics

APPROVED:

Eric P. Smith, Chairman	Klaus Hinkelmann
Raymond H. Myers	Jeffrey B. Birch
George R. Terrell	Camilla A. Brooks

June 23, 1988

Blacksburg, Virginia

Ill-Conditioned Information Matrices and the Generalized Linear Model:

An Asymptotically Biased Estimation Approach

by

Brian D. Marx

Eric P. Smith, Chairman

Statistics

(ABSTRACT)

In the regression framework of the generalized linear model (Nelder and Wedderburn (1972)), iterative maximum likelihood parameter estimation is employed via the method of scoring. This iterative procedure involves a key matrix, the information matrix. Ill-conditioning of the information matrix can be responsible for making many desirable properties of the parameter estimates unattainable. Some asymptotically biased alternatives to maximum likelihood estimation are put forth which alleviate the detrimental effects of near singular information. Notions of ridge estimation (Hoerl and Kennard (1970a) and Schaefer (1979)), principal component estimation (Webster et al. (1974) and Schaefer (1986)), and Stein estimation (Stein (1960)) are extended into a regression setting utilizing any one of an entire class of response distributions.

Acknowledgements

First of all, I would like to extend my thanks to the members of my family who were consistently supportive and encouraged my graduate work, who stood by me in my efforts to reach a personal goal, who have allowed me to share a sense of accomplishment.

I would like to thank my wife with whom I have shared the past three years. She has been at my side keeping my attitude fresh, my outlook realistic, and my doctoral study a pleasant, exciting, enjoyable time of my life.

I would also like to thank all of my professors at Virginia Tech, Penn State, and Michigan State, especially my dissertation chairman, Dr. Eric Smith. Eric has helped me unravel and formulate a well defined, interesting, and tractable statistical problem which I will continue to pursue in the future. My thanks go to Dr. George Terrell. George's open door has helped me understand difficult concepts at difficult times. Dr. Ray Myers and Dr. Brad Skarpness have given me many of the statistical tools necessary to construct this work. has my thanks for her continual, most valuable computer help.

My thanks are not complete without mentioning two visiting professors during my stay in Virginia Tech's statistics department. , from University of New South

Wales, Australia and from Duke University were both instrumental in constructing the big picture of my dissertation. I have not forgotten my friends, fellow students, or whose diligence and competence at Script has given me a peace of mind.

Table of Contents

List of Illustrations

List of Tables

Chapter I

INTRODUCTION

1.1 SCOPE OF THE DISSERTATION

Nelder and Wedderburn (1972) have broadened the domain of the usual linear model. Their development of the generalized linear model (GLM) can accommodate a great variety of response variables, capturing distributional forms ranging from discrete to continuous, from symmetric to asymmetric. The model can be a design, regression, or a mixture. The GLM is extremely versatile and has a wealth of applications including standard multiple regression, analysis of variance, log-linear models, logistic and Poisson regression, among many others. A detailed overview of the generalized linear model is forthcoming in Chapter 2. Of course with a framework so well suited for a variety of applications, this dissertation must focus only on specific problems. These primarily include problems in the regression setting with all continuous explanatory variables. The GLM's regression parameters are typically estimated via an iterative maximum likelihood process. Consequently, a distributional form must be specified; one which is a member of the exponential family. Problems can exist with the maximum likelihood process particularly when a key matrix, entangled in the iterative procedure, is near-singular. The key matrix will be shown to be the information

matrix for the parameter estimates. Near-singular information matrices can often make desirable properties of accurate parameter estimation, precise prediction and testing with high power unattainable. Normal response data and the identity link simplify to least squares multiple regression and near singularity of the information matrix is equivalent to problems resulting from multicollinearity among the explanatory variables. However, this is not the case in general. Naturally, the next step is to develop alternate estimation procedures which alleviate problems associated with maximum likelihood in the presence of ill-conditioned information matrices and ideally restore desirable properties of the regression. Utilizing the fact that maximum likelihood estimates are asymptotically unbiased, various asymptotically biased estimation solutions will be developed and proposed as reasonable alternatives to maximum likelihood in the GLM. This dissertation concentrates on principal component, ridge and Stein estimation in the regression setting of the framework of the generalized linear model.

Schaefer (1979 and 1986) has had success in developing alternates to maximum likelihood for logistic regression. Recall that logistic regression assumes Bernoulli responses and is, in fact, a special case of the generalized linear model. In his 1979 dissertation, Schaefer has contributed a ridge estimate for the logistic model having all continuous explanatory variables. Somewhat later (1986), Schaefer further presented a principal component and a Stein estimation procedure, again for the logistic model with continuous regressors. Chapter 3 will show that these procedures tend to be particularly useful for accuracy in parameter estimation and can improve prediction abilities of the model for data combinations outside the mainstream of the original data. Nonetheless, maximum likelihood predicts well for internal data combinations. Also in Chapter 3, a review of literature and a comprehensive overview of logistic regression will be presented. Maximum likelihood, ridge and principal component estimators will be derived from likelihood theory and from Schaefer's techniques. In addition, the author has independently developed an iterative principal component technique which will be presented as an alternate to Schaefer's one step adjustment to maximum likelihood.

Chapter 4 stresses differences between multicollinearity among the explanatory variables and an ill-conditioned information matrix. Schaefer (1979) claims that the above are equivalent in logistic regression (as they are in standard multiple regression) in the limiting case of an exact deficiency among the regressors. It will be repeatedly noted that care must be taken in understanding the true relationship of collinearity to an ill-conditioned information matrix. Moreover, Chapter 4 presents some diagnostic tools for determining the severity of the ill-conditioning in the GLM. Among the diagnostics are generalizations to variance inflation factors, variance proportion decompositions, and condition indices. Some details regarding centering and scaling the data are also advised.

Chapter 5 extends the alternate parameter estimation techniques, given in Chapter 3 for logistic regression, to the framework of the generalized linear model. In particular, Chapter 5 discusses both of the mentioned principal component techniques (i.e. the one step adjustment to maximum likelihood and the iterative process) for the GLM. Also variable deletion is discussed as a viable option. Developments for hypothesis testing and rules for deletion of principal components are presented. Lastly, applications for principal component estimation are given for a variety of Bernoulli models, including logistic, probit, linear and extreme value.

Chapter 6 presents the development of a ridge estimator for the GLM. Various methods for choosing a shrinkage parameter, including a C_p based criterion, are generalized. Further, a general Stein estimation procedure is suggested.

Chapter 7 attempts to unify all the biased estimation techniques of the generalized linear model into one general class. The class is termed the Generalized Fractional Principal Component Estimators (GFPC). Some comparisons will be made among these estimators in a very broad manner.

Chapter 8 will present a simulation study to investigate the relative improvements using one estimation technique when compared to another. Parameter estimation techniques will be judged

on variance, bias, and mean square error. Other factors, such as sample size, number of explanatory variables, and severity of ill-conditioning of the information matrix will be examined for their respective impacts. An assortment of experimental settings are investigated, incorporating distributional forms of the response variable.

The concluding chapter of this dissertation will present some additional problems in the GLM which have not been addressed. Some suggestions will also be made as to present the GLM as a reasonable option to least squares regression.

1.2 NOTATIONS OF THE DISSERTATION

For the most part the author has tried to maintain a notation consistent with that of standard text books and major journals in the field of statistics. However due to a lack of consistency among statisticians, it is necessary to to mention some conventions used in the upcoming chapters.

The symbol N is reserved for the total number of observations, whereas the number of explanatory variables (regressors) is given by p (not including the constant term). The matrix of centered and scaled regressors is given by X and has dimension $N \times (p+1)$. In general, capital letters, such as A, denote a matrix with entries $\{a_{ij}\}$. The transpose of a matrix is given by A'. The inverse of a nonsingular square matrix, B , is given by B^{-1}, the generalized inverse is denoted as B^{-}, the trace is symbolized by tr(B). Lower case underscored letters, such as $\underline{x}_j$ or $\underline{a}_j$, are vectors. Typically observation regression vectors of the data X matrix are given by $\underline{x}'_i$ for $1 \leq i \leq N$.

Standard mathematical symbols are used throughout the dissertation. Among the most common are limit (lim), summation ($\sum$), integration ($\int$), differentiation ($d / d\beta$ or $\partial / \partial\beta$). The derivative with respect to θ of a function, $c(\theta)$, can be denoted by $c'(\theta)$ ($c''(\theta)$ for the second derivative, etc.).

Standard statistical notations include expected values of a random variable, $E(Y)$, variance of a random variable, $\mathrm{Var}(Y)$. Usually hatted vectors, e.g. $\hat{\underline{\beta}}$, refer to the maximum likelihood estimate of the unknown parameter vector, $\underline{\beta}$. Equality is denoted with '$=$', whereas approximate equality is denoted with '$\cong$'. The symbol '$\sim$' signifies 'is distributed as'. The symbol '$\dot{\sim}$' denotes 'is asymptotically distributed as'.

Chapter II

THE GENERALIZED LINEAR MODEL (GLM)

2.1 INTRODUCTION

Many regression problems can link the mean of the response variable's distribution to a linear combination of explanatory variables $x_1, x_2, \ldots, x_p$. If β_i are regression parameters and

$$\underline{x}'_i = [1, x_{i1}, x_{i2}, \ldots, x_{ip}]$$
$$\underline{\beta}' = [\beta_0, \beta_1, \ldots \beta_p],$$

then $\underline{x}'_i\underline{\beta}$ is a linear combination of the X explanatory variables.

When continuous responses, $\underline{Y}$, are modelled as a linear combination of explanatory variables, there may be cases when the data exhibits extreme nonnormal tendencies. For the model,

$$\begin{aligned} & E(Y) = \underline{\mu} = X\underline{\beta} \\ \text{where} \quad & \underline{Y} \sim N(X\underline{\beta}, \sigma^2 I), \end{aligned} \tag{2.1.1}$$

the assumption of normality in the response distribution may be inappropriate. Pregibon (1979) points out that, more often than not, least squares estimation is performed without regard to normality of the data. The distributional form of the response variable is not typically identified, unless large amounts of data are collected. Nonetheless, least squares can be an adequate estimation procedure if the data is reasonably symmetric, continuous and not heavy tailed.

However, there do exist various experimental settings when the standard linear model, defined in equation (2.1.1), is not appropriate for model building. For example, consider survival models utilizes the reciprocal mean lifetime expressed as a linear combination of explanatory variables, $\underline{\mu}^{-1} = X\underline{\beta}$. Log-linear models employ the log of cell means modelled as a linear function of parameters. Situations could arise when discrete responses are collected which are binomial in nature; in this case logistic regression,

$$\text{logit}(\underline{\pi}) = X\underline{\beta}, \tag{2.1.2}$$

would be appropriate for model building.

Responses having obvious asymmetry or a discrete nature require a method of estimation alternative to least squares. A more global approach to model building is given by the generalized linear model (Nelder and Wedderburn (1972) and McCullagh and Nelder (1983)). Equations (2.1.1) and (2.1.2) can be rewritten as

$$g(\underline{\mu}) = X\underline{\beta}, \tag{2.1.3}$$

where $\mu_i = E(Y_i)$. Notice that the function g links the systematic component, $\underline{x}'_i\underline{\beta}$, to the mean, μ_i.

As equation (2.1.3) suggests, the generalized linear model is formulated by

i) the distributional form of the response variable (in order to implement maximum likelihood estimation techniques);

ii) the choice of the linking function, which in most cases will be chosen to be the natural link (developed later);

iii) the choice of explanatory variables responsible for best linking the systematic component to the mean of the response variable.

A generalized linear model is constructed with the combination of the response's distributional form and the link function. If the response's distribution is nonnormal, then the mean response will be expressed nonlinear in $\underline{\beta}$.

2.2 EXPONENTIAL FAMILY OF DISTRIBUTIONS

A class of distributions capable of including many deal of discrete random variables (success-failure, counts, etc.), as well as a number of continuous distributions (normal, asymmetric, restricted on the domain, etc.) is the exponential class of distributions.

Consider a random variable, Y, having a distribution depending on a parameter θ of the form

$$f_Y(y;\theta) = \exp\{[a(y)b(\theta) + c(\theta)] / q(\phi) + d(y,\phi)\}, \qquad (2.2.1)$$

where a, b, c, d, q are known functions. If $b(\theta) = \theta$, then call θ the natural parameter. If $a(y) = y$, then equation (2.2.1) is in a simplified form developed later. The natural parameterization is presented in upcoming results. Let the nuisance parameter ϕ be a constant for all Y_i.

The exponential family is a rich class of distributions containing the normal, gamma, Poisson, binomial as well as many other distributions. Table 1 and Table 2 contain some distributions belonging to the exponential family. ϕ and w_i will be defined in equation (2.3.1). The Poisson model is commonly applied to log-linear models for contingency tables. The normal model is the basis

for analysis of variance, as well as testing in standard multiple regression. The binomial is often used in dose-response problems. Other particularly common exponential distributions are the gamma, used in life testing, and the inverse Gaussian, used in nonsymmetric regression.

The members of the exponential family have a general log-likelihood function of the form

$$l = [a(y)b(\theta) + c(\theta)] \,/\, q(\phi) + d(y, \phi). \tag{2.2.2}$$

The score, as defined in Bickel and Doksum (1976) with certain regulatory conditions, is

$$\begin{aligned} U &= \frac{\partial l}{\partial \theta} \\ E(U) &= 0 \\ \mathrm{Var}(U) &= E(U^2) = -E\left[\frac{\partial U}{\partial \theta}\right]. \end{aligned} \tag{2.2.3}$$

Notice

$$\begin{aligned} E\left[\frac{\partial l}{\partial \theta}\right] &= \int (1/f)(\partial f/\partial \theta) f \, dy \\ &= \frac{\partial}{\partial \theta}\int f \, dy = 0 \\ E\left[\frac{\partial^2 l}{\partial \theta^2}\right] &= \int (\partial^2 \ln f/\partial \theta^2) f \, dy \\ &= \int [\{f f'' - (f')^2\}/f^2] f \, dy \\ &= \int f'' \, dy - \int \{(f')^2/f\} \, dy \\ &= -\int \{(f')^2/f\} \, dy \end{aligned}$$

and

$$\begin{aligned} -E\left[\frac{\partial l}{\partial \theta}\right]^2 &= -\int (\partial \ln f/\partial \theta)^2 f \, dy \\ &= -\int (f'/f)^2 f \, dy \\ &= -\int \{(f')^2/f\} \, dy \,, \end{aligned}$$

where f denotes $f(y;\theta)$. For the exponential family, as given in equation (2.2.1),

Table 1. VARIOUS EXPONENTIAL FAMILY DISTRIBUTIONS

Distribution	Bounds
Poisson (λ)	$y \in N^+,\ \lambda > 0$
Neg. Binomial (r, p)	$y \in N,\ 0 < p < 1,\ r \in N^+$(known)
Binomial (n, π)	$0 \leq y \leq n \in N^+,\ 0 < \pi < 1$
Normal (μ, σ^2)	$-\infty \leq y \leq \infty,\ -\infty \leq \mu \leq \infty,\ \sigma^2 > 0$
Gamma (r, λ)	$y \geq 0,\ \lambda > 0,\ r > 0$
Unit Inverse Gaussian (μ, 1)	$y \geq 0,\ -\infty \leq \mu \leq \infty,\ \sigma^2 = 1$

Table 2. PARAMETERS IN THE GLM

Distribution	Natural Parameter $\theta = b(\theta)$	$c(\theta)$	$d(y, \phi)$	ϕ	w_i
Poisson(λ)	$\ln(\lambda)$	$e^{-\theta}$	$-\ln(y!)$	1	1
Neg. Binomial(r, p)	$\ln(1-p)$	$r\ln(1-e^{\theta})$	$\ln\binom{r+y-1}{y}$	1	1
Binomial(n, π)	$\ln\left(\frac{\pi}{1-\pi}\right)$	$-n\ln(1+e^{\theta})$	$\ln\binom{n}{y}$	1	1
Normal (μ, σ^2)	μ	$-\theta^2/2$	$-[y^2/\sigma^2 + \ln(2\pi\sigma^2)]/2$	σ^2	1
Gamma (r, λ)	$-\lambda/r$	$-\ln(-\theta)$	$(r-1)\ln(y) + r\ln(r)$ $-\ln(\Gamma(r))$	r^{-1}	1
Unit Inverse Gaussian (μ, 1)	$-\mu^{-2}/2$	$\sqrt{-2\theta}$	$-y^{-1}/2 - \ln(2\pi y^3)/2$	1	1

$$\frac{\partial l}{\partial \theta} = U = [a(y)b'(\theta) + c'(\theta)] \,/\, q(\phi)$$
$$\frac{\partial^2 l}{\partial \theta^2} = U' = [a(y)b''(\theta) + c''(\theta)] \,/\, q(\phi). \tag{2.2.4}$$

Thus

$$\mu = E[a(Y)] = -c'(\theta) \,/\, b'(\theta) \tag{2.2.5}$$

from equation (2.2.3). If $a(y) = y$ and $b(\theta) = \theta$, then $E(Y) = \mu = -c'(\theta)$.

It follows,

$$E(-U') = [-b''(\theta)\, E[a(y)] - c''(\theta)] \,/\, q(\phi)$$
$$\text{and} \quad \text{Var}(U) = E(U^2) = [b'(\theta) \,/\, q(\phi)]^2 \,\text{Var}[a(y)]. \tag{2.2.6}$$

Combining equations (2.2.5) and (2.2.6)

$$\text{Var}[a(Y)] = q(\phi)[b''(\theta)c'(\theta) - c''(\theta)b'(\theta)] \,/\, [b'(\theta)]^3. \tag{2.2.7}$$

Notice when $q(\phi) = 1$, $a(y) = y$ and $b(\theta) = \theta$, then $\text{Var}(Y) = -c'(\theta) = \frac{\partial}{\partial \theta} E(Y)$. Thus $\text{Var}(Y)$ can be thought of as the rate of change of the $E(Y)$ with respect to θ. For normally distributed responses, $\frac{\partial}{\partial \theta} E(Y)$ is constant, giving homogeneous fixed variance.

Naturally, if $Y_1, Y_2, ..., Y_N$ are independent random variables with the same exponential distribution, then the joint density is given by

$$f_Y(\underline{y}; \theta) = \exp\{[b(\theta)\sum_{i=1}^{N} a(y_i) + Nc(\theta)] \,/\, q(\phi) + \sum_{i=1}^{N} d(y_i, \phi)\}, \tag{2.2.8}$$

with $\sum_{i=1}^{N} a(y_i)$ the complete and sufficient statistic for $b(\theta)$ (Cox and Hinkley (1974)).

2.3 FORMULATION OF GENERALIZED LINEAR MODELS (GLM)

The framework of the GLM allows the response variable to have any distribution from the exponential family when $a(y) = y$. Thus Y can be from one of many discrete or continuous distributions. Furthermore, the relationship between the response, Y, and the explanatory variables does not have to be linear, as in the usual regression setting.

Consider $Y_1, Y_2, ..., Y_N$ as independent random variables each from the exponential family with the following conditions (Nelder and Wedderburn (1972)):

i) the scale parameter $q(\phi) = \phi / w_i$, where w_i are known weights;

ii) the distribution of each Y_i is such that $a(y) = y$ and depends on a single parameter θ_i, that is the Y_i are not identically distributed and

$$f(y_i; \theta_i) = \exp\{[y_i b_i(\theta_i) + c_i(\theta_i)]w_i / \phi + d_i(y_i, \phi)\}; \qquad (2.3.1)$$

iii) the form of the distribution of all the Y_is are the same so that the subscripts on b, c, d are not needed.

Write the joint probability density function of $Y_1, Y_2, ..., Y_N$ as

$$f(\underline{y}; \underline{\theta}) = \exp\left[\sum_{i=1}^{N}\{[y_i b(\theta_i) + c(\theta_i)]w_i / \phi + d(y_i, \phi)\}\right],$$
$$l(\underline{\theta}; \underline{y}) = \sum_{i=1}^{N}\{[y_i b(\theta_i) + c(\theta_i)]w_i / \phi + d(y_i, \phi)\}. \qquad (2.3.2)$$

Equation (2.3.2) is overspecified. That is there are as many parameters to estimate as there are observations. Thus, for the generalized linear model, consider a smaller set of parameters, as given

in equation (2.1.3), $\beta_0, \beta_1, \ldots, \beta_p$ $(p < N)$. Given the set of p explanatory variables, the generalized linear model utilizes the relationship,

$$g(\mu_i) = \underline{x}'_i\underline{\beta}, \tag{2.3.3}$$

satisfying:

i) $\mu_i = E(Y_i)$;

ii) g is a monotone, twice differentiable function with an inverse (i.e. $g^{-1} = h$ exists) called the link function;

iii) $\underline{x}'_i$ is a $(p+1) \times 1$ row vector of regressor variables (later developments will require continuous covariates and a constant);

iv) $\underline{\beta}$ is the unknown parameter vector;

v) the estimation of $\underline{\beta}$ does not depend on having an estimate of ϕ.

Notice that in the special case when the systematic component

$$g(\mu) = \underline{x}'_i\underline{\beta} = b(\theta) = \theta = -c'^{-1}(\mu),$$

then equation (2.3.2) can be expressed in terms of the natural parameter θ. Hence the natural link.

In terms of the $(p+1)$ dimensional $\underline{\beta}$ vector using the natural link, the log-likelihood becomes

$$l(X\underline{\beta}; \underline{y}) = \sum_{i=1}^{N} \{[y_i\underline{x}'_i\underline{\beta} + c(\underline{x}'_i\underline{\beta})]w_i / \phi + d(y_i, \phi)\}. \tag{2.3.4}$$

Using the natural link function $b(\theta) = \theta = \underline{x}'\underline{\beta}$ and setting the derivative of equation (2.3.4) to zero, "normal-like" equations can be given as

$$\begin{aligned} \underline{Q} = \frac{\partial}{\partial \underline{\beta}} l(X\underline{\beta}; \underline{y}) &= \sum_{i=1}^{N} x_{ij}(y_i + c'(\underline{x}'_i\underline{\beta}))w_i / \phi \\ &= \sum_{i=1}^{N} x_{ij}(y_i - h(\underline{x}'_i\underline{\beta}))w_i / \phi, \end{aligned} \tag{2.3.5}$$

for $j = 0, 1, 2, ..., p$ and $h = g^{-1}$. Equivalently,

$$\underline{Q} = X'(\underline{y} - h(X\underline{\beta})). \tag{2.3.6}$$

2.4 ESTIMATION OF PARAMETERS IN THE GLM

Recall that the exponential family log-likelihood function for independent Y_i is

$$\begin{aligned} l(\underline{\theta}; \underline{y}) &= \sum_{i=1}^{N} \{[y_i b(\theta_i) + c(\theta_i)]w_i / \phi\} + \sum_{i=1}^{N} d(y_i, \phi) \\ l(\theta_i; y_i) &= \{[y_i b(\theta_i) + c(\theta_i)]w_i / \phi\} + d(y_i, \phi), \end{aligned}$$

where $E(Y_i) = \mu_i = -c'(\theta_i) / b'(\theta_i)$. Further, g is a monotone and twice differentiable function such that

$$g(\mu_i) = \underline{x}'_i\underline{\beta} = \eta_i. \tag{2.4.1}$$

One advantage of employing maximum likelihood procedures for estimation of $\underline{\beta}$ is that the exponential family ensures an unique solution to the set of equations $\frac{\partial l}{\partial \underline{\beta}} = 0$ (Cox and Hinkley (1974)). Notice by the chain rule that

$$\frac{\partial l}{\partial \underline{\beta}} = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \underline{\beta}}$$
$$= \sum_{i=1}^{N} x_i h'(\eta_i) \frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} \qquad (2.4.2)$$
$$= \sum_{i=1}^{N} x_i ,$$

where $h = g^{-1}$. Equation (2.4.2) follows from

$$\frac{\partial l_i}{\partial \theta_i} = [y_i b'(\theta_i) + c'(\theta_i)] w_i / \phi$$
$$= b'(\theta_i)[y_i - \mu_i] w_i / \phi$$
$$\frac{\partial \theta_i}{\partial \mu_i} = \left[\frac{\partial \mu_i}{\partial \theta_i} \right]^{-1} = [b'(\theta_i)]^2 / [b''(\theta_i)c'(\theta_i) - b'(\theta_i)c''(\theta_i)] \qquad \text{from eq. (2.2.5)}$$
$$= \phi / [w_i b'(\theta_i)\mathrm{Var}(Y_i)] \qquad \text{from eq. (2.2.7)}$$
$$\frac{\partial \mu_i}{\partial \eta_i} = h'(\eta_i)$$
$$\frac{\partial \eta_i}{\partial \underline{\beta}} = x_i .$$

In equating equation (2.4.2) to zero, "normal-like" equations are formed for any general link function. The Newton-Raphson approach uses the following Taylor series expansion of $\partial l / \partial \underline{\beta}$ about $\underline{\beta}_0$,

$$\frac{\partial l}{\partial \underline{\beta}} \cong \frac{\partial l}{\partial \underline{\beta}} \Big|_{\underline{\beta} = \underline{\beta}_0} + \frac{\partial^2 l}{\partial \underline{\beta} \partial \underline{\beta}'} \Big|_{\underline{\beta} = \underline{\beta}_0} (\underline{\beta} - \underline{\beta}_0) = 0,$$

where

$$\frac{\partial^2 l}{\partial \underline{\beta} \partial \underline{\beta}'} = \sum_{i=1}^{N} x_i (y_i - \mu_i) \frac{\partial}{\partial \underline{\beta}'} [h'(\eta_i) / \mathrm{Var}(Y_i)] - \sum_{i=1}^{N} [x_i h'(\eta_i) / \mathrm{Var}(Y)] \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \underline{\beta}'} = H,$$

from equation (2.4.2). H is often termed the Hessian matrix. This implies the iterative scheme

$$\hat{\underline{\beta}}_t = \hat{\underline{\beta}}_{t-1} - \left[\frac{\partial^2 l}{\partial \underline{\beta} \partial \underline{\beta}'} \right]^{-1}_{\underline{\beta} = \hat{\underline{\beta}}_{t-1}} \frac{\partial l}{\partial \underline{\beta}} \Big|_{\underline{\beta} = \hat{\underline{\beta}}_{t-1}}, \tag{2.4.3}$$

Equation (2.4.3) can be simplified computationally by replacing

$$\frac{\partial^2 l}{\partial \underline{\beta} \partial \underline{\beta}'} \text{ with } E\left[\frac{\partial^2 l}{\partial \underline{\beta} \partial \underline{\beta}'} \right] = -E\left[\frac{\partial l}{\partial \underline{\beta}} \quad \frac{\partial l}{\partial \underline{\beta}'} \right] = -\sum_{i=1}^{N} x_i x_i' [h'(\eta_i)]^2 / \text{Var}(Y_i) = -\Phi,$$

where Φ is the information matrix, which is called the method of scoring. It follows that the method of scoring has the iterative scheme

$$\hat{\underline{\beta}}_t = \hat{\underline{\beta}}_{t-1} + [\Phi_{t-1}]^{-1} \frac{\partial l}{\partial \underline{\beta}} \Big|_{\underline{\beta} = \hat{\underline{\beta}}_{t-1}}, \tag{2.4.4}$$

where t denotes the iteration step. The method of scoring corresponds to the ordinary least squares solutions when the identity link is used with normal data.

Notice that the contribution of Y_i to Φ_{jk}, denoted by Φ^i_{jk}, where $\Phi_{jk} = \sum_{i=1}^{N} \Phi^i_{jk}$, is given by

$$\begin{aligned} \Phi^i_{jk} &= E\left[\frac{\partial l_i}{\partial \beta_j} \quad \frac{\partial l_i}{\partial \beta_k} \right] \\ &= E\left[\frac{(y_i - \mu_i)^2 x_{ij} x_{ik}}{[\text{Var}(Y_i)]^2} \left[\frac{\partial \mu_i}{\partial \eta_i} \right]^2 \right] \\ &= \frac{x_{ij} x_{ik} [h'(\eta_i)]^2}{\text{Var}(Y_i)}. \end{aligned} \tag{2.4.5}$$

Thus

$$\Phi_{jk} = \sum_{i=1}^{N} x_{ij} k_{ii}^{-1} x_{ik}, \tag{2.4.6}$$

where $k_{ii}^{-1} = [h'(\eta_i)]^2 / \text{Var}(Y_i)$. Therefore,

$$\Phi = X'K^{-1}X, \quad \text{where} \quad K^{-1} = \text{diag}\{k_{ii}^{-1}\}.$$

Notice when the natural link function is used the $K^{-1} = \text{diag}\{\text{Var}(Y_i)\}$ is particularly simple. This follows since $\eta = \theta = -c'^{-1}(\mu)$ and $\partial\mu / \partial\eta = \text{Var}(Y)$.

Now the iterative scheme in equation (2.4.4) can be re-expressed, using equation (2.4.2), as

$$\begin{aligned}
\hat{\underline{\beta}}_t &= \hat{\underline{\beta}}_{t-1} + (X'\hat{K}_{t-1}^{-1}X)^{-1}\left[\sum_{i=1}^{N} \underline{x}_i \hat{k}_{ii}^{-1}\,(y_i - \hat{\mu}_i)\,\frac{\partial\eta_i}{\partial\mu_i}\right]_{t-1} \\
&= (X'\hat{K}_{t-1}^{-1}X)^{-1}\left[\sum_{i=1}^{N} \hat{k}_{ii}^{-1}\underline{x}_i\left[\underline{x}'_i\hat{\underline{\beta}}_{t-1} + (y_i - \hat{\mu}_i)\,\frac{\partial\eta_i}{\partial\mu_i}\right]\right]_{t-1} \\
&= (X'\hat{K}_{t-1}^{-1}X)^{-1}X'\hat{K}_{t-1}^{-1}\underline{y}^*_{t-1},
\end{aligned} \tag{2.4.7}$$

where $y_i^* = \eta_i + (y_i - \mu_i)(\partial\eta_i / \partial\mu_i)$ evaluated at $\hat{\underline{\beta}}_{t-1}$. Note that in the most general setting, the estimate of K^{-1} and y_i^* must be updated at each iteration step until convergence of the parameter estimate since they are a function of the iterated $\underline{\eta}_{t-1}$. Observe that

$$\text{Var}(\underline{y}^*) \cong K \quad \text{and} \quad \hat{\underline{\beta}} \sim N(\underline{\beta}, (X'K^{-1}X)^{-1}). \tag{2.4.8}$$

Asymptotic distributional properties will be developed in section 2.5. The iterative scheme, given in equation (2.4.7), is consistent with the Gauss-Newton procedure outlined in section 3.5 for the logit model. Also notice the similarity of equation (2.4.7) to that of reweighted least squares. Nelder and Wedderburn (1972) have shown that the solutions to the "normal-like" equations given in equation (2.3.5) are equivalent to an iterative weighted least squares solution working with the variable $\underline{y}^*$. Previous to Nelder and Wedderburn's 1972 work, Fisher (1935) had used this iterative weighted least squares scheme in the special case of binomial data using the probit link. Somewhat later, Finney (1947) used a similar approach to that of Fisher's with binomial data but with a logit link function for fitting dose response curves.

As an example, consider the common case of the generalized linear model, that is the normal model,

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \tag{2.4.9}$$

where the $\varepsilon_i \sim N(0, \sigma^2)$ and independent. It follows that $Y_i \sim N(\underline{x}'_i\underline{\beta}, \sigma^2)$ and $\mu_i = \underline{x}'_i\underline{\beta}$. When $\phi = \sigma^2$ is known and $w_i = 1$ for all i, the normal distribution is a member of the one-parameter exponential family. In this example, $g(\mu_i) = \mu_i = \theta_i$; thus g is the identity link.

Consider, as another example, $Y_i \sim \text{Poisson}(\lambda_i)$. Using the natural parameter as the linking function, equation (2.4.10) holds.

$$\theta_i = g(\mu_i) = \ln(\lambda_i) = \underline{x}'_i\underline{\beta}. \tag{2.4.10}$$

Hence $\mu_i = \lambda_i = h(\underline{x}'_i\underline{\beta}) = \exp(\underline{x}'_i\underline{\beta})$.

As a third example, the binomial-logit model with the natural link function gives

$$g(\mu_i) = \text{logit}(\pi_i) = \underline{x}'_i\underline{\beta}. \tag{2.4.11}$$

It follows then that $\mu_i = h(\eta_i) = n(1 + \exp(-\eta_i))^{-1}$.

The negative binomial is also a member of the exponential family when the parameter r is treated as known.

$$f(y) = \exp\left[y_i \ln(1-p) + r\ln(p) + \ln\binom{r+y-1}{y}\right].$$

$$\theta_i = g(\mu_i) = \ln\left[p_i(1-p_i)/p_i\right] = \underline{x}'_i\underline{\beta} = \eta_i.$$

$$\mu_i = (1-p_i)/p_i = h(\eta_i) = e^{\eta_i}/(1-e^{\eta_i}).$$

Notice that when $r = 1$ the negative binomial is reduced to the geometric distribution.

The use of the Gamma (r, λ), as an illustrative example, is instructive but not as straight forward as the previous examples. Recall that for $y > 0$, $\lambda > 0$, and $r > 0$,

$$\begin{aligned} f(y) &= \lambda^r \, y^{r-1} e^{-y\lambda} / \Gamma(r) \\ &= \exp\{-y\lambda + (r-1)\ln(y) + r\ln(\lambda) - \ln\Gamma(r)\}. \end{aligned} \tag{2.4.12}$$

Observe that $E(Y) = r/\lambda = -c'(\theta)$ and $\mathrm{Var}(Y) = r/\lambda^2 = -q(\phi)c''(\theta)$. See equation (2.2.7). From the form of the exponential family,

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

the above equations yield $\theta / q(\phi) = -\lambda$. Thus $-\theta c''(\theta) = c'(\theta) = -r/\lambda$. This implies that $c'(\theta) \propto \theta^{-1}$. Hence $\theta = -\lambda / r$, $q(\phi) = r^{-1} = -\mu^{-1}$, $c(\theta) = -\ln(-\theta)$ giving

$$f(y) = \exp\{[y\theta - c(\theta)] / q(\phi) + (r-1)\ln y + r\ln r - \ln\Gamma(r)\}.$$

A common link function is $\mu = \eta^{-1}$. Note that $\mathrm{Var}(Y) = q(\phi)\mu^2$ (McGilchrist (1987)). See Table 3.

2.5 INFERENCES CONCERNING THE GENERALIZED LINEAR MODEL

For the generalized linear model, define the score with respect to β_j to be

$$U_j = \frac{\partial l}{\partial \beta_j} \qquad j = 0, 1, \ldots, p. \tag{2.5.1}$$

In obtaining the maximum likelihood parameter estimates, $\underline{U} = (U_0, U_1, \ldots, U_p)'$ is set to zero, where

$$E(\underline{U}) = \underline{0} \quad \text{and} \quad E(\underline{U}\,\underline{U}') = \Phi. \tag{2.5.2}$$

Table 3. NATURAL LINK FUNCTIONS

Distribution	Natural Link $g(\mu) = \theta = \eta$	$h(\eta)$	k_{ii}^{-1}
Poisson (λ)	$\ln(\lambda)$	$\exp(\eta)$	$\exp(\eta) > 0$
Neg. Binomial (r, p)	$\ln(1-p)$	$-(1-\exp(-\eta))^{-1}$	$e^{\eta}/(1-e^{\eta})^2 > 0$
Binomial (n, π)	$\ln\left(\frac{\pi}{1-\pi}\right)$	$(1 + \exp(-\eta))^{-1}$	$0 < e^{\eta}/(1+e^{\eta})^2 < .25$
Normal (μ, σ^2)	μ / σ^2	$\eta\sigma^2$	$\sigma^2 > 0$
Unit Gamma ($1,\lambda$)	$-\lambda$	$-\eta^{-1}$	$\eta^{-2} > 0$
Unit Inverse Gaussian	$-\mu^{-2}/2$	$(-2\eta)^2$	$64\,\eta^2 > 0$

By an extension of the Central Limit Theorem (Feller (1966)), the asymptotic distribution of $\underline{U}$ is multivariate $N(\underline{0}, \Phi)$; hence

$$\underline{U}'\Phi^{-1}\underline{U} \sim \chi^2_{p+1,0}. \tag{2.5.3}$$

The multivariate central limit theorem can be found in Rao (1967). The application to the vector of scores follows. Consider equation (2.4.2). Let $\underline{z}_1, \ldots, \underline{z}_N$ be a sequence of independent $(p+1)$ dimensional random vectors such that $E(\underline{z}_i) = 0$ and the dispersion matrix $D(\underline{z}_i) = \Phi$. Define

$$\underline{z}_i = x_i h'(\eta_i) \frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} ,$$

for $1 \le i \le N$. Now $E(\underline{z}_i) = 0$ since $\underline{z}_i$ is a vector of scores (see equation (2.2.3)). Further $D(\underline{z}_i) = \Phi \neq 0$ (see equation (2.4.6)). The $\underline{z}_i$ are independent since the Y_i are independent. Moreover, $\Phi = X'K^{-1}X$ is finite, nonnull by assumptions given in section 3.6. Standard Lindeberg conditions outlined by Rao are met. Hence, the above result, given in equation (2.5.3),

$$\underline{U} = \sum_{i=1}^{N} \underline{z}_i \sim N(\underline{0}, \Phi).$$

Asymptotic normality of scores give asymptotic normality of maximum likelihood estimates of the parameters, as shown in equation (2.5.7).

When convergence is obtained using the iterative equation (2.4.7), consider the unique maximum likelihood estimate, $\hat{\underline{\beta}}$. Define $\underline{\beta}$ to be the true parameter vector. The Taylor series expansion of $U(\underline{\beta})$ about $\hat{\underline{\beta}}$ (Dobson (1983)) is

$$U(\underline{\beta}) \cong U(\hat{\underline{\beta}}) + H(\hat{\underline{\beta}})(\underline{\beta} - \hat{\underline{\beta}}), \tag{2.5.4}$$

where H is the Hessian matrix evaluated at the maximum likelihood estimates, $\hat{\underline{\beta}}$. Thus,

$$U(\underline{\beta}) \cong U(\hat{\underline{\beta}}) - \Phi\,(\underline{\beta} - \hat{\underline{\beta}}), \tag{2.5.5}$$

since $\Phi = E(-H)$. This implies that

$$(\hat{\underline{\beta}} - \underline{\beta}) \cong \Phi^{-1} \underline{U}, \tag{2.5.6}$$

since $U(\hat{\underline{\beta}}) = 0$ by definition. By taking expectations of both sides of equation (2.5.6),

$$E(\hat{\underline{\beta}}) = \underline{\beta} \quad \text{asymptotically,}$$

since $E(U) = 0$ from equation (2.2.3). Similarly

$$E[(\hat{\underline{\beta}} - \underline{\beta})(\hat{\underline{\beta}} - \underline{\beta})'] = \Phi^{-1}\, E(\underline{U}\,\underline{U}')\Phi^{-1} = \Phi^{-1},$$

since $\Phi = E(\underline{U}\,\underline{U}')$ and symmetric (provided that Φ is nonsingular). Thus for large samples

$$\begin{aligned} \hat{\underline{\beta}} &\dot{\sim} N(\underline{\beta},\ \Phi^{-1}) \\ (\hat{\underline{\beta}} - \underline{\beta})'\,\Phi\,(\hat{\underline{\beta}} - \underline{\beta}) &\dot{\sim} \chi^2_{p+1,\,0}, \end{aligned} \tag{2.5.7}$$

where the mean is of order N^{-1} and the variance of order N^{-2} (Bartlett (1953)). Note that for normal response data, the distributions are exact rather than asymptotic (Dobson (1983)).

2.6 HYPOTHESIS TESTING FOR THE GENERALIZED LINEAR MODEL

Define the overspecified or maximal model as having as many parameters as observations. Thus the maximal model can be thought of as having the parameter vector

$$\underline{\beta}_{\max} = [\beta_1, \beta_2, \cdots, \beta_N]'.$$

To determine whether another model with ($p + 1 < N$) parameters $\underline{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]'$ is adequate relative to the maximal model, it is reasonable to compare their likelihood functions. If $L(\underline{\beta}; \underline{y}) \cong L(\underline{\beta}_{max}; \underline{y})$, then the model describes the data well. However, if $L(\underline{\beta}; \underline{y}) << L(\underline{\beta}_{max}; \underline{y})$, then the model is poor relative to the maximal model. This suggests the likelihood ratio test using the statistic

$$\begin{aligned} \lambda &= L(\hat{\underline{\beta}}_{max}; \underline{y}) \,/\, L(\hat{\underline{\beta}}; \underline{y}) \\ \text{or} \quad \ln \lambda &= l(\hat{\underline{\beta}}_{max}; \underline{y}) - l(\hat{\underline{\beta}}; \underline{y}) \end{aligned} \tag{2.6.1}$$

If λ is large, then claim $\underline{\beta}$ is a poor model.

The sampling distribution of ln λ can be approximated by the following Taylor series expansion of $l(\underline{\beta}; \underline{y})$ about the maximum likelihood estimator $\hat{\underline{\beta}}$.

$$l(\underline{\beta}; \underline{y}) \cong l(\hat{\underline{\beta}}; \underline{y}) + (\underline{\beta} - \hat{\underline{\beta}})' \, U(\hat{\underline{\beta}}) + (1/2)(\underline{\beta} - \hat{\underline{\beta}})' \, H(\hat{\underline{\beta}}) \, (\underline{\beta} - \hat{\underline{\beta}}) \,, \tag{2.6.2}$$

where $H(\hat{\underline{\beta}})$ is the Hessian matrix evaluated at the maximum likelihood estimate. Recall that $U(\hat{\underline{\beta}}) = \underline{0}$ by definition and $\Phi = -E(H)$ for large samples. In giving a distributional result, equation (2.6.2) can be rewritten as

$$2\,[l(\hat{\underline{\beta}}; \underline{y}) - l(\underline{\beta}; \underline{y})] = (\underline{\beta} - \hat{\underline{\beta}})' \, \Phi \, (\underline{\beta} - \hat{\underline{\beta}}) \sim \chi^2_{p+1,\,0}, \tag{2.6.3}$$

from equation (2.5.7).

Utilizing the asymptotic result in equation (2.6.3), a goodness-of-fit measure can be constructed. Nelder and Wedderburn (1972) define the scaled deviance as

$$S = 2 \ln \lambda = 2\,[l(\hat{\underline{\beta}}_{max}; \underline{y}) - l(\hat{\underline{\beta}}; \underline{y})], \tag{2.6.4}$$

where $\hat{\underline{\beta}}$ is maximum likelihood based on p explanatory variables and a constant. Notice that the scaled deviance can be written as

$$S = 2\sum_{i=1}^{N} [l(x'_i\hat{\underline{\beta}}_{\max};\ y_i) - l(x'_i\hat{\underline{\beta}}; y_i)]$$
$$= 2\sum_{i=1}^{N} d_i^2.$$

As an example, the scaled deviance for Poisson responses, as defined by Bishop et al. (1975), is the G^2 -statistic (since $\phi = w_i = 1$),

$$G^2 = S_{PSN} = 2\sum_{i=1}^{N} \{y_i \ln[c'(x'_i\hat{\underline{\beta}}) / y_i] - c'(x'_i\hat{\underline{\beta}}) - y_i\}.$$

The expression $\sum_{i=1}^{N} [-c'(x'_i\hat{\underline{\beta}}) - y_i]$ sums to zero if the natural link is used. The scaled deviance can be broken down into the following components

$$S = 2\{[l(\hat{\underline{\beta}}_{\max};\ \underline{y}) - l(\underline{\beta}_{\max};\ \underline{y})] - [l(\hat{\underline{\beta}};\ \underline{y}) - l(\underline{\beta};\ \underline{y})] + [l(\underline{\beta}_{\max};\ \underline{y}) - l(\underline{\beta};\ \underline{y})]\}$$
$$\dot{\sim}\ \chi^2_{N-p-1,0}, \tag{2.6.5}$$

when $l(\underline{\beta}_{\max};\ \underline{y}) \cong l(\underline{\beta};\ \underline{y})$ (i.e. the data represents the maximal model well); otherwise, equation (2.6.5) has an asymptotic noncentral χ^2 distribution.

In testing a current model against a full model, an useful hypothesis test is of the form

$$H_0: \underline{\beta} = \underline{\beta}_C \quad (q+1)$$
$$H_1: \underline{\beta} = \underline{\beta}_F \quad (p+1), \tag{2.6.6}$$

where $q < p < N$ and H_0 is nested in H_1. The subscript F denotes the full model whereas the C denotes the current model of interest. H_0 is tested against the alternative by using the difference in the log-likelihood statistics, producing a scaled deviance

$$S^* = S_C - S_F = 2[l(\hat{\underline{\beta}}_F; \underline{y}) - l(\hat{\underline{\beta}}_C; \underline{y})]. \tag{2.6.7}$$

If both H_0 and H_1 describe the data adequately relative to the maximal model, then

$$\begin{aligned} S_C &\dot{\sim} \chi^2_{N-q-1,0}, \\ S_F &\dot{\sim} \chi^2_{N-p-1,0}, \\ \text{and } S^* &\dot{\sim} \chi^2_{p-q,0}, \end{aligned} \tag{2.6.8}$$

provided S^* and S_F are independent. Notice that if $q+1=p$, then $S^* \dot{\sim} \chi^2_{1,0}$. The degrees of freedom, $p-q$, can be thought of as the number of restrictions imposed on the null hypothesis.

Perhaps a more common test in practice would be the one of the form

$$H_o: C\underline{\beta} = \underline{0}, \tag{2.6.9}$$

where C is a $q \times (p+1)$ matrix of constants. In particular, the test for the deletion of a single parameter would yield the choice of $C = (0, ...,0, 1, 0, ...,0)$.

It follows under H_o,

$$\hat{\underline{\beta}}' C'(C\Phi^{-1}C')^{-1}C\hat{\underline{\beta}} \dot{\sim} \chi^2_q. \tag{2.6.12}$$

Hence, the test for a single parameter simplifies to

$$\hat{\beta}_j^2 / \Phi_{jj}^{-1} \dot{\sim} \chi^2_1. \tag{2.6.13}$$

The above statistic is compared to the appropriate percentage point of the asymptotic chi-square distribution.

2.7 DEVIANCE VERSUS SCALED DEVIANCE

Notice that the scaled deviance in equation (2.6.7) can be rewritten as

$$
\begin{aligned}
S^* &= 2\,[l_F - l_C] \\
&= (2/\phi)\sum_{i=1}^{N} w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) + c(\tilde{\theta}_i) - c(\hat{\theta}_i)] \\
&\dot{\sim} \chi^2_{N-k},
\end{aligned}
\tag{2.7.1}
$$

where $\tilde{\underline{\theta}} = X_F\tilde{\underline{\beta}}_F$ and $\hat{\underline{\theta}} = X_C\hat{\underline{\beta}}_C$. A difficulty in using S^* as a practical measure of goodness-of-fit is that it is a function of ϕ, which is unknown for two-parameter families. Hence an estimate of ϕ is desired. Define

$$
D = \phi S^* = 2\sum_{i=1}^{N} w_i[\, y_i(\tilde{\theta}_i - \hat{\theta}_i) + c(\tilde{\theta}_i) - c(\hat{\theta}_i)]. \tag{2.7.2}
$$

D is termed the deviance of the current model relative to the full model. D is a known quantity when given the data and the maximum likelihood estimates, $\hat{\underline{\beta}}$. The deviance will be shown to be a common measure of goodness-of-fit. To estimate ϕ, D is computed using an overspecified full model of rank N and a current model of rank k. The dimension k is determined by choosing the largest reasonable current model. Since S^* is distributed χ^2_{N-k}, an estimate of ϕ is given by

$$
\hat{\phi} = D_{N,k} / (N-k) \ . \tag{2.7.3}
$$

2.8 GOODNESS OF FIT IN THE GENERALIZED LINEAR MODEL

In the usual linear model with the identity link, certainly one of the most common measures of goodness-of-fit is the deviance which simplifies to the sum of squared error, SSE . SSE is a measure of how well the data is represented by the model. The quantity SSE is given by

$$SSE = \sum_{i=1}^{N} (y_i - x'_i\hat{\beta})^2 = \sum_{i=1}^{N} (y_i - \hat{\mu}_i)^2. \tag{2.8.1}$$

SSE will be zero and without degrees of freedom when the data is perfectly fit by an overspecified model which assigns one parameter for each observation. On the other hand, *SSE* reaches its other extreme when only a constant term is fit. That is in the absence of explanatory variables, the model fits the mean response and *SSE* is the total variance of Y with $N-1$ degrees of freedom. An intermediate number of parameters, p, is typically fit to the data, where $1 < p < N$. Of course, other statistics and diagnostics based on prediction and parameter estimation will have to be taken into consideration when choosing an appropriate model for the researcher's use.

Pregibon (1979) points out a natural extension of the *SSE*, in the generalized linear model, is the relative sum of squared deviations, given by

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2} \\ &= -\sum_{i=1}^{N} \frac{(y_i + c'(x'_i\hat{\beta}))^2}{c''(x'_i\hat{\beta})}. \end{aligned} \tag{2.8.2}$$

The natural link function is required for the last expression to hold. Apart from σ^2, equation (2.8.2) simplifies to *SSE* for normal data. Moreover, the versatility of χ^2 is seen for binomial and Poisson responses. For these data, χ^2 is the standard goodness-of-fit test statistic used in log-linear models, multinomial data, and contingency tables. That is

$$\chi^2 = \sum_{i=1}^{N} (o_i - e_i)^2 / e_i, \tag{2.8.3}$$

where o_i and e_i denote the observed and expected cell frequencies respectively.

Certain distributional properties do hold for the χ^2 statistic. For normal responses with the identity link, χ^2 has an exact χ^2_{N-p-1} distribution. For some nonnormal exponential-family response data, χ^2 varies in how well an approximate χ^2 distribution is followed. Despite the fact that the χ^2 statistic can be poorly approximated by a χ^2 distribution, Pregibon (1979) gives examples where this is of little consequence. Asymptotic arguments suggest that if the primary use of χ^2 is to compare competing models rather than an assessment of fit for a particular model, then $\Delta\chi^2 = \chi^2_F - \chi^2_C$ is approximated well by a χ^2_{p-q} distribution, where $p-q$ is the number of restrictions imposed on the full model.

2.9 SCREENING REGRESSORS IN THE GENERALIZED LINEAR MODEL

In least squares regression with normal responses, various diagnostics have been developed to effectively screen explanatory variables in the pursuit of the best possible subset model. Among these diagnostics is the all possible regressions computer routine which can effectively and quite quickly entertain up to $k = 10$ regressors. All of the $2^k - 1$ possible regressions are computed while giving several criteria for the researcher to base his or her choice on. The criteria given in the *SAS* pressall macro, developed at the Department of Statistics, Virginia Polytechnic Institute and State University (Myers, S. (1984)), are *MSE*, C_p, R^2, and *PRESS*. Combining the pressall routine with collinearity and outlier diagnostics (with a $k \leq 10$), more often than not the model selected is superior to that selected by a forward selection or backward elimination stepwise procedure which ignores problems associated with multicollinearity.

Lawless and Singhal (1978) have developed an all possible regressions routine to efficiently screen explanatory variables in nonnormal regression models. Since the algorithm is quite general, it certainly includes the class of generalized linear models. Scaled deviance is used as a criterion for the best subset model. Two approximations to scaled deviance are presented to speed up the computation. The paper includes examples from exponential, Poisson and logistic regression and

with $k = 8$. The examples show the accuracy of the two approximations to scaled deviance in determining the best subset model.

Alternatively to the all possible regressions routine, C_p has been mentioned as a method to assess the quality of a subset model (see Mallows (1973)). The C_p statistic is oriented toward the predictive capabilities of the model by giving the mean squared error for a p-regressor candidate model. Define

$$\begin{aligned} C_p &= \sum_{i=1}^{N} \frac{MSE(\hat{y}(x_i))}{\sigma^2} \\ &= \sum_{i=1}^{N} \frac{\mathrm{Var}(\hat{y}(x_i)) + [\mathrm{Bias}(\hat{y}(x_i))]^2}{\sigma^2} \\ &= (p+1) + \frac{(s^2 - \sigma^2)(N-p-1)}{\sigma^2} \\ &\cong \frac{SSE_p}{\hat{\sigma}^2} - N + 2(p+1), \end{aligned} \tag{2.9.1}$$

where SSE_p is the sum of squares error for the p regressor subset model and $\hat{\sigma}^2$ is the mean squared error for the full k regressor model, and $p+1$ corresponds to the $\mathrm{tr}(H) = \mathrm{tr}(X(X'X)^{-1}X')$.

C_p attempts to strike the proper balance between the impact of overfitting (i.e. inflation of $\mathrm{Var}(\hat{y})$) and the impact of underfitting (i.e. inflation of $\mathrm{Bias}(\hat{y})$). In fact, C_p is a compromise between the complexity of the model (p) and goodness-of-fit (SSE_p). Plots of C_p vs. p will summarize the candidate models. The value $C_p = p$ suggests that bias is absent. However, candidate models having a C_p less than p usually suggests that s_p^2 is less than $\hat{\sigma}^2$. Since C_p has an approximate expectation of $p+1$, models with $C_p > p$ usually suggest that the data is not represented well by the model.

Pregibon (1979) presents a very interesting and natural generalization to Mallow's C_p statistic for the generalized linear model. Define

$$C_p^* = (D_{N,p} / \hat{\phi}) - N + 2(p+1), \tag{2.9.2}$$

where $\hat{\phi} = D_{N,k} / (N-k)$ presented in equation (2.6.14) and k is the full number of regressors. Notice that $C_p^* = C_p$ for normal responses with the identity link.

The value $D_{N,p} / \hat{\phi}$ given in equation (2.9.2) can be replaced by the χ^2 statistic given previously in equation (2.8.2). See equations (2.7.1) and (2.7.2). In Pregibon's (1979) development of C_p^*, he chooses to use $D_{N,p} / \hat{\phi}$ over χ^2 for nonnormal models to make a connection between C_p^* and Akaike's (1974) Information Criterion (AIC). Akaike developed

$$\begin{aligned} AIC_p &= -2\, l(\hat{\underline{\beta}}; \underline{y}) + 2(p+1) \\ &= D_{N,p} - 2\, l(\hat{\underline{\beta}}_{\max}; \underline{y}) + 2(p+1) \\ &= C_p^* - 2\, l(\hat{\underline{\beta}}_{\max}; \underline{y}) + N, \end{aligned} \tag{2.9.3}$$

when $\hat{\phi} = 1$. The statistic AIC_p will punish models with large numbers of explanatory variables in the same way C_p^* does.

2.10 CENTERING AND SCALING THE EXPLANATORY VARIABLES

Quite often it is convenient to look at standardized columns of the X matrix so that the variables are unitless (Myers (1986)). Centering and scaling often help the analysis in, for example, principal components regression to make sense. Generalized ridge regression also has an appeal to centering and scaling. Further, inversion problems can exist for $X'X$ when the explanatory variables have extremely different magnitudes. For sake of consistency, throughout this dissertation the explanatory variables will be presented or assumed as centered and scaled. The columns of the X matrix can be represented as

$$X = (1 \;\; x_1 \;\; x_2 \;\; \ldots \;\; x_p),$$

where x_i represents an independent variable. Redefine

$$X = (x_1^* \;\; x_2^* \;\; \ldots \;\; x_p^*), \tag{2.10.1}$$

where

$$x_i^* = SS_i^{-1/2}(x_i - \bar{X}_i 1)$$

$$\bar{X}_i = N^{-1}\sum_{j=1}^{N} x_{ij}$$

$$SS_i = \sum_{j=1}^{N} (x_{ij} - \bar{X}_i)^2.$$

X^* is $(p \times p)$ matrix and $R = X^{*\prime}X^*$ is in the usual correlation form.

Chapter III

LOGISTIC REGRESSION

3.1 INTRODUCTION

Up to this point, this dissertation has been quite general. That is, within the framework of the generalized linear model, the response variable can be from any distributional form in the exponential family. To broaden an already general scenario, the researcher is not required to use the natural link even though it is the function most often used to connect the mean response to the systematic linear component. Upon data collection, both the distributional form of the response and the link function must be determined in order to implement an iterative maximum likelihood estimation technique.

As a specific case of the GLM, Schaefer (1979 and 1986) considers alternatives to iterative reweighted maximum likelihood parameter estimation when the response is Bernoulli. Schaefer has developed both ridge and principal component techniques for logistic regression. Maximum likelihood is particularly to be avoided in the presence of an ill-conditioned information matrix. Schaefer's work is much of the motivation behind this dissertation. The author will expand on

Schaefer's principal component logistic regression (PCLR) parameter estimates. Specifically, an iterative PCLR technique will be developed as an alternative to Schaefer's one step adjustment to maximum likelihood. Both one step and iterative principal component estimators, as well as ridge estimators, will be extended to the GLM in upcoming chapters.

3.2 DEVELOPMENT OF LOGISTIC REGRESSION

It is not uncommon in research to obtain dichotomous data on a number of individuals. For example, each individual may be given or denied a car loan, may favor or oppose a political issue, may or may not acquire a disease. Usually this binary datum is recorded along with a set of the individual's characteristics; perhaps levels of blood glucose, antibodies, and urine protein are on a medical record. Since the outcome frequently depends on the individual's set of characteristics, logistic regression equations can be developed to model and predict the probability for a future individual's outcome when given his set of characteristics. The work here will specifically deal with continuous explanatory variables. Two important and desirable properties of the logistic regression model are good prediction of the probability and good estimates of regression coefficients. These properties may be unattainable while using maximum likelihood estimation.

The objective is to develop a method that will reduce or eliminate damage that a near singular information matrix poses to binary regressions, while maintaining accurate probability predictions. In addition, if theoretical equations are specified, then more trustworthy estimates can be given for regression coefficients, that is for the rate and direction of change in this probability when one of the characteristics is increased or decreased. This research is especially important in medical issues since inaccurate prediction may be catastrophic. In some instances, applying alternate estimation techniques in the logistic setting allows, for example, the construction of a reliable probability equation to predict whether a person actually has a disease when given a set of explanatory variables results in an ill-conditioned information matrix.

Consider a binomial random variable Y with parameters n and π . $f(y)$ is a member of the exponential family and has the natural parameter θ such that

$$f_Y(y;\theta) = \exp\{y\theta + c(\theta) + d(y)\} I_{\{0,1,2,\ldots n\}}(y), \tag{3.2.1}$$

where $\theta = \text{logit}(\pi) = \ln\left[\frac{\pi}{(1-\pi)}\right]$, $c(\theta) = -n\ \ln(1+e^{\theta})$, $d(y) = \ln\binom{n}{y}$, and $n \in Z^{+}$. Note that $q(\phi) = 1$. Thus y is a complete sufficient statistic with

$$\begin{aligned} E(Y) &= -c'(\theta) = n\pi \\ \text{Var}(Y) &= -c''(\theta) = n\pi(1-\pi). \end{aligned} \tag{3.2.2}$$

Given a sample of N independently distributed binomial random variables Y_i with parameters n_i and π_i respectively, the log-likelihood function becomes

$$\begin{aligned} l(\underline{\theta}; \underline{y}) &= \sum_{i=1}^{N} l(\theta_i, y_i) \\ &= \sum_{i=1}^{N} \{y_i\theta_i + c(\theta_i) + d(y_i)\}. \end{aligned} \tag{3.2.3}$$

Just as in the framework of the GLM, notice that there are as many parameters to estimate as there are observations; $l(\underline{\theta}, \underline{y})$ is overspecified. However, given a set of p covariates $\{X_1, X_2, \ldots, X_p\}$ for each Y_i one could model the parameter vector $\underline{\theta}$ by

$$\underline{\theta} = \text{logit}(\underline{\pi}) = X\underline{\beta}, \quad (p+1) << N, \tag{3.2.4}$$

where X is a $N \times (p+1)$ matrix of covariates including the constant term and logit links the systematic linear component $\underline{x}'\underline{\beta}$ to the mean response $n\pi$. Now the log-likelihood function can be written as

$$l(X\underline{\beta}; \underline{y}) = \sum_{i=1}^{N} \{y_i\underline{x}'_i\underline{\beta} + c(\underline{x}'_i\underline{\beta}) + d(y_i)\}, \tag{3.2.5}$$

where $\underline{x}'_i$ is an observation vector. Consider the maximum likelihood estimates of $\underline{\beta}$

$$0 = \frac{\partial}{\partial \underline{\beta}} \, l(X\underline{\beta}; \underline{y}) = \sum_{i=1}^{N} x_{ij}(y_i + c'(\underline{x}'_i\underline{\beta})) \qquad j = 0, 1, \ldots, p \tag{3.2.6}$$

where $-c'(\underline{x}'_i\hat{\underline{\beta}}) = n_i\hat{\pi}_i = \hat{y}_i$ and $\hat{\pi}_i = [\exp(\underline{x}'_i\hat{\underline{\beta}})] / [1 + \exp(\underline{x}'_i\hat{\underline{\beta}})]$. This leads to a set of "normal-like" equations which are nonlinear in $\hat{\underline{\beta}}$

$$X'(\underline{y} - \hat{\underline{y}}) = 0. \tag{3.2.7}$$

This set of equations does not have a closed form solution and iterative methods are usually employed to solve for the maximum likelihood estimates (MLE) $\hat{\underline{\beta}}$. Albert and Anderson (1984) discuss nonuniqueness and nonexistence of the logit MLE's for the coefficient vector. Although maximum likelihood estimation is available, if the data are grouped so that there are multiple observations at various levels of the covariates, then empirical weighted least squares can be used as a one step estimation procedure as $\hat{K}$ need not be obtained via iteration.

3.3 GROUPED DATA

One method of empirically solving for $\hat{\underline{\beta}}$ when $n_i > 1$ is by means of modeling z^*_i as a linear function of the parameters and employing weighted least squares where

$$\text{logit}(\pi^*_i) = z^*_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i \ . \tag{3.3.1}$$

The observed $z^*_i = \ln\left[\frac{y_i}{n_i - y_i}\right]$, with $E(\varepsilon_i) = 0$, and $\text{Var}(\varepsilon_i) \cong (n_i\pi_i(1 - \pi_i))^{-1} = \Gamma_{ii}$. Define the diagonal matrix

$$\begin{aligned} \Gamma &= E(\underline{\varepsilon}\underline{\varepsilon}') = \text{diag}\{\Gamma_{ii}\} \\ \hat{\Gamma} &= \text{diag}\{n_i [y_i(n_i - y_i)]^{-1}\}. \end{aligned} \tag{3.3.2}$$

Thus a one step weighted least squares estimate for the parameter vector is

$$\hat{\underline{\beta}}^* = (X'\hat{\Gamma}^{-1}X)^{-1}X'\hat{\Gamma}^{-1}\underline{z}^*. \tag{3.3.3}$$

Inferences on $\underline{\beta}$ are based on approximate normality of the error term.

Cox (1970) suggested an improvement to the estimation of parameters given in equation (3.3.3). To help the small sample properties of estimation without affecting the asymptotic results, use

$$\ln\left[\frac{y_i+\frac{1}{2}}{n_i-y_i+\frac{1}{2}}\right] \tag{3.3.4}$$

with estimated variances

$$\tilde{\Gamma}_{ii} = \frac{(n_i+1)(n_i+2)}{n_i(y_i+1)(n_i-y_i+1).}$$

In the ungrouped case, $n_i = 1$ for all i, notice that

$$\ln\left[\frac{y_i+\frac{1}{2}}{n_i-y_i+\frac{1}{2}}\right] = \begin{cases} \ln\ \frac{1}{3} \cong -1 & y_i = 0 \\ \ln\ 3 \cong 1 & y_i = 1. \end{cases} \tag{3.3.5}$$

In terms of inferences on $\underline{\beta}$, this is not a satisfactory method in the ungrouped ($n_i = 1$) case since normality of z^*_i is not a reasonable assumption due to the discrete nature of the logit function in equation (3.3.5). In this setting, another method of estimation is needed since there does not exist an initial estimate of Γ.

3.4 UNGROUPED DATA

In the case where $n_i = 1$ for all i, this is a regression setting with continuous covariates without replication. Walker and Duncan (1967) modeled this Bernoulli Y using

$$y_i = \pi(\underline{x}'_i\underline{\beta}) + \varepsilon_i \quad y_i = 0,\ 1 \ \text{ independent,} \tag{3.4.1}$$

where Y_i is the binary datum, $\pi(\underline{x}'_i\underline{\beta})$ is interpreted as the Bernoulli parameter $P(Y_i = 1 | \underline{x}'_i\underline{\beta})$ and $\varepsilon_i \sim (0,\ \pi_i(1-\pi_i))$. Notice that $\pi_i = \pi(\underline{x}'_i\underline{\beta}) = [\exp(\underline{x}'_i\underline{\beta})] / [1 + \exp(\underline{x}'_i\underline{\beta})]$ is nonlinear in the parameters and in the ungrouped setting the additive error term is assumed. $\underline{\beta}$ is an unknown $(p+1) \times 1$ coefficient vector and π_i is constrained to the unit interval. Since $e^z \geq 0$ for all $z \in R$, $0 \leq \pi_i \leq 1$. Also $\pi_i = F(\underline{x}'_i\underline{\beta})$ where $F(\cdot)$ is the cumulative distribution of the logistic family. $\pi_i = F(z_i) = F(\underline{x}'_i\underline{\beta})$. The rate and direction of change in the probability per unit change in x_{ij} can be estimated by using $\hat{\underline{\beta}}$ with

$$\frac{\partial \pi_i}{\partial x_{ij}} = \frac{\partial F}{\partial z_i}\ \frac{\partial z_i}{\partial x_{ij}} = f(z_i)\,\beta_j \quad j = 0,\ 1,\ 2,\ \ldots,\ p, \tag{3.4.2}$$

where $f(z_i)$ is the logistic density function evaluated at the scalar index $z_i = \underline{x}'_i\underline{\beta} \in R$. The standard logistic density closely resembles the t-distribution with seven degrees of freedom (Pindyck and Rubinfeld (1981)). It is convenient to think of the monotone increasing function π mapping the index $z_i \in R$ into the unit interval.

3.5 ITERATIVE GAUSS-NEWTON SOLUTIONS

In Chapter 2, the method of scoring was developed as a means of maximum likelihood parameter estimation. It is also instructive to view logistic regression from a nonlinear model of

Walker and Duncan (1967). An alternate method of iteration is the Gauss-Newton procedure using the Taylor series expansion up to the linear term. Iteration is continued until some specified degree of convergence is obtained. Denote t as the iteration step.

Consider the Taylor series expansion of a general function $h(x)$ about a constant b, then

$$h(x) = h(b) + h'(x)|_{x=b}(x-b) + \frac{h''(x)}{2!}|_{x=b}(x-b)^2 + \cdots \tag{3.5.1}$$

In the case where h is a function of several covariates, then

$$h(x_1, x_2, \ldots, x_p) \cong h(X, \underline{\beta}_0) + \sum_{j=0}^{p} \frac{\partial h}{\partial \beta_j}(\beta_j - \beta_{j,0}). \tag{3.5.2}$$

Hence a linear model can be formulated (Capps (1985))

$$y_i - F(X, \underline{\beta})|_{\underline{\beta}=\underline{\beta}_t} = \sum_{j=0}^{p} \frac{\partial F}{\partial \beta_j}|_{\underline{\beta}=\underline{\beta}_t}(\beta_j - \beta_{j,t}) + \varepsilon_i, \tag{3.5.3}$$

where this additive error term is more tenable. Equation (3.5.3) may be rewritten as

$$y_i - \hat{\pi}_{i,t-1} = \sum_{j=0}^{p} w_{ij} \gamma_j + \varepsilon_i, \tag{3.5.4}$$

where

$$w_{ij,t} = \frac{\partial \pi(\underline{x}_i'\underline{\beta})}{\partial \beta_j}|_{\underline{\beta}=\underline{\beta}_{t-1}} \qquad \begin{matrix} i = 1, 2, \ldots, N \\ j = 0, 1, \ldots, p, \end{matrix} \tag{3.5.5}$$

t is the iteration step and $W = (w_{ij})$ has dimension of $N \times (p+1)$, and $\hat{\underline{\pi}} = \underline{\pi}(X\hat{\underline{\beta}}) = \hat{\underline{y}}$. Denote

$$\gamma_j = (\beta_j - \beta_{j,t-1})$$

and thus $\hat{\underline{\gamma}}_t = (W'\hat{V}_{t-1}^{-1}W)^{-1}W'\hat{V}_{t-1}^{-1}(\underline{y} - \hat{\underline{y}}_{t-1})$.

Define the variance-covariance matrix of Y to be $V = \text{diag}\{\pi_i(1-\pi_i)\}$, $i = 1, 2, \ldots, N$. The development of V follows (Myers (1986)).

$$\begin{aligned}
\text{Var}(\varepsilon_i) &= E(\varepsilon_i^2) \quad \text{since} \quad E(\varepsilon_i) = 0 \\
&= E[y_i - \pi(\underline{x}'_i\underline{\beta})]^2 \\
&= P(y_i = 1)[1 - P(y_i = 1)]^2 + P(y_i = 0)[-P(y_i = 1)]^2 \\
&= \pi_i(1-\pi_i)^2 + (1-\pi_i)(-\pi_i)^2 \\
&= \pi_i(1-\pi_i) \\
&= v_i \ . \\
E(\underline{\varepsilon}\,\underline{\varepsilon}') &= \text{diag}\{v_i\} \\
&= V = \Gamma^{-1}.
\end{aligned}$$

Call $\hat{V} = \text{diag}\{\hat{\pi}_i(1-\hat{\pi}_i)\}$. Utilizing that $\gamma_j = \beta_j - \beta_{j,t-1}$, the weighted iterative maximum likelihood scheme leads to

$$\hat{\underline{\beta}}_t = \hat{\underline{\beta}}_{t-1} + (W'_{t-1}\hat{V}_{t-1}^{-1}W_{t-1})^{-1}W'_{t-1}\hat{V}_{t-1}^{-1}(\underline{y} - \hat{\underline{y}}_{t-1}), \tag{3.5.6}$$

Note $W = VX$ since

$$\begin{aligned}
\frac{\partial\pi(\underline{x}'_i\underline{\beta})}{\partial\beta_j} &= \begin{cases} e^{\underline{x}'_i\underline{\beta}}(1+e^{\underline{x}'_i\underline{\beta}})^{-2} & j = 0 \\ x_{ji}e^{\underline{x}'_i\underline{\beta}}(1+e^{\underline{x}'_i\underline{\beta}})^{-2} & j = 1, 2, \ldots, p \end{cases} \\
&= \begin{cases} \pi_i(1-\pi_i) & j = 0 \\ \pi_i(1-\pi_i)x_{ji} & j = 1, 2, \ldots, p \end{cases}
\end{aligned}$$

and $i = 1, 2, \ldots, N$. Thus equation (3.5.6) can be re-expressed as

$$\hat{\underline{\beta}}_t = \hat{\underline{\beta}}_{t-1} + (X'\hat{V}_{t-1}X)^{-1}X'(\underline{y} - \hat{\underline{y}}_{t-1}). \tag{3.5.7}$$

It is interesting to note in the case for logistic regression that the Gauss-Newton approach developed in equation (3.5.7) is completely consistent with the method of scoring in equation (2.4.7) when Bernoulli data is used in the GLM. Thus equation (3.5.7) is an iterative maximum likelihood

solution to the re-expressed model given in equation (3.4.1) which can be used for grouped data as well. Maximum likelihood estimation has the large sample properties of consistency and asymptotic normality of $\hat{\underline{\beta}}$ allowing conventional tests of significance.

3.6 PROPERTIES OF LOGISTIC REGRESSION

Schaefer (1979) points out that most of the theoretical work on the asymptotic properties of maximum likelihood estimators for independent, nonidentically distributed responses has already been done (Bradley and Gart (1962)). Bradley and Gart's work essentially require that the following two assumptions hold:

i) $|x_{ij}|$ is bounded for all i and j;

ii) $\lim_{N\to\infty} N^{-1}(X'VX) = Q$, for Q positive definite with finite determinant.

The first assumption is perfectly reasonable for regression data sets. In unconventional circumstances when an element of the X matrix takes an arbitrarily large value, then set $x_{ij} = K^*$ for $|x_{ij}| \geq K^*$, where K^* is a large constant. The second assumption given above is equivalent to requiring the distribution of the $\underline{x}$'s to have a finite second moment. Moreover, the second assumption implies that the elements of $X'VX$ are of order $O(N)$ which follows directly from the definition and (ii) above.

Once these two assumptions are satisfied, the groundwork is set for the following to hold:

i) $\hat{\underline{\beta}}$, the maximum likelihood estimated parameter vector, is consistent for $\underline{\beta}$, the true parameter vector.

ii) $(\hat{\underline{\beta}} - \underline{\beta})$ converges in distribution to a $(p+1)$ multivariate normal distribution with mean $\underline{0}$ and variance-covariance matrix Q^{-1}. Q is defined above.

Based on the above results from Schaefer (1979) and the development given in section 2.5, for large N, $\hat{\underline{\beta}}$ is asymptotically unbiased for $\underline{\beta}$ with variance-covariance matrix $(X'VX)^{-1}$. In practice V is usually unknown and is also estimated via maximum likelihood. In fact, $X'\hat{V}X$ is a consistent estimate for $X'VX$. Hence the following asymptotic results are commonly used in practice:

i) $E(\hat{\underline{\beta}} - \underline{\beta}) \cong 0$;

ii) $\text{Var}(\hat{\underline{\beta}}) \cong (X'\hat{V}X)^{-1}$;

iii) $MSE(\hat{\underline{\beta}}) \cong \text{tr}(X'\hat{V}X)^{-1} + \text{Bias}^2(\hat{\underline{\beta}}) = \sum_{i=0}^{p} \hat{\lambda}_i$,

where the $\hat{\lambda}_i$ are the eigenvalues of $X'\hat{V}X$.

Following from the above asymptotic results of consistency, efficiency and normality, the usual t-tests and confidence intervals can be applied. To test the significance of all or a subset of the regression coefficients, a χ^2-test is used rather than a F -test. See section 2.6. For example, suppose that the significance of the logit model is tested,

$$\begin{aligned} &H_o\colon\ \beta_0 \neq 0,\ \beta_1 = \ \ldots\ = \beta_p = 0 \\ &H_1\colon\ \text{not}\ \ H_o\,. \end{aligned} \tag{3.6.1}$$

Equation (3.6.1) suggests using

$$\begin{aligned} \lambda &= L(\hat{\underline{\beta}}_{\max}; \underline{y}) \,/\, L(\hat{\underline{\beta}}_o; \underline{y}) \\ &\sim \chi^2_p. \end{aligned}$$

In fact, the above test can be used in developing a measure of goodness-of-fit which is analogous to a R^2 measure. McFadden's R^2 (Likelihood Ratio Index) is given by

$$R^2_{LR} = 1 - \lambda^{-1}, \tag{3.6.2}$$

(see Pindyck and Rubinfeld (1981)). Notice that the statistic given in equation (3.6.2) is identically equal to zero when there is no increase in the likelihood function, given p additional regressors. However, R^2_{LR} increases toward one as the regressors explain the true model and hence deviates away from the constant model. Other measures of goodness-of-fit include Efron's R^2,

$$R^2_{EF} = 1 - \left[\sum_{i=1}^{N} (y_i - \hat{\pi}_i)^2 \Big/ \sum_{i=1}^{N} (y_i - \bar{y})^2\right] \tag{3.6.3}$$

(which is analogous to $1 - (SSE / SSTOT)$) and the square of the Pearson Product Moment Correlation coefficient,

$$\text{PPMC} = \left[\sum_{i=1}^{N} (y_i - \bar{y})\hat{\pi}_i\right]^2 \Big/ \left[(\sum_{i=1}^{N} (y_i - \bar{y})^2)(\sum_{i=1}^{N} (\hat{\pi}_i - \bar{\pi})^2)\right]. \tag{3.6.4}$$

It should be noted that values between .1 and .3 for R^2_{EF} or PPMC are not at all uncommon for a reasonably good fit (Capps (1985)). Perhaps the most widely used method of goodness-of-fit for the logit model is one of Proportion Correct Classification. That is if the estimated probability is greater than (less than) 1/2 and the first (second) alternative is selected, then the decision is correctly classified. Thus

$$\text{Proportion Correct} = \text{No. of Correctly Classified} / N. \tag{3.6.5}$$

The χ^2-statistic is quite common among computer packages as a goodness-of-fit measure in logistic regression. Define

$$\chi^2 = \sum_{i=1}^{N} (y_i - \hat{\pi}_i)^2 [\hat{\pi}_i (1 - \hat{\pi}_i)]^{-1/2}, \tag{3.6.6}$$

where y_i refers to the observed 0,1 responses.

Deviance is also customary as a model fitting statistic. Define

$$D = 2\{l(\hat{\underline{\theta}}; \underline{y}) - l(X\hat{\underline{\beta}}; \underline{y})\}, \tag{3.6.7}$$

where $l(\hat{\underline{\theta}}; \underline{y})$ refers to the maximum likelihood of the log-likelihood function when fitting each data observation exactly. In this case, $\hat{\theta}_i = \text{logit}(\hat{\pi}_i)$ and is undefined for 0,1 responses. However, D is defined for all values of y_i even though $\hat{\theta}_i$ may not be. Using l'Hospital's rule

$$D = \sum_{i=1}^{N} d_i^2, \tag{3.6.8}$$

where

$$d_i^2 = -2\ln(1 - \hat{\pi}_i) \quad \text{for } y = 0$$
$$d_i^2 = -2\ln(\hat{\pi}_i) \quad \text{for } y = 1.$$

Both χ^2 and D are excellent goodness-of-fit measures, and asymptotic arguments suggest that they both have a limiting χ^2_{N-p-1} distribution. There are still other possibilities for measuring goodness-of-fit, such as unweighted sum of squared residuals and Akaike's Information Criterion (AIC).

3.7 WEIGHTED COLLINEARITY

Collinearity among the $\underline{w}_j$ in equation (3.5.4) can give estimates of the coefficients of the $\underline{x}_j$ which are unstable and sensitive to small perturbations to the data. Not only does collinearity give

imprecise estimates of γ_j's, but may give estimates which have the wrong sign. Consequently, the problem of identifying the effects of the explanatory variables is now compounded by the possibility of a damaged rate of convergence in this iterative procedure.

Notice equation (3.5.4) can be rewritten

$$\underline{z} = W\underline{\gamma} + \underline{\varepsilon}, \tag{3.7.1}$$

where $E(\underline{\varepsilon}) = \underline{0}$ and $E(\underline{\varepsilon}\,\underline{\varepsilon}') = V$. Consider the transformation to (3.7.1) of the form (Burdick (1987))

$$\begin{aligned} V^{-1/2}\underline{z} &= V^{-1/2}W\underline{\gamma} + V^{-1/2}\underline{\varepsilon} \\ \text{or} \quad \underline{z}^* &= W^*\underline{\gamma} + \underline{\delta}^* \end{aligned} \tag{3.7.2}$$

and where now $E(\underline{\delta}^*) = \underline{0}$ and $E(\underline{\delta}^*\underline{\delta}^{*\prime}) = I$. Hence the homogeneous error covariance matrix is the identity. Thus a "correlation" form of the $\underline{w}_j$ can be expressed by

$$W^{*\prime}W^* = W'V^{-1}W = X'VX. \tag{3.7.4}$$

Let M be an orthogonal matrix such that $MM' = M'M = I = MM^{-1}$ and

$$M'(X'\Gamma^{-1}X)M = M'(X'VX)M = \Lambda. \tag{3.7.4}$$

M is a set of eigenvectors of $X'VX$ and Λ is a diagonal matrix of associated eigenvalues, λ_i for $i = 0,\ 1,\ 2,\ \ldots,\ p$. In up coming chapters, M will represent an orthogonal matrix which gives a spectral decomposition of the information matrix. Thus if the positive definite matrix $X'VX$ is near singular then $|X'VX| = \prod_{i=0}^{p}\lambda_i \cong 0$ and $\lambda_i \cong 0$ for some i . Details regarding ill-conditioning of $X'VX$ are developed for the GLM in Chapter 4. Notice that in either estimation technique (empirical weighted least squares or maximum likelihood), the collinearity among the X variables may or may not be relevant, rather, the collinearity of the weighted X variables is the issue.

3.8 DAMAGING CONSEQUENCES OF ILL-CONDITIONED INFORMATION

Ill-conditioning of $X'VX$ leads to the demise of many desirable aspects of the logistic regression. Perhaps the most obvious damage done by small eigenvalues of $X'VX$ is the inflation of the trace of $(X'VX)^{-1}$.

$$\begin{aligned}\sum_{i=0}^{p}\mathrm{Var}(\hat{\beta}_i) &\cong \mathrm{tr}(X'VX)^{-1} \\ &= \mathrm{tr}[MM'(X'VX)^{-1}] \\ &= \mathrm{tr}[\Lambda^{-1}] \\ &= \sum_{i=0}^{p}\frac{1}{\lambda_i}.\end{aligned}$$

Thus a near zero λ_i severely increases the sum of the variances of the estimated coefficients. Further, it can be shown that $\lambda_{\min} \to 0$ implies $\mathrm{Var}(\hat{\beta}_j) \to \infty$ for some j. Near singularity of $X'VX$ does indeed inflate at least one variance of the estimated parameters. Consequently, interpretations of the meaning of the magnitude and sign of a coefficient must be made with extreme caution.

Secondly, another variance which may be inflated due to near singularity of $X'VX$ is that of $\mathrm{Var}(\hat{y}(x_0))$. Given a new observation vector, x'_0, the variance of the predicted probability of a success will be inflated if the vector x'_0, is outside the mainstream of collinearity among the $V^{1/2}X$ data. The variance can be expressed as

$$\mathrm{Var}\ (\hat{y}(x_0)) \cong [\pi_{1,0}(1-\pi_{1,0})]^2 \sum_{i=0}^{p}\frac{z_{i,0}^2}{\lambda_i}, \tag{3.8.1}$$

where the z_i are the coordinates of the transformed orthogonal principal axes, $Z = XM$. Figure 1 presents orthogonal principal component axes for data which are collinear in a weighted sense. An argument for equation (3.8.1) follows.

$$
\begin{aligned}
\mathrm{Var}(\hat{\underline{\beta}}) &\cong (X'VX)^{-1} \qquad \text{(asymptotically)} \\
\mathrm{Var}(\underline{x}'_0\hat{\underline{\beta}}) &\cong \underline{x}'_0(X'VX)^{-1}\underline{x}_0 \\
&= \underline{x}'_0MM'(X'VX)^{-1}MM'\underline{x}_0 \\
&= \underline{z}'_0\Lambda^{-1}\underline{z}_0 \\
&= \sum_{i=0}^{p} \frac{z_{i,0}^2}{\lambda_i}.
\end{aligned}
$$

Let

$$
\begin{aligned}
\hat{\eta} &= \underline{x}'_0\hat{\underline{\beta}} \\
h(\hat{\eta}) &= \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}}.
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathrm{Var}[h(\hat{\eta})] &\cong \mathrm{Var}(\hat{\eta})[\,h'(\hat{\eta})]^2|_{\eta} \\
&= \left[\sum_{i=0}^{p} \frac{z_{i,0}^2}{\lambda_i}\right][\pi_{i,0}(1-\pi_{i,0})]^2.
\end{aligned}
$$

Thus

$$
\hat{\mathrm{Var}}[h(\hat{\eta})] \cong \left[\sum_{i=0}^{p} \frac{z_{i,0}^2}{\lambda_i}\right][\hat{y}_{i,0}(1-\hat{y}_{i,0})]^2.
$$

Thus if $z_{i,0}^2$ is relatively large when the corresponding value of λ_i is small, then the $\mathrm{Var}[h(\hat{\eta})]$ is inflated. The data point represented as '*' in Figure 1 demonstrates a region in the weighted X space where prediction can be poor due to the variance argument given in equation (3.8.1). Notice

$$
0 \le \mathrm{Var}(\hat{y}(\underline{x}_0)) \le .0625\sum_{i=0}^{p} \frac{z_{i,0}^2}{\lambda_i}.
$$

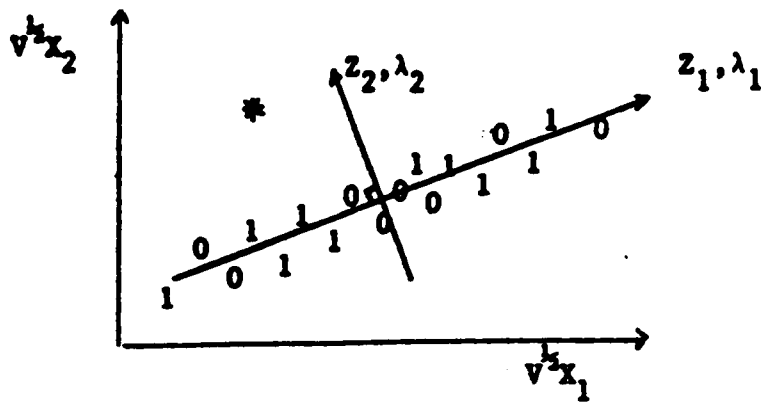

Figure 1. POOR PREDICTION WITH WEIGHTED COLLINEARITY

A third damaging consequence of an ill-conditioned $X'VX$ matrix is that the power of certain tests may be damaged due to a deflation of the test statistic. Consider the argument below.

$$\begin{aligned} H_0&: \underline{\beta} = \underline{\beta}_C \\ H_1&: \underline{\beta} = \underline{\beta}_F \ , \end{aligned} \tag{3.8.2}$$

where C and F denote current and full models respectively. Haberman (1978) and Jennings (1986) give conditions for the test. A summary of the conditions necessary for the test is that as $N \to \infty$ then $\lim_{N\to\infty} N^{-1}(X'_N X_N)$ exists and is positive definite, where X_N is the matrix under H_1. Define $l(\cdot) = l(X\underline{\beta}; \underline{y}) = \sum_{i=1}^{N} [y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)]$. The test statistic of the above test can be shown to be (see section 2.6)

$$\begin{aligned} \chi^2 &= (\hat{\underline{\beta}}_F - \hat{\underline{\beta}}_C)' X' V X (\hat{\underline{\beta}}_F - \hat{\underline{\beta}}_C) \\ &= (\hat{\underline{\alpha}}_F - \hat{\underline{\alpha}}_C)' \Lambda (\hat{\underline{\alpha}}_F - \hat{\underline{\alpha}}_C) \\ &= \sum_{i=0}^{p} (\hat{\alpha}_{i,F} - \hat{\alpha}_{i,C})^2 \lambda_i . \end{aligned} \tag{3.8.3}$$

Notice as $\lambda_i \to 0$ for some i then χ^2 decreases thus damaging the power of the test. The i^{th} dimension does not contribute to the test statistic.

Moreover, collinearity among the $\underline{w}'_i s$ could be the direct cause of parameter estimates failing to converge during the iterative process. Schaefer (1984) notes

$$\begin{aligned} E(\hat{\underline{\beta}}'\hat{\underline{\beta}}) &= \underline{\beta}'\underline{\beta} + \text{tr}[\text{Var}(\hat{\underline{\beta}})] + \text{Squared Bias}(\hat{\underline{\beta}}) \\ &\geq \underline{\beta}'\underline{\beta} + \sum_{j=0}^{p} (\lambda_j)^{-1} \end{aligned}$$

and hence if the columns of $V^{1/2}X$ are collinear, the maximum likelihood estimate vector will be too long on the average. Also note that

$$\begin{aligned} V &= \text{diag}\{\pi_i(1-\pi_i)\} \\ \hat{V} &= \text{diag}\{\hat{y}_i(1-\hat{y}_i)\}, \end{aligned}$$

where $\hat{y}(x_0) = [\exp(x'_0\hat{\beta})] / [1 + \exp(x'_0\hat{\beta})]$. However $|x'_0\hat{\beta}|$ is likely to be quite large in magnitude; with the presence of collinearity resulting in either $\exp(x'_0\hat{\beta}) \to 0$, or ∞ and hence giving $\hat{V}_{\mu} \cong \emptyset$ or perhaps diagonal elements blowing up at the first step if limits are not imposed during exponentiation. Next some alternate estimation techniques will be suggested which will shrink $|x'_0\hat{\beta}|$.

3.9 INTRODUCTION TO PRINCIPAL COMPONENT REGRESSION

Certainly in standard multiple least squares regression, multicollinearity among the explanatory variables poses difficulty in parameter estimation even though Gauss-Markov properties of minimum variance among unbiased estimates hold. Various options have been proposed to overcome these problems. Variable deletion or subset regression is one option discussed in Chapter 5. Biased estimation techniques are also procedures to reduce the ill effects of collinearity. One such biased method is a ridge procedure developed by Hoerl and Kennard (1970a). Ridge estimation will be discussed for standard multiple regression, as well as for the GLM. The beauty of PC regression is that the X matrix of explanatory variables is transformed to a set of uncorrelated principal components. Hence collinearities are in some sense eliminated. In fact, if all the PC's are used in the regression problem, then the model is equivalent to the one obtained using least squares. However, the problems associated with multicollinearity have not faded into thin air. PC regression simply redistributes the large variances associated with the estimated coefficients. In situations when some of the principal components are deleted, the result is that the computed parameter estimates are biased yet at the same time have associated variance which can be greatly reduced. PC regression can effectively remove the ill effects of collinearity.

Jolliffe (1986) provides excellent coverage of principal component analysis, including principal component regression. Consider the standard multiple regression model,

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \tag{3.9.1}$$

where $\underline{y}$ is a $N \times 1$ vector of independent responses, X is a $N \times p$ matrix of explanatory variables which will be augmented by a constant vector of ones later for logistic regression, $\underline{\beta}$ is a $p \times 1$ unknown parameter vector, and $\underline{\varepsilon}$ is vector of independent random errors with mean $\underline{0}$ and common variance σ^2. For convenience, as well as consistency with the literature, let $X'X$ be in correlation form and X without a constant entry. Let M^* be the matrix such that its columns are the eigenvectors of $X'X$. The principal components are defined as

$$Z = XM^*, \tag{3.9.2}$$

where Z_{ij} is the value of the j^{th} PC on the i^{th} observation.

Jolliffe uses the fact that M^* is orthogonal and hence

$$\begin{aligned} \underline{y} &= X\underline{\beta} + \underline{\varepsilon} \\ &= XM^*M^{*\prime}\underline{\beta} \\ &= Z\underline{\alpha}, \end{aligned} \tag{3.9.3}$$

which replaces the explanatory variables by the PC's. If r of the PC's are deleted leaving $s = p - r$ components in the model, then the following notation is used

$$\underline{y} = Z_s\underline{\alpha}_s + \underline{\varepsilon}_s.$$

In fact, if all the components are kept in the model, then finding a least squares estimate for α is equivalent to finding an estimate for β. That is

$$\hat{\underline{\beta}} = M^*\hat{\underline{\alpha}}. \tag{3.9.4}$$

PC regression can give insight to the contribution of each explanatory variable even when collinearity is not present. However, the advantages of PC regression are most apparent with multicollinerity in the data. More stable estimates can be found for $\underline{\beta}$ in many cases when PC's

associated with small eigenvalues are deleted. An illustration of this follows from letting μ^2 be the eigenvalues of $X'X$. Define $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_p$. Notice that

$$\begin{aligned}\hat{\underline{\beta}} &= M^*(Z'Z)^{-1}Z'\underline{y} \\ &= M^*D^{-2}M^{*\prime}X'\underline{y} \\ &= \sum_{i=1}^{p} \mu_i^{-2}\underline{m}_i^*\underline{m}_i^{*\prime}X'\underline{y},\end{aligned} \tag{3.9.5}$$

where $D^2 = \text{diag}\{\mu_i^2\}$.

To give an understanding of how multicollinearities produce large variances in the estimates of $\hat{\beta}_i$, consider the variance-covariance matrix of $\hat{\underline{\beta}}$ (again disregarding the constant),

$$\begin{aligned}\sigma^2(X'X)^{-1} &= \sigma^2 M^*D^{-2}M^{*\prime} \\ &= \sigma^2\sum_{i=1}^{p} \mu_i^{-2}\underline{m}_i^*\underline{m}_i^{*\prime}.\end{aligned} \tag{3.9.6}$$

Hence any explanatory variable which has a large coefficient in any of the PCs associated with small eigenvalues has a large variance in that coefficient.

Naturally, a way to reduce the ill effects of multicollinearities would be to use the following PC estimator,

$$\underline{b}_S^{PC} = \sum_{i=r+1}^{p} \mu_i^{-2}\underline{m}_i^*\underline{m}_i^{*\prime}X'\underline{y}, \tag{3.9.7}$$

where the deleted μ_i are the very small ones. Further discussion of the number to delete will be presented later. Notice also that the

$$\begin{aligned} \mathrm{Var}(\underline{b}_s^{pc}) &= \sigma^2 \sum_{i=r+1}^{p} \mu_i^{-2} \underline{m}_i^* \underline{m}_i^{*\prime} \\ &= M_s^* D_s^{-1} M_s^{*\prime} \\ &= \mathrm{Var}(\hat{\beta}) - M_r^* D_r^{-1} M_r^{*\prime}, \end{aligned} \tag{3.9.8}$$

which can be significantly reduced when components associated with small eigenvalues are deleted.

Also consider the variance of a predicted response, say $\mathrm{Var}(\hat{y}(\underline{x}_o))$.

$$\mathrm{Var}(\hat{y}(\underline{x}_o)) / \sigma^2 = \underline{x}'_o (X'X)^{-1} \underline{x}_o,$$

for the standard multiple regression model. Equivalently,

$$\begin{aligned} \mathrm{Var}(\hat{y}(\underline{x}_o)) / \sigma^2 &= \underline{x}'_o (X'X)^{-1} \underline{x}_o \\ &= \underline{z}'_o D^{-2} \underline{z}_o \\ &= \sum_{i=1}^{p} z_{o,i}^2 \mu_i^{-2}. \end{aligned} \tag{3.9.9}$$

Notice that a large coordinate value of a principal component which is associated with a small eigenvalue can yield an inflated variance in equation (3.9.9). However for a subset of principal components, prediction variance can greatly be reduced. That is

$$\mathrm{Var}^{pc}(\hat{y}(\underline{x}_o)) / \sigma^2 = \sum_{r+1}^{p} z_{o,i}^2 \mu_i^{-2}. \tag{3.9.10}$$

Principal component estimation is a viable option for reducing prediction variance for new observations outside the mainstream of the original data points.

The bias associated with the principal component estimator can be quantified as follows

$$
\begin{aligned}
E(\underline{b}_S^{PC}) &= E(M_s^* \hat{\underline{\alpha}}_s) \\
&= E(M_s^* M_s^{*\prime} \hat{\underline{\beta}}) \\
&= \underline{\beta} - M_r \underline{\alpha}_r .
\end{aligned} \tag{3.9.11}
$$

Note that $I = M_s^* M_s^{*\prime} + M_r^* M_r^{*\prime}$ and if $\underline{\alpha}_r \cong \underline{0}$, then the bias is minimal. In fact the decrease in variance of the estimated coefficients can certainly outweigh the induced bias. Some suggestions as to the number and choice of principal components to delete are given in section 3.15 and will be developed further in section 5.5 for PCA in the GLM.

3.10 PRINCIPAL COMPONENT LOGISTIC REGRESSION (PCLR)

When using maximum likelihood techniques, principal component logistic regression (PCLR) introduces an additional bias in estimating the already biased coefficient vector. However, if PCLR is used successfully then some of the damaging consequences of an ill conditioned $X'VX$ matrix can be eliminated with only minimal additional bias. As mentioned, the variance of predicted probabilities for data outside the mainstream of collinearity can be reduced along with variance reductions in the estimated coefficients and greater power in certain tests.

Consider a data matrix X which has been centered and scaled or by design has variables of the same units. Details for such centering and scaling were given in section 2.10. Further, augment X to a vector of ones associated with the constant term. M is the orthogonal matrix such that it yields the spectral decomposition of $X'\Gamma^{-1}X = X'VX$. PCLR does not utilize the spectral decomposition the correlation matrix of the correlation matrix. The concern of PCLR is that the X matrix is composed of a set of p independent variables having the same scale to give some interpretation to linear combinations of variables. If, by design, the columns of the X matrix are originally the same units, then centering and scaling X may not be necessary to allow a more natural interpretation of the results.

3.11 SCHAEFER'S PCLR FOR UNGROUPED DATA

Using an approach much like that of Webster, Gunst and Mason (1974), Schaefer (1986) has developed a principal component logistic procedure. Define

$$X'X = \sum_{i=0}^{p} \lambda_i^* \underline{m}_i^* \underline{m}_i^{*\prime}$$
$$\text{and} \quad X'\hat{V}_t X = \sum_{i=0}^{p} \lambda_{it} \underline{m}_{it} \underline{m}'_{it}, \tag{3.11.1}$$

where λ_i and $\underline{m}$ denote the ordered eigenvalues and eigenvectors respectively of $X'\hat{V}_t X$ and λ_i^* and $\underline{m}_i^*$ the ordered eigenvalues and eigenvectors respectively of $X'X$. Again t denotes the iteration step and $\hat{V}_t$ is a maximum likelihood estimate.

Starting with the least squares estimator, $\hat{\underline{\beta}}_0$, Schaefer defines the logistic estimator as

$$\hat{\underline{\beta}} = \hat{\underline{\beta}}_0 + \sum_{t=0}^{L} (X'\hat{V}_t X)^{-1} X'(\underline{y} - \hat{\underline{\pi}}_t), \tag{3.11.2}$$

where L is the iteration of convergence. This leads to the principal component estimator (assuming a single collinearity)

$$\hat{\underline{\beta}}_{pc} = \sum_{i=1}^{p} [(\lambda_i^*)^{-1} \underline{m}_i^* \underline{m}_i^{*\prime} X'\underline{y} + \sum_{t=0}^{L} (\hat{\lambda}_{it})^{-1} \underline{m}_{it} \underline{m}'_{it} X'(\underline{y} - \hat{\underline{\pi}}_t)]. \tag{3.11.3}$$

Notice that the $\sum_{i=1}^{p}$ denotes the sum over $(p+1)-1$ components. $\hat{\underline{\pi}}_0$ is the probability of a "success" given the starting values, $\underline{\beta}_0$.

To simplify notation, Schaefer defines

$$(X'V_tX)^+ = \sum_{j=1}^{p} \lambda_{jt}^{-1} m_{jt} m'_{jt}$$
$$\text{and } (X'X)^+ = \sum_{j=1}^{p} (\lambda_j^*)^{-1} m_j^* m_j^{*\prime}. \tag{3.11.4}$$

Therefore, Schaefer's PC estimator can be written

$$\hat{\underline{\beta}}_{pc} = (X'X)^+ X'\underline{y} + \sum_t (X'\hat{V}_tX)^+ X'(\underline{y} - \hat{\underline{\pi}}_t). \tag{3.11.5}$$

Schaefer notes that $(X'\hat{V}_tX) \cong (X'\hat{V}_{ML}X)$ and $(X'\hat{V}_tX)^+ \cong (X'\hat{V}_{ML}X)^+$ since $\hat{V}_{ML}$ is a function of predicted data points which are not severely affected by ill-conditioning. Thus, Schaefer gives the one step estimate

$$\hat{\underline{\beta}}^*_{pc} = (X'\hat{V}_{ML}X)^+ (X'\hat{V}_{ML}X)\hat{\underline{\beta}}_{ML}. \tag{3.11.6}$$

The justification for $\hat{\underline{\beta}}_{pc}$ follows from Schaefer (1986). $\hat{V}_{ML}$ is a maximum likelihood estimate.

$$\begin{aligned}
\hat{\underline{\beta}}_{pc} &= (X'\hat{V}X)^+ (X'\hat{V}X)\hat{\underline{\beta}}_{ML} \\
&= (X'\hat{V}X)^+ (X'\hat{V}X)(X'X)^{-1}X'\underline{y} + \sum_{t=0}^{L} (X'\hat{V}X)^+ (X'\hat{V}X)(X'\hat{V}_tX)^{-1}X'(\underline{y} - \hat{\underline{\pi}}_t) \\
&\cong (X'\hat{V}X)^+ X'\underline{y} + \sum_{t=0}^{L} (X'\hat{V}X)^+ X'(\underline{y} - \hat{\underline{\pi}}_t) \\
&\cong (X'X)^+ X'\underline{y} + \sum_{t=0}^{L} (X'\hat{V}X)^+ X'(\underline{y} - \hat{\underline{\pi}}_t).
\end{aligned}$$

Let C^* be a constant. Note that Schaefer approximates $X'\hat{V}X$ with $C^*X'X$ and $(X'\hat{V}X)^+$ with $(C^*)^{-1}(X'X)^+$. If such approximations are reasonable, then the one step principal component logistic estimator has nice properties that would not require drastic changes to existing softwares.

3.12 AN ITERATIVE PCLR FOR UNGROUPED DATA

Recall the maximum likelihood iterative method in section 3.5, equation (3.5.7)

$$\hat{\underline{\beta}}_t = \hat{\underline{\beta}}_{t-1} + (X'\hat{V}_{t-1}X)^{-1}X'(\underline{y} - \hat{\underline{y}}_{t-1}).$$

Since V is a function of the unknown parameter vector $\underline{\beta}$, $\hat{V}_{t-1}$ is used, where

$$\begin{aligned}\hat{V}_{t-1} &= \text{diag}\{\pi(\underline{x}'_i\hat{\underline{\beta}}_{t-1})(1 - \pi(\underline{x}'_i\hat{\underline{\beta}}_{t-1}))\} \\ &= \text{diag}[\hat{\pi}_i(1 - \hat{\pi}_i)]_{t-1} \\ &= \text{diag}[\hat{y}_i(1 - \hat{y}_i)]_{t-1}.\end{aligned} \tag{3.12.1}$$

Again t denotes the iteration step.

Even if collinearity is severe among the $\underline{w}_j$, as mentioned, prediction is fairly good for the locations of the original data points. Despite the good estimation of V using maximum likelihood, perhaps $\hat{V}_{ML}$ can be fined tuned via principal component estimation. Using the fact that $M'X'VXM = \Lambda$, rewrite

$$\begin{aligned}&\text{logit}(\pi_i) = \underline{x}'_i\underline{\beta} = \underline{z}'_i\underline{\alpha} \\ &\text{and} \quad \pi_i = e^{\underline{x}'_i\underline{\beta}}(1 + e^{\underline{x}'_i\underline{\beta}})^{-1} \\ &\qquad\quad = e^{\underline{z}'_i\underline{\alpha}}(1 + e^{\underline{z}'_i\underline{\alpha}})^{-1},\end{aligned} \tag{3.12.2}$$

where $\underline{z}'_i = \underline{x}'_iM$ and $\underline{\alpha} = M'\underline{\beta}$. The $\hat{\underline{\alpha}}$ are the iterative reweighted least squares estimates with diagonal variance-covariance matrix Λ^{-1}. Hence the transformed variables, $\underline{z}'_i = \underline{x}'_iM$, or principal components (PC's) are orthogonal and uncorrelated. The total variance of the coefficients have been redistributed in such a way that a small eigenvalue of $X'\hat{V}X$ will flag a large variance for some $\hat{\alpha}_j$. Thus an elimination of at least one principal component $\underline{z}_j$ associated with $\lambda_{\min}$ could reduce the variability considerably in the model and perhaps repair some of the damage to various prop-

erties of the regression due to an ill-conditioned $X'\hat{V}X$. Thus the principal component iterative equation now becomes

$$\begin{aligned}\hat{\underline{\alpha}} &= \hat{\underline{\alpha}}_{t-1} + (Z'\hat{V}_{t-1}Z)^{-1}Z'(\underline{y} - \hat{\underline{y}}_{t-1}) \\ &= \hat{\underline{\alpha}}_{t-1} + \hat{\Lambda}_{t-1}^{-1}M'X'(\underline{y} - \hat{\underline{y}}_{t-1}),\end{aligned} \tag{3.12.3}$$

where t denotes the iteration step. Thus a natural iterative principal component scheme becomes

$$\hat{\underline{\alpha}}_{t,s}^{pc} = \hat{\underline{\alpha}}_{t-1,s}^{pc} + \hat{\Lambda}_{t-1,s}^{-1}M'_sX'(\underline{y} - \hat{\underline{y}}_{t-1}), \tag{3.12.4}$$

where $\hat{\Lambda}$ and M must be re-iterated at each step since $\hat{V}$ is changing and s denotes the number of principal components kept, $s = p + 1 - r$. The updating of the diagonal matrix of weights is for principal component logistic regression since $0 \leq k_{ii}^{-1} \leq .25$. However, empirical results suggest using the spectral decomposition of the fixed maximum likelihood estimate of the information in the general case since $0 \leq k_{ii}^{-1} \leq \infty$.

Typically, principal component analysis is used as a device to effectively reduce the dimensionality of the logistic regression. Since the principal components are artificial variables and often are difficult to interpret, the model will ultimately be converted back to one using the original variables. Suppose by choice, the principal component model is reduced by r dimensions. Thus the reduction in dimensionality is equivalent to the elimination of r eigenvectors of $X'\hat{V}X$ or setting r of the $\alpha's$ equal to zero using some rule. The transformation back to the original variables follows

$$\underline{b}_s^{pc} = M_s\hat{\underline{\alpha}}^{pc}. \tag{3.12.5}$$

Notice that this PCLR procedure differs from Schaefer's by convergence required in the $\hat{\underline{\alpha}}$ rather than the $\hat{\underline{\beta}}$. From empirical results, there do exist experimental situations when maximum likelihood estimates do not converge (thus Schaefer's PC approach does not converge), whereas the iterative PC approach does converge due to the reduction in dimensionality.

3.13 EMPIRICAL PCLR FOR GROUPED DATA

Recall the model in equation (3.3.1)

$$\begin{aligned} z^* &= X\underline{\beta} + \underline{\varepsilon} \\ &= XMM'\underline{\beta} + \underline{\varepsilon} \\ &= Z\underline{\alpha} + \underline{\varepsilon}. \end{aligned} \tag{3.13.1}$$

$\underline{\alpha}$ can be estimated by

$$\begin{aligned} \hat{\underline{\alpha}} &= (Z'\hat{\Gamma}^{-1}Z)^{-1}Z'\hat{\Gamma}^{-1}z^* \\ &= (M'X'\hat{\Gamma}^{-1}XM)^{-1}M'X'\hat{\Gamma}^{-1}z^* \\ &= \hat{\Lambda}^{-1}M'X'\hat{\Gamma}^{-1}z^*. \end{aligned} \tag{3.13.2}$$

Recall that Γ is estimated by

$$\hat{\Gamma} = \text{diag}\left\{ \frac{n_i}{y_i(n_i - y_i)} \right\}$$

and M are the eigenvectors of $X'\hat{\Gamma}^{-1}X$. The $\hat{\underline{\alpha}}$ are weighted least squares estimates with variance-covariance matrix Λ^{-1}.

Let $p + 1 = s + r = \dim(X'\hat{\Gamma}^{-1}X)$. Note the deletion of r principal components does not imply the deletion of any original regression variables and this is shown by

$$\underline{b}^{pc} = M_s\hat{\underline{\alpha}}_s, \tag{3.13.3}$$

where M_s is $(p + 1) \times s$ and $\hat{\alpha}_s$ is $s \times 1$. Thus

$$\hat{y}^{pc}(x_0) = (1 + \exp(-x'_0\underline{b}^{pc}))^{-1}. \tag{3.13.4}$$

3.14 PAYOFFS OF PCLR WITH ILL-CONDITIONED INFORMATION

The asymptotic variance-covariance matrix of the maximum likelihood estimates of $\underline{\beta}$ can be shown to be (Cox (1970))

$$\text{Var}(\hat{\underline{\beta}}) \cong (X'VX)^{-1}$$
$$\text{and} \quad \hat{\text{Var}}(\hat{\underline{\beta}}) \cong (X'\hat{V}X)^{-1}.$$

To quantify the magnitude of the decrease in $\text{Var}(\underline{b}^{pc})$, consider

$$\begin{aligned} \text{Var}(\hat{\underline{\beta}}) &\cong M\Lambda^{-1}M' \\ &= M_s\Lambda_s^{-1}M'_s + M_r\Lambda_r^{-1}M'_r. \end{aligned} \tag{3.14.1}$$

Thus

$$\text{Var}(\underline{b}^{pc}) \cong M_s\Lambda_s^{-1}M'_s. \tag{3.14.2}$$

Recall that the r eliminated principal components were the ones most likely to be associated with small eigenvalues of $X'VX$. Therefore, equation (3.14.2) illustrates that a considerable amount of coefficient variance can be eliminated using PCLR. The induced bias is quantified by a similar expression to that of equation (3.9.11).

3.15 ELIMINATING PRINCIPAL COMPONENTS

Certainly a difficulty with PCLR is determining how many principal components need to be eliminated, if any at all. Consider Figure 2. Notice that $V^{1/2}X_1$ and $V^{1/2}X_2$ are highly correlated and λ_2 is likely to be quite small. However $\hat{\alpha}_2$ is likely to be quite significant since the slope is carried in the Z_2 direction. Hence one would expect

$$t_1^* = \hat{\alpha}_2\sqrt{\hat{\lambda}_2} > 2,$$

where $\hat{\alpha}_2$ is essentially the discriminate function in this case. Hence, PCLR would not be appropriate.

As another example, consider $V^{1/2}X_1$ and $V^{1/2}X_2$ in Figure 3. In this setting, PCLR may be more appropriate since $\hat{\alpha}_2$ and $\hat{\lambda}_2$ are both relatively small, thus making

$$t_2^* = \hat{\alpha}_2\sqrt{\hat{\lambda}_2} < 2.$$

Hence it is now more reasonable to delete the principal component Z_2 which contains little information in this regression. When several variables are in the analysis then obviously the problem becomes more complex.

As another suggestion to determine the number of principal components to delete, a graph of $\sum_i \text{Var}(b^{pc})$ or $\sum_i \text{Var}(\hat{y}^{pc})$ vs. number of PC's deleted. The order of deletion of the principal components (PC's) can be done by the researchers choice. Some common rules are deleting "small" eigenvalues of $X'VX$ or by a stepwise procedure using a t -statistic, $t_i^* = \hat{\alpha}_i\sqrt{\hat{\lambda}}$. Developments of hypothesis testing and deletion of principal components for the GLM are given in section 5.5.

3.16 INTRODUCTION TO RIDGE REGRESSION

In standard multiple regression, as mentioned, other biased estimation techniques exist as an alternative to principal component regression in the quest to accurately estimate the true parameter vector, β. Quite often ridge regression is used as a plausible alternative to variable deletion or principal component regression. It has been repeatedly noted that the variance of the coefficients swell when collinear explanatory variables are used in their estimation. Moreover,

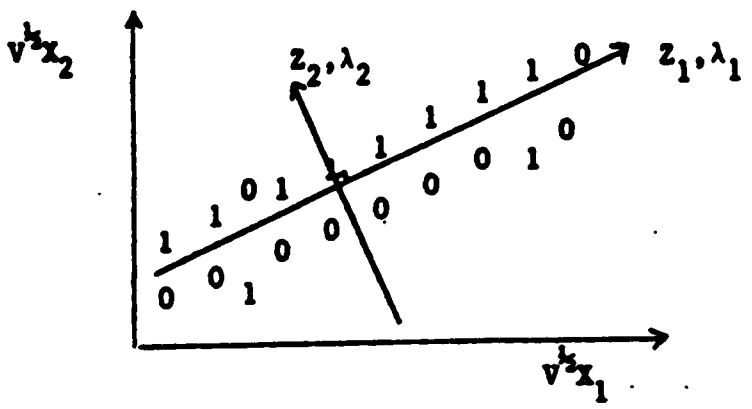

Figure 2. PRINCIPAL COMPONENT PLOT: SIGNIFICANT Z2 SLOPE

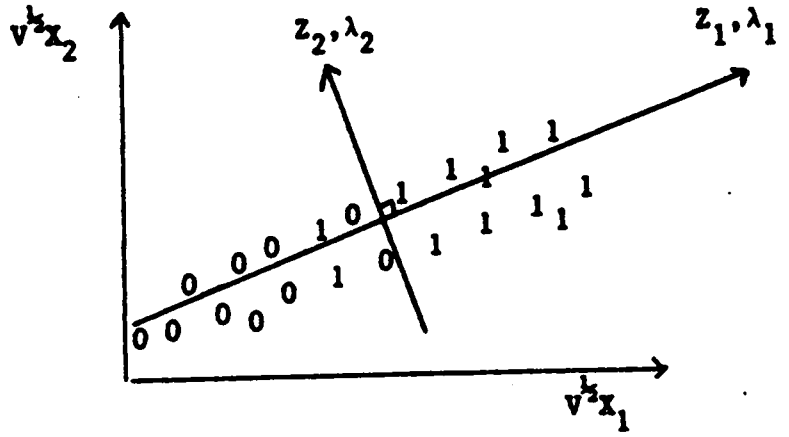

Figure 3. PRINCIPAL COMPONENT PLOT: INSIGNIFICANT Z2 SLOPE

$$E(\hat{\underline{\beta}}'\hat{\underline{\beta}}) = \sum_{l=0}^{p} \lambda_l^{-1} + \underline{\beta}'\underline{\beta}\,, \tag{3.16.1}$$

which is unbounded. When using ridge estimation, the purpose is to bound $\sum_{l=0}^{p} \hat{\beta}_l^2$. Thus $\sum_{l=0}^{p} \hat{\beta}_l^2$ is subject to the constraint that it must be equal to δ. The ridge solution for the estimated parameter vector minimizes the following Lagrange

$$Q = (\underline{y} - X\underline{\beta}^R)'(\underline{y} - X\underline{\beta}^R) + d(\underline{\beta}^{R'}\underline{\beta}^R - \delta). \tag{3.16.2}$$

In setting $(\partial Q / \partial \underline{\beta}^R) = 0$, the following normal equations are found

$$(X'X + dI)\underline{\beta}^R = X'\underline{y}.$$

For sake of simplicity consider X as centered and scaled without a constant term. Hence the ridge solution is

$$\begin{aligned} \underline{\beta}^R &= (X'X + dI)^{-1}X'\underline{y} \\ &= (X'X + dI)^{-1}X'X\hat{\underline{\beta}}, \end{aligned} \tag{3.16.3}$$

for $d \geq 0$.

In his book, Myers (1986) illustrates by example that the ridge solutions are sensible ones. Nonorthogonal explanatory variables typically create large VIF's. However, the VIF's can be greatly reduced by artificially creating a near orthogonal system simply by attacking the diagonal of $X'X$. The eigenvalues of a matrix can be increased by adding a small increment to the diagonal and hence the condition index will usually be deflated with collinear data.

3.17 PROPERTIES OF THE RIDGE ESTIMATORS

Just as in principal component regression, the goal with ridge estimation is to reduce the ill effects of collinearity. The researcher would like to reduce the variance of the coefficients, hence lowering the VIF's, as well as decrease the variance of predicted values among other improvements in the regression. For simplicity, neglect the column of ones associated with the constant. First note that the matrix M^*, composed of eigenvectors of $X'X$, yields the following diagonalization

$$\begin{aligned} \Lambda^*_d &= M^{*\prime}(X'X + dI)M^* \\ &= \text{diag}\{\lambda^*_i + d\}\,. \end{aligned} \tag{3.17.1}$$

Hence the variance-covariance matrix for for $\underline{\beta}^R$ is

$$\begin{aligned} \sigma^{-2}\,\text{Var}(\underline{\beta}^R) &= (X'X + dI)^{-1}X'X(X'X + dI)^{-1} \\ &= M^*\Lambda^{*-1}_d\Lambda^*\Lambda^{*-1}_d M^{*\prime}. \end{aligned} \tag{3.17.2}$$

Equivalently,

$$\sum_{i=1}^{p} \frac{\text{Var}(\beta^R_i)}{\sigma^2} = \sum_{i=1}^{p} \frac{\lambda^*_i}{(\lambda^*_i + d)^2}\,. \tag{3.17.3}$$

Observe that equation (3.17.3) goes to zero as $d \to \infty$.

Along with a decrease in the $\sum_{i=1}^{p} \text{Var}(\beta^R_i)$, the length of $\underline{\beta}^R$ itself can be considerably decreased when compared to that of $\hat{\underline{\beta}}$. Recall in standard multiple regression that the least squares estimates are

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y} = M^{*\prime}\Lambda^{*-1}M^*X'\underline{y} = \sum_{i=1}^{p} \underline{m}^*_i\underline{m}^{*\prime}_i\lambda^{*-1}_i X'\underline{y}. \tag{3.17.4}$$

Thus $\hat{\underline{\beta}} \to \infty$ as any $\lambda^*_i \to 0$. A similar decomposition of $\underline{\beta}^R$ demonstrates

$$\underline{\beta}^R = \sum_{i=1}^{p} m^*_i(\lambda^*_i + d)^{-1} m^{*\prime}_i X' \underline{y}.$$

Hence $\underline{\beta}^R$ is shrunk toward zero for given $d > 0$.

Similar gains are present for expressions of variance inflation factors (VIF's). In standard multiple regression, recall

$$\text{VIF}_i = \sigma^{-2} \text{Var}(\hat{\beta}_i) = (1 - R_i^2)^{-1} = \sum_{j=1}^{p} \lambda^{*-1}_j m^{*2}_{ij}, \tag{3.17.5}$$

which are clearly inflated for small eigenvalues. In the ridge setting

$$\text{VIF}_i = \sigma^{-2} \text{Var}(\beta_i^R) = \sum_{j=1}^{p} m^{*2}_{ij} \lambda^*_j (\lambda^*_j + d)^{-2}, \tag{3.17.6}$$

using the fact that the variance-covariance matrix of $\underline{\beta}^R$ is $(X'X + dI)^{-1}X'X(X'X + dI)^{-1}$ apart from σ^2. The VIF in equation (3.17.6) does not have a standard of unity and can have values less than one.

However, with a decrease in variance in parameter estimates comes an increase in bias. In fact, variance is a strictly decreasing function in d, whereas $\text{Bias}^2(\underline{\beta}^R)$ is a strictly increasing function of d. Hoerl and Kennard (1970a) plot variance, bias and mean squared error (MSE) as a function of d. There exists a window, say $[0, \omega]$, where the decrease in variance outweighs the $\text{Bias}^2(\underline{\beta}^R)$. Thus, in using the mean squared error criterion, the ridge approach appears to be perfectly reasonable so long as $0 \leq d \leq \omega$. To quantify the bias portion consider the following argument. Premultiplying the normal equations by $(X'X)^{-1}$, the following holds:

$$(X'X)^{-1}(X'X + dI)\underline{\beta}^R = (X'X)^{-1}X'\underline{y} = \hat{\underline{\beta}}.$$

The following expectation results.

$$\begin{aligned} E(\underline{\beta}^R) &= (X'X + dI)^{-1}X'X\underline{\beta} \\ &= [(X'X + dI)^{-1}(X'X + dI) - d(X'X + dI)^{-1}]\underline{\beta} \\ &= [I - d(X'X + dI)^{-1}]\underline{\beta}. \end{aligned} \tag{3.17.7}$$

Thus consider the following expression for bias

$$\sigma^{-2}\sum_{i=1}^{p} \text{Bias}^2(\beta_i^R) = \sigma^{-2}d^2\,\underline{\beta}'(X'X + dI)^{-2}\underline{\beta}.$$

Hence MSE can be expressed as

$$\begin{aligned} \sigma^{-2}\sum_{i=1}^{p} MSE(\beta_i^R) &= \sigma^{-2}\sum_{i=1}^{p} \text{Var}(\beta_i^R) + \sigma^{-2}\sum_{i=1}^{p} \text{Bias}^2(\beta_i^R) \\ &= \sum_{i=1}^{p} \lambda^*_i(\lambda^*_i + d)^{-2} + \sigma^{-2}d^2\underline{\beta}'(X'X + dI)^{-2}\underline{\beta} \\ &= \sum_{i=1}^{p} \lambda^*_i(\lambda^*_i + d)^{-2} + \sigma^{-2}d^2\sum_{i=1}^{p} \alpha_i^2(\lambda^*_i + d)^{-2}, \end{aligned} \tag{3.17.8}$$

where $\underline{\alpha}_i = M'\underline{\beta}$. The values of d for which mean squared error will be improved over the least squares estimator is $0 < d < (\sigma^2 / \alpha^2_{max})$.

Also comparisons in prediction abilities of the standard multiple model can be made. It has been noted that

$$\frac{\text{Var}\,\hat{y}(x_o)}{\sigma^2} = \sum_{i=1}^{p} z_{i,o}^2\,\lambda^{*-1}_i, \tag{3.17.9}$$

where z_i are the coordinates corresponding to the principal components. The counterpart to the above prediction variance using ridge estimates is

$$\frac{\text{Var}\, y^R(x_o)}{\sigma^2} = \sum_{i=1}^{p} z_{i,o}^2 \lambda^*_i (\lambda^*_i + d)^{-2}. \tag{3.17.10}$$

Notice how d dominates the small eigenvalues and that the prediction variance can be greatly reduced for even a small value of d.

3.18 METHODS FOR CHOOSING THE SHRINKAGE PARAMETER

Section 3.20 and Chapter 6 will discuss various options for choosing d in the GLM framework. The results given are simplified if the identity link function is used with normal data. Since the goal of this dissertation is not to develop standard ridge regression, the theoretical development for selection of d will not be given. Perhaps the most elementary method is one termed ridge trace (Hoerl and Kennard (1970b)). Quite simply, this procedure plots the estimated coefficients as a function of d. Choose d at a point where they have stabilized. Other more prediction oriented methods have incorporated a PRESS or C_p^R statistic. The C_p^R statistic is developed and generalized in section 6.4. Further, one step and iterative harmonic mean methods have been established as a conservative technique. In section 6.5, a DF-trace procedure is generalized from Tripp's (1983) dissertation. There exist literally hundreds of variations in methods of selecting d based on sundry criteria from prediction to estimation.

3.19 GENERALIZATIONS IN RIDGE REGRESSION

By no means is the above introduction to ridge regression a comprehensive one. It is not within the scope of this dissertation to address ridge regression in its entirety. A natural extension to the work given in section 3.17 is generalized ridge regression. The generalized ridge regression solutions have the form

$$\underline{\beta}^{GR} = (X'X + \Delta)^{-1}X'\underline{y}, \tag{3.19.1}$$

where $\Delta = \text{diag}\{d_{ii}\}$. Hence each eigenvalue, λ^*_i, is artificially increased by its own respective d_{ii} in an attempt to create an orthogonal system. Typically $0 \leq d_{ii} \leq (\sigma^2 / \alpha_i^2) =$ minimum value of $MSE(\hat{\beta}_i)$. It should also be mentioned that generalized ridge is usually performed to the set of explanatory variables which has been orthogonally transformed via the spectral decomposition of the correlation matrix.

Heavily relying on Hoerl and Kennard's (1970a) development of ridge regression, Schaefer (1979) extends ridge estimators into logistic regression. The construction of the ridge logistic estimator will be presented in the next section. Some suggestions for choosing a shrinkage parameter, d, will also be given. In Chapter 6, the logistic logistic estimator will be shown to be a member of the broader GLM class. The GLM ridge estimators will then eventually be shown to be a member of an even broader class of shrinkage estimators, termed generalized fractional principal component estimators, given in Chapter 7.

3.20 RIDGE LOGISTIC ESTIMATORS

With Bernoulli response data, a competitor to PCLR is the ridge logistic estimator developed by Schaefer (1979). Schaefer proposes that a reasonable alternative estimate would be one with a smaller norm than that of maximum likelihood. Recall that the ML estimates can be too long on the average in the presence of an ill-conditioned information matrix. Of course the null vector has

the shortest norm, but is not of any information. Hence a good start in developing a ridge estimator would be to take a parallel approach to that of Hoerl and Kennard (1970a) used in standard multiple regression. Hoerl and Kennard utilize the definition of least squares solutions, $\hat{\underline{\beta}}$, which minimize the sum of squared error, SSE, in standard multiple regression. Using the result that $\underline{\beta}^R = (X'X + dI)^{-1}X'X\hat{\underline{\beta}}$, it follows by substitution that

$$SSE(\underline{\beta}^R) = SSE(\hat{\underline{\beta}}) + (\hat{\underline{\beta}} - \underline{\beta}^R)'X'X(\hat{\underline{\beta}} - \underline{\beta}^R). \tag{3.20.1}$$

Note then that

$$SSE(\underline{\beta}^R) = SSE(\hat{\underline{\beta}}) + \delta, \tag{3.20.2}$$

for $\delta > 0$. Thus a constraint can be written as

$$\delta = (\hat{\underline{\beta}} - \underline{\beta}^R)'X'X(\hat{\underline{\beta}} - \underline{\beta}^R). \tag{3.20.3}$$

The appropriate counterpart constraint in logistic regression is

$$\delta = (\hat{\underline{\beta}}_{ML} - \underline{\beta}^R)'X'VX(\hat{\underline{\beta}}_{ML} - \underline{\beta}^R), \tag{3.20.4}$$

since logistic regression is developed in a weighted sense. In fact, logistic ridge regression inflates the weighted SSE ($WSSE$) by an increment $\delta > 0$. That is

$$WSSE(\underline{\beta}^R) = WSSE(\hat{\underline{\beta}}_{ML}) + \delta. \tag{3.20.5}$$

The development of the ridge estimator is explained in detail in Schaefer's (1979) dissertation. The notion of ridge estimation in a general weighted sense will be developed in Chapter 6 for the generalized linear model. Schaefer requires $\underline{\beta}^R$ to be a consistent estimator of $\underline{\beta}$ and cleverly re-expresses equation (3.20.5) using a first order Taylor series expansion in deriving equation (3.20.4). Hence the ridge estimator developed by Schaefer (1979) is given by

$$\underline{\beta}^R(d) = (X'\hat{V}X + dI)^{-1}X'\hat{V}X\hat{\underline{\beta}}, \tag{3.20.6}$$

where $\hat{V}$ is the estimate of V using the maximum likelihood estimates, $\hat{\underline{\beta}}$. The shrinkage parameter d is the Lagrange multiplier. Schaefer's methods for choosing d in practice relied on the similarities between multiple and logistic regression. Three analogs were investigated

$$\begin{aligned} d_1 &= 1 / (\hat{\underline{\beta}}'\hat{\underline{\beta}}) \\ d_2 &= [\max_j |\hat{\alpha}_j|]^{-2} \\ d_3 &= (p+1) / (\hat{\underline{\beta}}'\hat{\underline{\beta}}). \end{aligned} \tag{3.20.7}$$

Some of the more sophisticated methods of choosing d are based on predictive abilities, as well as accurately estimate parameters. These methods will be presented with the development of ridge estimation in the generalized linear model.

To evaluate the accuracy of parameter estimation for the logistic ridge estimator, Schaefer et al. (1984) presented some examples. The measure of closeness to the true parameter vector was determined by

$$SQE(\hat{\underline{\beta}}) = (\hat{\underline{\beta}} - \underline{\beta}^*)'(\hat{\underline{\beta}} - \underline{\beta}^*).$$

$\hat{\underline{\beta}}$ is the maximum likelihood estimate using a subset of observations and $\underline{\beta}^*$ is the maximum likelihood estimate using all the observations. Schaefer's justification to this approach is that since the bias of the maximum likelihood estimates is $o(N^{-2})$, then $\underline{\beta}^*$ is a reasonable estimate of $\underline{\beta}$. Certainly an argument can be given that $\underline{\beta}^*$ is nearly unbiased; however Schaefer et al. do not address the variance of $\underline{\beta}^*$ which is likely to be quite large especially in the presence of collinearity which is purposely induced. Unbiasedness of $\underline{\beta}^*$ does not guarantee $\|\underline{\beta}^* - \underline{\beta}\|^{1/2} \cong 0$. From a MSE point of view, this given measure of closeness is not a reasonable one. Simulations for logistic and Poisson regressions will be presented in chapter 8.

Chapter IV

ILL-CONDITIONED INFORMATION MATRICES

4.1 COLLINEARITY VS. AN ILL-CONDITIONED INFORMATION MATRIX

In standard multiple regression, a near-deficiency in the X matrix of explanatory variables can result in problems for estimation of the least squares parameter vector, $\underline{\beta}$. When an explanatory variable does not provide any more information that is already inherent in the other regressors, it becomes difficult to separate the influence due to each individual variable on the response (Belsley, Kuh, Welsch (1980)). Multicollinearity, in the above sense, can lead to inversion problems of the information matrix and can further result in large variances associated with the estimated coefficients, as well as wrong signs and magnitude of estimated coefficients, insignificant t-statistics for important regressors, extreme sensitivity to small perturbations to the data, and poor prediction outside the main stream of collinearity (Myers (1986)).

Although collinearity diagnostics and corrective actions have been thoroughly developed for standard multiple regression, little work has been done for such problems in the generalized linear model other than the special case given previously for logistic regression. Schaefer (1979) contends that problems in maximum likelihood parameter estimation can also exist in logistic regression. The contention is that $X'X$, in standard multiple regression, and $X'VX$, in logistic regression, both suffer from collinearity among the explanatory variables, X. Schaefer has an argument, given below, that the elements of $(X'VX)^{-1}$ are large in absolute value when the degree of multicollinearity, in the matrix X, is severe. An outlined proof to Schaefer's conjecture will be presented. Comments will also be forthcoming in an attempt to clear up some implications of the argument. Further some diagnostic techniques will be developed for guidance in variable deletion.

Schaefer's conjecture, mentioned above, can be found in his 1979 Ph.D. dissertation. A sketch of the proof follows. Let $K^{-1} = V = \text{diag}\{\pi_i(1-\pi_i)\}$. Partition the information matrix as follows,

$$X'VX = \begin{bmatrix} x'_i V x_i & x'_i V X_{-i} \\ X'_{-i} V x_i & X'_{-i} V X_{-i} \end{bmatrix}. \tag{4.1.1}$$

X_{-i} is the matrix of explanatory variables without the i^{th} column. The inversion of the partitioned matrix $X'VX$ is

$$(X'VX)^{-1} \equiv t^{-1} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}, \tag{4.1.2}$$

where

$$
\begin{aligned}
t^{-1} &= [x'_i V x_i - x'_i V X_{-i}(X'_{-i} V X_{-i})^{-1} X_{-i} V x_i]^{-1} \\
T_{11} &= 1 \\
T_{21} &= T_{12} = -x'_i V X_{-i}(X'_{-i} V X_{-i})^{-1} \\
T_{22} &= t(X'_{-i} V X_{-i})^{-1} \\
&\quad + (X'_{-i} V X_{-i})^{-1} X'_{-i} V x_i x'_i V X_{-i}(X'_{-i} V X_{-i})^{-1}.
\end{aligned}
\tag{4.1.3}
$$

With severe collinearity, the regression

$$
x_i = X_{-i} f + \xi_i, \tag{4.1.4}
$$

has $SSE = \xi'_i \xi_i \to 0$ as x_i nears an exact linear combination of the columns of X_{-i} (see Schaefer (1979) or Myers (1986)). Schaefer substitutes equation (4.1.4) into equation (4.1.3). Thus

$$
\begin{aligned}
t &= x'_i V x_i - x'_i V X_{-i}(X'_{-i} V X_{-i})^{-1} X'_{-i} V x_i \\
&= \xi'_i [V - V X_{-i}(X'_{-i} V X_{-i})^{-1} X'_{-i} V] \xi_i \\
&= \xi'_i \Omega \xi_i.
\end{aligned}
\tag{4.1.5}
$$

In showing that the diagonal elements of Ω are bounded, Schaefer claims that as $\xi'_i \xi_i \to 0$ then $t \to 0$ and thus $t^{-1} \to \infty$ with severe collinearity among the X's . Note that the diagonal elements of Ω are bounded since the V_{ii} are trivially bounded by (0, .25) and $X_{-i}(X'_{-i} V X_{-i})^{-1} X'_{-i}$ is a projection matrix which always has finite elements. Further, by assumptions, X_{-i} is finite and $N^{-1} \lim_{N \to \infty} (X'_N V X_N) = Q$, for Q positive definite with finite determinant.

Next Schaefer demonstrates that T_{11}, T_{12}, T_{22} are also bounded and with $t^{-1} \to \infty$ the result is complete. $T_{11} = 1$ is trivially bounded. Recall

$$
\begin{aligned}
T_{12} &= -f'_i + \xi'_i V X_{-i}(X'_{-i} V X_{-i})^{-1} \\
T_{21} &= T_{12}.
\end{aligned}
\tag{4.1.6}
$$

As $\xi'_i \xi_i \to 0$, then $T_{12} \to -f'_i$. f'_i is nonnull since X_{-i} is full column rank. Also $T_{22} \to f_i f'_i$ which is bounded away from null for similar reasons.

Schaefer concludes that the matrix $(X'VX)^{-1}$ has large elements in absolute value when the degree of collinearity is severe in X; hence the same problems that occurred with collinearity among the explanatory variables in standard multiple regression, also exist in logistic regression. $\underset{\sim}{\xi}'_j\underset{\sim}{\xi}_j \to 0$ ultimately yields poor precision of the estimated coefficients. In summary, Schaefer states that as collinearity of the independent variables (the columns of the X matrix not the $V^{1/2}X$ matrix) becomes more severe then the following are equivalent:

(i) R_j^2, the coefficient of determination from the regression of the j^{th} independent variable on the remaining independent variables, tends to one for some j.

is equivalent to

(ii) $(\underset{\sim}{\xi}'_j\underset{\sim}{\xi}_j)$, the SSE from (i), tends to zero for some j.

is equivalent to

(iii) λ_{min}, the smallest eigenvalue of $X'VX$, tends to zero.

When the X data exhibits collinearity, Schaefer's argument for the existence of similar problems with $X'VX$ in logistic regression as with $X'X$ in standard multiple regression is convincing. However, care must be taken in understanding the true role of X in the ill-conditioning of $X'VX$. Consider the following example (Burdick (1987)),

$$X = \begin{bmatrix} 1.00 & .98 \\ -.02 & .02 \\ -.98 & -1.00 \end{bmatrix}, \quad X'X = \begin{bmatrix} 1.9608 & 1.9596 \\ 1.9596 & 1.9608 \end{bmatrix}. \tag{4.1.7}$$

Clearly X has near column deficiency and $X'X$ is nearly singular with condition index (of the centered and scaled data) $\lambda_{max} / \lambda_{min} = 1.99939 / .0006119 = 3267.511$. Condition indices are defined formally in section 4.3. Let

$$V = \begin{bmatrix} .00005 & 0 & 0 \\ 0 & .245 & 0 \\ 0 & 0 & .00005 \end{bmatrix}. \tag{4.1.8}$$

Thus

$$X'VX = \begin{bmatrix} .00019602 & 2.033E-20 \\ 2.033E-20 & .00019602 \end{bmatrix}. \tag{4.1.9}$$

The condition index of $X'VX$ (centered and scaled $V^{1/2}X$ data) is $\lambda_{max} / \lambda_{min} = 1.19321 / .806793 = 1.4789$. Thus $X'VX$ need not be near singular when $X'X$ is near singular. In terms of the condition index, $X'X$ is ill-conditioned, whereas $X'VX$ is not. Schaefer is correct in stating that the diagonal elements of $(X'VX)^{-1}$ or the variances of the estimated coefficients, along with the off-diagonals, will be large in absolute value. Examples can be contrived with the same phenomena but the off-diagonals are zero. Hence $V^{1/2}X$ may have orthogonal columns and be well conditioned with extremely collinear X's . Schaefer's result only shows that as the collinearity becomes more and more severe with a fixed V will the off diagonals of $(X'VX)^{-1}$ become large.

In general, given a matrix X that is near deficient in column rank, $X'X$ is near singular. Thus $R_j^2 \to 1$ for some j ; the coefficient of determination from the regression of the j^{th} independent variable on the remaining independent variables tends to one for some j. There exists $a_1, a_2, \ldots, a_p$ such that $a_i \neq 0$ for all i and $\sum_{i=1}^{p} a_i x_i \cong 0$. Recall the information $X'K^{-1}X = X'K^{-1/2}K^{-1/2}X = S'S$. Let $K^{-1/2} = \text{diag}\{\gamma_i\}$. If $X'K^{-1}X$ is also near singular, then $c_1, c_2, \ldots, c_p$ may be found such that $c_i \neq 0$ for all i and $\sum_{i=1}^{p} c_i s_i \cong 0$. Without loss of generality, consider the example of X having dimensions 3×2 .

$$U = \begin{bmatrix} \gamma_1 & 0 & 0 \\ 0 & \gamma_2 & 0 \\ 0 & 0 & \gamma_3 \end{bmatrix} \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \end{bmatrix}$$
$$= \begin{bmatrix} \gamma_1 x_{11} & \gamma_1 x_{21} \\ \gamma_2 x_{12} & \gamma_2 x_{22} \\ \gamma_3 x_{13} & \gamma_3 x_{23} \end{bmatrix}$$

Assume there exists c_i defined above. Then

$$c_1 \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} \# x_1 + c_2 \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} \# x_2 \cong 0,$$

where # indicates elementwise multiplication. Notice if the γ_i are nearly all equal then $c_i \underline{\gamma} \# x_i$ can be redefined as $a_i x_i$. However, in general, the γ_i are not all equal; in fact the γ_i vary considerably in the generalized linear model. Hence by contradiction, if $X'X$ is near singular then $X'K^{-1}X$ does not necessarily need to be near singular, unless K^{-1} is nearly proportional to the identity matrix. Interestingly enough, it can be shown in logistic regression that $\lambda_i \geq 4\lambda_i^*$, where λ_i^* and λ_i are the i^{th} ordered eigenvalue of $X'X$ and $X'K^{-1}X$ respectively. This does imply that if λ_{min} is "small", then λ_{min}^* is necessarily "small". This follows from the fact that $0 \leq \gamma_i \leq .5$.

Expanding on the fact that Schaefer strictly speaks of the effect of a limiting exact deficiency in X on the elements of $(X'VX)^{-1}$ and not the rate at which damaging inflations occur, for illustration, let a matrix X be of the form

$$X = \begin{bmatrix} a & ka \\ c & (k+\varepsilon)c \end{bmatrix}, \tag{4.1.10}$$

where $\varepsilon \in (-\infty, \infty)$ and k some arbitrary constant. When $\varepsilon = 0$, X is deficient; $X'X$ and $X'VX$ are singular. As $|\varepsilon|$ deviates from zero, X becomes less and less deficient. For some fixed a, c, ε, k, X may appear to be of no real threat in terms of deficiency and $X'X$ is not considered

near singular. For the same fixed a, c, ε, k, $X'VX$ can have severe inversion problems. Typically this will happen when several of the data points are predicted quite well, say with $\pi \cong 0$ or 1, and thus V has several diagonal elements near zero. This is not uncommon. In strictly relying on a well behaved $X'X$ matrix to determine if $(X'VX)^{-1}$ is well behaved, then there is a risk of being misled; i.e. the diagonal elements of $(X'VX)^{-1}$ can be inflated. On the other hand, the example given previously in equation (4.1.7) suggests that the information matrix in standard multiple regression can be deemed ill-conditioned, whereas the information matrix in logistic regression may not be. Anything can happen. For a given X matrix of explanatory variables with some fixed severity of collinearity, one cannot make general statements on whether $(X'VX)^{-1}$ will be deemed ill-conditioned or not.

Schaefer's conjecture (iii) above can be alternatively viewed as an argument for the continuity of $\lambda_{\min}$ at the point zero. Since $X'VX$ is positive definite, $\lambda_{\min} > 0$. In the case of an exact deficiency in X, then both $X'X$ and $X'VX$ are singular and $\lambda_{\min} = 0$. However, in departing from the exact collinearity in X, then one can think of $\lambda_{\min}$ as a continuous function. Let X be of the form in equation (4.1.10). For fixed a, c and k and $V = \text{diag}\{v_i\}$, then the eigenvalues of $X'VX$ are given by the following solutions to the quadratic equation,

$$|X'VX - \lambda I| = 0.$$

The

$$\text{Roots} = \frac{b}{2} \pm \frac{1}{2}(b^2 - 4v_1v_2a^2c^2\varepsilon^2)^{1/2}, \tag{4.1.11}$$

where

$$b = v_1a^2(k^2 + 1) + v_2c^2((k + \varepsilon)^2 + 1).$$

Thus

$$\lambda_{\min} = f(\varepsilon) = \frac{b}{2} - \frac{1}{2}(b^2 - 4v_1v_2a^2c^2\varepsilon^2)^{1/2}. \tag{4.1.12}$$

$\lambda_{\min}$ is continuous at zero by the following argument. Notice f: $R \to R^+$ and $0 \in R$. f is continuous at zero if and only if for each $\varepsilon^* > 0$, there exists an $\varepsilon^{**} > 0$ such that if

$$|\varepsilon| < \varepsilon^*, \tag{4.1.13}$$

then

$$0 \le f(\varepsilon) < \varepsilon^{**}.$$

Continuity at zero holds for $\lambda_{\min}$ in equation (4.1.12). All nontrivial examples yield $\lambda_{\max}$ bounded away from zero. Schaefer's claim of $\xi_i'\xi_i \to 0$ is also equivalent to the condition index going to infinity. In terms of the example given in equation (4.1.10), Schaefer essentially points out that regardless of V, an ε can be found arbitrarily close to zero such that $X'VX$ is ill-conditioned leading to inflated elements in $(X'VX)^{-1}$.

In a more global setting such as in the generalized linear model (Nelder and Wedderburn (1972)), recall the general weight matrix is K^{-1} where

$$K^{-1} = \text{diag}\{k_{ii}^{-1}\}, \tag{4.1.14}$$

where $0 < k_{ii}^{-1} = [h'(\eta_i)]^2 / \hat{\text{Var}}(Y) < \infty$. It is interesting to note that the smallest eigenvalue of the information $X'K^{-1}X$ also goes to zero as the collinearity among the columns of X becomes more severe. Hence Schaefer's conjecture holds for the generalized linear model even though the diagonal elements of K^{-1} are not bounded, but fixed. For the generalized linear model, the diagonal elements of K^{-1} may vary considerably, living anywhere on the positive real line; for example, in the Poisson response in the discrete case and the Gamma response in the continuous case (see Table 3). In fact, boundedness of the k_{ii}^{-1} is not a key factor to Schaefer's proof, as presented. Diagnostic

tools should be developed to determine if an alternate estimation procedure is needed based strictly on the condition of $X'K^{-1}X$ rather than that of $X'X$.

4.2 COLUMN SCALING FOR DIAGNOSTICS

In terms of model building, the choice of scale for the X explanatory variables is usually the units of the researcher's convenience. Essentially equivalent model structures can be built whether the researcher chooses, for example, units of ounces, milliliters or cubic inches. However, as in standard multiple regression, scale changes do in fact change the diagnostic's numerical properties when we try to assess the conditioning for the information matrix of the generalized linear model. In particular, a change in scale can result in very different singular value decompositions.

In order to make a comparison of condition indexes meaningful, it is necessary to standardize the information matrix in an effort to obtain a stable diagnostic. Again the standardization cannot be simply done on the explanatory variables, but rather on the weighted $\hat{S} = \hat{K}^{-1/2}X$ variables. A natural scaling method, giving the columns of $\hat{S}$ unit length, is given in equation (4.2.1). The author also chooses to center $\hat{S}$. Note, for diagnostic purposes mentioned above, that the $\hat{S}$ data only needs to be scaled and not centered. Essentially, the matrix $\hat{S} = \{s_{ij}\}$ is centered and scaled by letting

$$s_{ij} = (s_{ij} - \bar{s}_j)\left[\sum_{j=1}^{N}(s_{ij} - \bar{s}_j)^2\right]^{-1/2}. \tag{4.2.1}$$

The given approach to scaling is a natural one because under ideal conditions, that is, when the columns of $\hat{S}$ are mutually orthogonal, then the condition index is unity. Any other choice of scale fails to meet this desirable property (see Belsley, Kuh and Welsch (1980)). Consequently, any condition index or variance proportion decomposition mentioned in a diagnostic sense or presented

in an example will be introduced in a standard form by first estimating $\hat{K}$ via maximum likelihood and scaling $\hat{S} = \hat{K}^{-1/2}X$ to have unit column length.

4.3 DIAGNOSTIC TOOLS FOR THE INFORMATION MATRIX

In standard multiple regression, examination of the spectral decomposition of the correlation matrix of explanatory variables probably has been one of the most fruitful techniques of detecting and combating collinearity. See Kendall (1957) and Myers (1986). In the past (Silvey (1969)), collinearity was often diagnosed when a "small" eigenvalue in the correlation matrix was observed. Of course, the smallest eigenvalue can be made arbitrarily small or large depending on which scale the researcher wishes to use in data collection. In effect, this is equivalent to claiming that a square matrix B is ill-conditioned when the determinant is small. This is simply not true since any well conditioned matrix $C = 10^{k}B$ is likely to have a small determinant when $k = -20$. Belsley, Kuh, and Welsch (1980) point out that if a "small" eigenvalue is used as a collinearity diagnostic, then there is a natural tendency to compare small to the wrong standard, namely zero. Perhaps a collinearity can be easier to identify if a "small" eigenvalue is small in relation to the other eigenvalues. The ratio of the largest eigenvalue to the smallest eigenvalues is one such indicator of an ill-conditioned matrix.

As early as 1952, Hartree pointed out the importance of a condition index as a means of determining the ill-conditioning of a general matrix, B. Currently there are several numerical indicators available to determine a measure of ill-conditioning of a square matrix, B (if B is rectangle, then form $B'B$). Some of the many variations of the condition index (ψ) include:

$$\psi_1 = \frac{\lambda_{\max}}{\lambda_{\min}} \quad (>1000)$$
$$\psi_2 = \sqrt{\psi_1} \quad (>30)$$
$$\psi_3 = \frac{\lambda_{\min}}{\mathrm{tr}(X'K^{-1}X)} = \frac{\lambda_{\min}}{\Sigma\lambda_i} \tag{4.3.1}$$
$$\psi_4 = \frac{\sqrt{\lambda_{\min}}}{\Sigma\sqrt{\lambda_i}}.$$

λ_i is an eigenvalue and $\lambda_i^{1/2}$ is a singular value of $X'K^{-1}X$. Recall from section 2.10 that the matrix X has already been centered and scaled. For the condition index diagnostic measure, the columns of the matrix $\hat{S} = \hat{K}^{-1/2}X$ will further be centered and scaled as presented in section 4.2; this gives a standard of unity for ψ_1. In forming a "correlation" matrix $\hat{S}'\hat{S}$, the eigenvalue decomposition spectrum will determine the conditioning of the estimated information matrix, $X'\hat{K}^{-1}X$. Notice ψ_3 and ψ_4 appeal to the proportion of variability in a principal component context. The $\sum_{i=0}^{p} \lambda_i$ quantifies the total variation of matrix $X'K^{-1}X$. That is the overall spatial variation of a cloud of points, depicted by $X'K^{-1}X$, is quantified by their total inertia. Greenacre (1984) connects the formulation of inertia in the physical sense to that of one in a statistical sense. The moment of inertia is often thought as the integral of mass times squared distance to the centroid. With categorical data, Greenacre views inertia as Pearson's mean squared contingency coefficient.

A justification for using a condition index as a diagnostic measure is outlined below. The notion of an ill-conditioned square matrix is often of one which is near singular, and for which an inflation of its inverse occurs. The motivation for the development of the condition index as a measure of ill-conditioning is presented by Belsly, Kuh, and Welsch (1980). Consider the singular value decomposition of any matrix, $T_{n\times p}$

$$T = UDV', \tag{4.3.2}$$

where $U'U = I_p = V'V$ and D is a diagonal matrix with the nonnegative singular values of T, $\mu_1 < \mu_2 < \ldots < \mu_p$. U and V are the eigenvectors of TT' and $T'T$ respectively (see Good (1969)). The general Euclidean norm of any $p \times p$ matrix B is defined by the specified norm, denoted by $\|B\|$, where

$$\|B\| = \sup_{\|\underline{b}\|=1} \|B\underline{b}\|. \tag{4.3.3}$$

The spectral norm is relevant to the nonsingular solution to the linear system $B\underline{b} = \underline{c}$ given by $\underline{b} = B^{-1}\underline{c}$. Belsley, Kuh, and Welsch (1980) consider how much the solution $\underline{b}$ will change due to small perturbations in the elements of B or $\underline{c}$. Consider δB and $\delta\underline{c}$. Let B be fixed and $\underline{c}$ change to $\delta\underline{c}$. Thus

$$\begin{aligned} \delta\underline{b} &= B^{-1}\delta\underline{c} \\ \|\delta\underline{b}\| &\le \|B^{-1}\|\,\|\delta\underline{c}\|. \end{aligned} \tag{4.3.4}$$

Further

$$\begin{aligned} \underline{c} &= B\underline{b} \\ \|\underline{c}\| &\le \|B\|\,\|\underline{b}\|. \end{aligned} \tag{4.3.5}$$

From equations above,

$$\frac{\|\delta\underline{b}\|}{\|\underline{b}\|} \le \|B\|\,\|B^{-1}\|\,\frac{\|\delta\underline{c}\|}{\|\underline{c}\|}. \tag{4.3.6}$$

The quantity $\|B\|\,\|B^{-1}\|$ provides a bound for the impact of relative changes in $\underline{c}$ on the solution $\underline{b}$. It can be shown that $\|B\| = \mu_{\max}$ of B and $\|B^{-1}\| = \mu_{\min}$ of B. Thus

$$\|B\|\,\|B^{-1}\| = \mu_{\max} / \mu_{\min} = (\lambda_{\max} / \lambda_{\min})^{1/2}. \tag{4.3.7}$$

A similar argument can be developed for perturbations in the matrix B.

This result is not only useful in the context of least squares solutions, as given by Belsley, Kuh, and Welsch, but also lends itself to maximum likelihood solutions of the generalized linear model. Recall at each iteration, the solution is of the form

$$\begin{aligned}\hat{\underline{\beta}}_t &= \hat{\underline{\beta}}_{t-1} + (X'\hat{K}_{t-1}^{-1}X)^{-1}\left[\sum_{i=1}^{N} x_i \hat{k}_{ii}^{-1}(y_i - \hat{\mu}_i)\frac{\partial \eta_i}{\partial \mu_i}\right]_{t-1} \\ &= [(X'\hat{K}^{-1}X)^{-1}X'\hat{K}^{-1}\underline{y}^*]_{t-1},\end{aligned} \tag{4.3.8}$$

where $y^*_i = \eta_i + (y_i - \mu_i)(\partial \eta_i / \partial \mu_i)$ evaluated at $\hat{\underline{\beta}}_{t-1}$. Letting $B = X'K^{-1}X$, $\underline{b} = \hat{\underline{\beta}}_t$ and $\underline{c} = X'K^{-1}\underline{y}^*$, the above argument suggests that if the ratio of $\lambda_{max} / \lambda_{min}$ for $X'K^{-1}X$ is large, then small changes in the vector $\underline{c}$ can adversely affect the maximum likelihood solution, at each step. Maximum likelihood solutions are sensitive to small perturbations to the data. Equations (4.3.6) and (4.3.7) suggest that the condition index for the information matrix can be a good indicator of ill-conditioning.

A more geometric measure of ill-conditioning of $X'K^{-1}X$ (Burdick (1987)) is measured by an index (ψ_5), where

$$\psi_5 = \frac{|\Phi|}{\prod_{i=0}^{p} \Phi_{ii}} \quad \text{where} \quad \Phi = X'K^{-1}X. \tag{4.3.9}$$

Define $|\Phi|$ as the determinant of Φ. The index ψ_5 is necessarily in the unit interval. The interpretation of ψ_5 can be visualized as the ratio of two volumes. The numerator is the volume of a parallelopiped composed of the vectors $K^{-1/2}\underline{x}_i \quad i = 0, 1, 2, \ldots, p$ starting at the origin. The denominator is the volume of a $p + 1$ dimensional general rectangle in an orthogonal setting to yield maximum volume. Thus if there is a near singularity in $X'K^{-1}X$ then the numerator parallel piped is quite flat in at least one dimension, thus reducing the index ψ_5.

4.4 GENERAL VARIANCE INFLATION FACTORS

In addition to condition indices for $X'\hat{K}^{-1}X$, where

$$\hat{K}^{-1} = \text{diag}\{\hat{k}_{ii}^{-1}\} \text{ and } \hat{k}_{ii}^{-1} = [h'(\hat{\eta}_i)]^2 / \hat{\text{Var}}(Y_i), \tag{4.4.1}$$

other diagnostic tools can be developed. In ordinary least squares with normal data, for example, variance inflation factors (VIF's) are available for the correlation matrix (Belsley, Kuh, and Welsch (1980)). Recall that if the regressors are centered and scaled in multiple regression, then $X'X$ is the information matrix, as well as the correlation matrix of the explanatory variables (ignoring the column of ones for the constant term). Under orthogonality of the explanatory variables, the correlation matrix will then be the identity. This is the ideal. The $\text{Var}(\hat{\beta}_i) = 1.0$ apart from σ^2 , for all i. Hence, in taking the inverse of the centered and scaled data matrix, the diagonal elements of the inverse denote a measure of the inflation of the variances of the coefficients, $\text{Var}(\hat{\beta}_i)$, apart from σ^2. For least squares standard multiple regression, VIF's can also be expressed as

$$\text{VIF}_i = \frac{1}{1 - R_i^2}, \tag{4.4.2}$$

where R_i^2 is the coefficient of multiple determination of the regression produced by regressing X_i on X_{-i} .

The development of a VIF for the generalized linear model is not as cut and dried. It is not proper to look at the inverse of the matrix of correlations as a diagnostic tool. For one, the information matrix is not a scalar multiple of $X'X$; hence the condition of the correlation matrix of explanatory variables may not always coincide with the condition of the information matrix. Recall that the condition of $X'K^{-1}X$ is of interest. Perhaps the most obvious solution for the construction of general VIF's or GVIF is to think of $\hat{S} = \hat{K}^{-1/2}X$ as a new data matrix. Thus, by centering and scaling $\hat{S}$, $\hat{S}'\hat{S}$ will be in the correlation form and will be the identity under ideal conditions (inter-

cept included). Typically, $\hat{K}^{-1/2}$ will have to be estimated via maximum likelihood and assumed to be well estimated for the original data points, even in the presence of collinearity among the $\hat{K}^{-1/2}X$. The $\text{Var}(\hat{k}_{ii}^{-1})$ will be developed in section 5.4 to give some justification to this approach. Nevertheless, the GVIF's seem like a reasonable means to get a measure of asymptotic variance inflation due to the nonorthogonality among the $\hat{K}^{-1/2}X$.

$$\text{GVIF}_i = \text{diagonal elements of } \{(\hat{S}'\hat{S})^{-1}\}. \tag{4.4.3}$$

The general VIF's, in equation (4.4.3), reduce to equation (4.4.2) for maximum likelihood estimation with normal response data and an identity link function.

4.5 GENERAL VARIANCE PROPORTION DECOMPOSITION

Recall the orthogonal matrix, M such that $M'M = MM' = I$ and

$$M'X'K^{-1}XM = \Lambda. \tag{4.5.1}$$

M are a set of eigenvectors for the information matrix and Λ is a diagonal matrix of the $(p+1)$ corresponding eigenvalues. Let $S = K^{-1/2}X$ be centered and scaled. $(X'K^{-1}X)^{-1} = M\Lambda^{-1}M'$. Define, asymptotically,

$$c_{jj} = \text{Var}(\hat{\beta}_j) = \sum_{u=0}^{p} m_{ju}^2 / \lambda_u. \tag{4.5.2}$$

Myers (1986) points out that it is easy to illustrate that a small eigenvalue deposits its influence, to some degree, on all variances. The proportion of variance associated to the j^{th} estimated coefficient, attributed to the i^{th} eigenvalue of the sum in equation (4.5.2), can be expressed as

$$P_{ij} = \frac{m_{ji}^2 / \lambda_i}{c_{jj}}. \qquad (4.5.3)$$

A matrix of proportions can be formatted as in Table 4. Hence a small eigenvalue i (relative to the maximum eigenvalue) responsible for at least two large proportions of variance $P_{i,j}$ and $P_{i,j'}$ suggests precision of estimation may be damaged.

4.6 EXAMPLE USING GENERAL DIAGNOSTICS

The data supplied in Appendix A concerns the prediction of a cancer remission when given six continuous explanatory variables. Hence the response is Bernoulli in nature which lends itself to a logistic model. The explanatory variables are first centered and scaled by the procedure outlined in section 2.10 and then the data matrix is augmented by a column of ones associated with the constant term. Maximum likelihood estimation is employed yielding the following estimates (standard errors):

Maximum Likelihood
(11 Iterations)

Intercept	:	-2.311	(1.800)
X1	:	23.012	(45.975)
X2	:	20.050	(61.358)
X3	:	-22.382	(71.784)
X4	:	9.511	(4.536)
X5	:	-6.527	(4.909)

First notice the large standard errors associated with the parameter estimates for X1, X2, and X3. A first suspicion would be small eigenvalues of the information matrix. Recall that the estimated information matrix for logistic regression is of the form, $X'\hat{V}X$, where

Table 4. WEIGHTED VARIANCE PROPORTION DECOMPOSITION

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

$\hat{V} = \text{diag}\{\hat{\pi}_i(1-\hat{\pi}_i)\}$. $\hat{\pi}_i$ is the maximum likelihood estimate for the probability of a cancer remission given in the i^{th} row of explanatory variables, $\underline{x}'_i$.

In constructing $\hat{S} = \hat{V}^{1/2}X$ and centering and scaling the columns of $\hat{S}$, the information eigenvalues are as follows:

$$\begin{array}{lcl} \lambda_0 & = & 2.41382 \\ \lambda_1 & = & 1.51930 \\ \lambda_2 & = & 1.06446 \\ \lambda_3 & = & .86062 \\ \lambda_4 & = & .15073 \\ \lambda_5 & = & .00106 \end{array}$$

From a rough benchmark of .01 for a small eigenvalue, $X'\hat{V}X$ can be deemed ill-conditioned and deficient in at least one dimension. The condition index is

$$\psi_1 = \frac{\lambda_{\max}}{\lambda_{\min}} = 2277.19,$$

which is considerably above the recommended cutoff of 1000 mentioned in section 4.3.

The general variance inflation factors (GVIF's) developed for weighted data in section 4.4 are:

Intercept	:	4.60
X1	:	63.68
X2	:	407.97
X3	:	471.14
X4	:	2.59
X5	:	2.42

Observe that there exists a GVIF associated with the intercept. The explanation for its presence is due to the fact that the columns of $V^{1/2}X$ are centered and scaled in the construction of a correlation matrix. It is evident that problems exists with GVIF's for parameter estimates again associated with X1, X2 and X3.

By investigating the general variance proportion decompositions outlined in section 4.5, problems can be immediately identified. See Table 5. Via routine analyses of variance proportion decompositions, it is quite obvious that severe collinearity problems exists between $V^{1/2}X_1$, $V^{1/2}X_2$ and $V^{1/2}X_3$. Large proportions of variance associated with a small eigenvalue are corresponding to large GVIF's. Certainly, subset regression is a viable option, and will be discussed in section 5.2, to alleviate problems associated with weighted collinearity. However, since asymptotically biased estimation is the topic, a comparison will be made using alternate estimation procedures outlined in Chapter 3.

Table 6 consists of a variety of estimation techniques for the cancer remission example. One purpose of this table is to demonstrate how much estimation techniques can vary in the logistic setting. A point of interest is the reduction in the standard errors of the coefficients for any biased technique when compared to maximum likelihood. As the eigenvalue structure and deviance measure suggest, it is quite obvious that PC estimation minus two dimensions is not necessary. The shrinkage parameters d_1, d_2, and d_3 are discussed in equation (3.20.7). The shrinkage methods d_{CP} and d_{DF} will be developed in sections 6.4 and 6.5 respectively.

Table 5. VARIANCE PROPORTION DECOMPOSITIONS CANCER EXAMPLE

Eigenvalue	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
2.41382	0.01656	0.00078	0.00030	0.00032	0.00391	0.00204
1.51930	0.00443	0.00053	0.00000	0.00000	0.11226	0.12887
1.06446	0.04659	0.00529	0.00062	0.00017	0.01821	0.00325
.85063	0.06631	0.00497	0.00001	0.00006	0.09245	0.11728
.15073	0.42456	0.01399	0.00001	0.00035	0.71374	0.73266
.00106	0.44153	0.97443	0.99906	0.99910	0.05943	0.01611

Table 6. VARIOUS ESTIMATION TECHNIQUES FOR CANCER EXAMPLE

Estimate (standard error)

	DEV	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
ML	21.755	-2.311 (1.800)	23.012 (44.975)	20.050 (61.359)	-22.382 (71.784)	9.512 (4.536)	-6.527 (4.909)
Schaefer PC(-1)	21.894	-1.798 (1.080)	7.154 (6.209)	-1.774 (2.683)	3.156 (2.802)	9.117 (4.399)	-6.314 (4.872)
Schaefer PC(-2)	31.388	- .343 (0.594)	-2.019 (2.499)	.817 (2.150)	-.489 (1.658)	3.693 (2.838)	-.315 (3.151)
Iterative PC(-1)	21.892	-1.847 (1.080)	7.337 (6.209)	-1.794 (2.683)	3.257 (2.802)	9.282 (4.399)	-6.454 (4.872)
Iterative PC(-2)	42.998	-.968 (0.594)	-8.144 (2.499)	3.889 (2.150)	-1.514 (1.658)	9.713 (2.838)	-6.037 (3.151)
Ridge d_1 = .00064	21.868	-1.803 (1.065)	8.807 (8.208)	1.071 (8.164)	-0.205 (9.444)	8.920 (4.283)	-6.081 (4.737)
Ridge d_2 = .00072	21.874	-1.788 (1.057)	8.546 (7.785)	.787 (7.379)	.123 (8.513)	8.882 (4.267)	-6.043 (4.721)
Ridge d_3 = .00382	22.048	-1.510 (0.931)	5.815 (5.207)	- .799 (2.829)	1.853 (2.950)	7.877 (3.811)	-5.009 (4.199)
Ridge d_{CP} = .0080	22.384	-1.286 (0.831)	4.427 (4.336)	-0.744 (2.358)	1.666 (2.307)	6.937 (3.388)	-4.049 (3.707)
Ridge d_{DF} = .0003	21.836	-1.890 (1.117)	10.703 (12.091)	3.386 (14.610)	-2.897 (17.035)	9.091 (4.350)	-6.239 (4.810)

Chapter V

AN ILL-CONDITIONED INFORMATION MATRIX IN THE GLM

5.1 INTRODUCTION

The damaging effects of multicollinearity are well documented for the generalized linear model when the identity link function is used with normal response data. See Hoerl and Kennard (1970), Webster, Gunst and Mason (1974), and Myers (1986). Schaefer (1986) has further suggested ridge, principal component, as well as Stein estimation procedures for logistic regression when the logit explanatory variables form an ill-conditioned X matrix. Recall in section 4.1, Schaefer (1979) developed an argument that the variance-covariance matrix $(X'K^{-1}X)^{-1}$ for $\hat{\underline{\beta}}$ of the logit model has large elements in absolute value when the degree of

multicollinearity becomes more and more severe in the X data. As noted in Chapter 4, Schaefer's argument of $X'X$ being near singular does not imply in general that $X'K^{-1}X$ is near singular.

The iterative equation for parameter estimates of the GLM suggests that if the information matrix is near singular, then perhaps some alternate estimation technique can be employed with the generalized linear model to improve properties, for example:

i) $\sum_{i=0}^{p} \text{Var}(\hat{\beta}_i) = \text{tr}(\Phi^{-1}) = \sum_{i=0}^{p} (\lambda_i)^{-1} \rightarrow \infty$ as $\lambda_i \rightarrow 0$;

ii) $\text{Var}[\hat{y}(\underline{x}_o)] \cong \left[\frac{\partial \mu_o}{\partial \eta_o}\right]^2 \sum_{i=0}^{p} z_{i,o}^2 \lambda_i^{-1} \rightarrow \infty$ for predictions of new observations outside the mainstream of weighted collinearity when combined with a small λ_i;

iii) For the test

$$H_o: \underline{\beta} = \underline{\beta}_C$$
$$H_1: \underline{\beta} = \underline{\beta}_F,$$

the test statistic, $\chi^2 = \sum_{i=0}^{p} (\hat{\alpha}_{i,C} - \hat{\alpha}_{i,F})^2 \lambda_i \rightarrow 0$ as $\lambda_i \rightarrow 0$, is deflated and hence reduces power (Kendall and Stuart (1973)), where C and F denote the current and full model respectively.

Notice how these damages of a near singular information matrix generalize from the logistic regression setting in section 3.8.

Since the iterative solution for the coefficients relies heavily on the information matrix, $X'K^{-1}X$; condition indices are excellent indicators for a deficiency in this matrix. Action should the be taken accordingly. If $X'K^{-1}X$ is deemed ill-conditioned, then several approaches for alternate estimation will be suggested. The first two alternate parameter estimation procedures developed for the generalized linear model are a ridge and principal component approach sim-

ilar to the work of Schaefer (1979) and (1984), respectively. Forthcoming will be a general class of biased estimators termed generalized fractional principal component estimators. This class of estimators will be shown to be particularly useful when the generalized linear model is in the canonical form.

5.2 VARIABLE DELETION

A common resort to the reduction of multicollinearity in standard least squares multiple regression models is variable deletion (Myers (1986)). The idea is to simply remove the explanatory variables that are inherently collinear with the remaining explanatory variables in the data matrix. The choice of deletion can be done quickly by looking at the correlation matrix of the data or, perhaps more appropriately, by examining the variance inflation factors (VIF's) along with the variance decomposition proportions. The researcher hopes to find a reduction in VIF's along with reductions in variances of regression coefficients with a stable subset model. The subset model should reduce collinearities with minimal loss of pertinent information. The predictive capabilities of the subset model can be compared to that of the original model by examining

i) $\text{PRESS} = \sum_{i=1}^{N} e_{i,-i}^2 = \sum_{i=1}^{N} \frac{e_i^2}{(1-h_{ii})^2}$, where h_{ii} are the diagonal elements of the hat matrix, $H = X(X'X)^{-1}X'$

ii) $H = \frac{\text{Var}(\hat{y})}{\sigma^2}$.

If theoretical models are not specified, then variable deletion is thought of as a convenient means for collinearity reduction in standard least squares multiple regression. However, in the generalized linear model, there exists regressions where deletions based on the collinearities of the columns of X may or may not have an impact on the collinearities of the columns of

$K^{-1/2}X$. Due to the fact that variables which are collinear in the weighted sense damage the information matrix, deletion of variables should be based on weighted collinearities. Chapter 4 has suggested various diagnostics for variable deletion in the GLM.

Schaefer (1979) points out that even though the concept of variable deletion is the most straightforward and easiest to implement, it may be better to use some other technique to remove multicollinearity among the explanatory variables. The variable deletion process removes variables solely on the interdependence of the X data without taking into account the dependent variable. Generalized principal component and ridge estimators will be discussed in this context.

5.3 GENERALIZED PRINCIPAL COMPONENT ANALYSIS (GPCA)

Principal components regression has been introduced in Chapter 3 for both standard multiple and logistic regression. Natural extensions are put forth to the generalized linear model. Examples and simulations will show that GPCA can be applied successfully in a variety of experimental settings. Moreover, the upcoming development has a certain elegance.

Consider the generalized linear model given in equation (2.4.1).

$$\eta_i = g(\mu_i) = \underline{x}'_i\underline{\beta} = \underline{z}'_i\underline{\alpha}\,, \tag{5.3.1}$$

where $\underline{x}'_iM = \underline{z}'_i$, $M'\underline{\beta} = \underline{\alpha}$ and M is the orthogonal matrix yielding the spectral decomposition of the information matrix. The point of view for logistic regression, given in section 3.10, is the same for generalized principal component regression (GPCA). The concern is not so much the exact form of the matrix X in the construction of the information matrix but rather that X is composed of a set of p independent variables having the same scale. The researcher may accomplish this by standardizing, as mentioned in equation (2.10.1). However, if, by design, the

columns of the X matrix are originally the same units, then further standardization may be avoided to allow a more natural interpretation of the results. The work given in this dissertation consistently views X as both centered and scaled. The fact that $X'K^{-1}X$ will not be in a correlation form is not a real issue. Despite the rather natural method of removing scale dependence in ordinary least squares principal components analysis, by means of a matrix in the correlation form, GPCA can accomplish scale removal in a different fashion.

The generalized principal component procedure involves the deletion of some of the components in equation (5.3.1) and finding the maximum likelihood estimates of the remaining components. Consider rewriting the model in equation (5.3.1) in the canonical form

$$\underline{\eta} = Z\underline{\alpha} = (Z_r\, Z_s)\begin{bmatrix}\underline{\alpha}_r \\ \underline{\alpha}_s\end{bmatrix}. \tag{5.3.2}$$

Note that the columns of Z_r represent the deleted principal components. Thus, the restricted canonical model follows

$$\underline{\eta}^{pc} = Z_s\underline{\alpha}_s.$$

A natural approach, from the point of view of GPCA, would be to maximize the likelihood function of $\underline{\alpha}$ given the orthogonally transformed data or the principal components, Z. Thus, it follows from equation (2.4.2) that $\partial l / \partial \underline{\alpha}$ has an unique maximum found by equating the following expressions to zero.

$$\begin{aligned} 0 = \frac{\partial l}{\partial \underline{\alpha}} &= \sum_{i=1}^{N} \frac{\partial l_i}{\partial \theta_i}\,\frac{\partial \theta_i}{\partial \mu_i}\,\frac{\partial \mu_i}{\partial \eta_i}\,\frac{\partial \eta_i}{\partial \underline{\alpha}} \\ &= \sum_{i=1}^{N} z_i\, h'(\eta_i)\,\frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)}. \end{aligned} \tag{5.3.3}$$

Thus, by a similar argument developed in equation (2.4.7), an iterative scheme for $\underline{\alpha}$ can be constructed using the method of scoring.

$$\begin{aligned}\hat{\underline{\alpha}}_t &= \hat{\underline{\alpha}}_{t-1} + \hat{\Lambda}_{t-1}^{-1}\left[\sum_{i=1}^{N} z_i \hat{k}_{ii}^{-1}(y_i - \hat{\mu}_i)\frac{\partial \eta_i}{\partial \mu_i}\right]_{t-1} \\ &= \hat{\Lambda}_{t-1}^{-1}\left[\sum_{i=1}^{N} z_i \hat{k}_{ii}^{-1}\left[z'_i \hat{\underline{\alpha}}_{t-1} + (y_i - \hat{\mu}_i)\frac{\partial \eta_i}{\partial \mu_i}\right]\right]_{t-1} \\ &= [\hat{\Lambda}^{-1} Z' \hat{K}^{-1} \underline{y}^*]_{t-1},\end{aligned} \tag{5.3.4}$$

where $y_i^* = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$ evaluated at $\hat{\underline{\alpha}}_{t-1}$. Note μ_i must be updated at each iteration step. M may be updated at each step, however empirical results suggest using a fixed spectral decomposition of the maximum likelihood estimate of the information. If all the principal components are kept, then $M\hat{\underline{\alpha}}$ is identical to the maximum likelihood estimate of $\hat{\underline{\beta}}$. However, if by choice $r = p + 1 - s$ principal components are deleted, then the iterative scheme becomes

$$\hat{\underline{\alpha}}_{s,t}^{pc} = [\hat{\Lambda}^{-1} Z' \hat{K}^{-1} \underline{y}^*]_{s,t-1}. \tag{5.3.5}$$

A conversion can be made from the principal component parameter estimates to one using the original centered and scaled explanatory variables while improving regression properties with virtually no loss in information. The transformation back to the original variables follows as in equation (3.12.5),

$$\underline{b}_s^{pc} = M_s \hat{\underline{\alpha}}_s^{pc} \tag{5.3.6}$$

Note that, as outlined from equation (3.14.2), the

$$\mathrm{Var}(\underline{b}_s^{pc}) = M_s \Lambda_s^{-1} M'_s. \tag{5.3.7}$$

The bias can be quantified as

$$E(\underline{b}_s^{pc}) = \underline{\beta} - M_r \underline{\alpha}_r. \tag{5.3.8}$$

Discussions of variance and bias carry over from section 3.9. The asymptotic distribution of $\hat{\underline{\alpha}}^{pc}$ will be derived in equation (5.5.9).

As an example, consider the common identity link when $\varepsilon_i \sim N(0, \sigma^2)$.

$$\begin{aligned} y_i &= \underline{x}'_i\underline{\beta} + \varepsilon_i \\ \mu_i &= \underline{x}'_i\underline{\beta} = \underline{z}'_i\underline{\alpha}. \end{aligned} \tag{5.3.9}$$

Thus from equation (5.3.5), the iterative principal component scheme becomes

$$\begin{aligned} \hat{\underline{\alpha}}^{pc}_{s,t} &= \sigma^2(Z'Z)^{-1}Z'\sigma^{-2}\underline{y} \\ &= (Z'Z)^{-1}Z'\underline{y}, \end{aligned} \tag{5.3.10}$$

since $y_i^* = \mu_i + (y_i - \mu_i)\partial\mu_i / \partial\mu_i = y_i$. Notice that this is the usual one step principal component least squares estimator in regression analysis with common variance.

As a second example, consider $Y_i \sim$ binomial(n_i, π_i), and thus

$$g(\mu_i) = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \underline{x}'_i\underline{\beta}. \tag{5.3.11}$$

Thus the iterative principal component scheme becomes, from equation (5.3.4),

$$\begin{aligned} \hat{\underline{\alpha}}_t &= \hat{\underline{\alpha}}_{t-1} + \hat{\Lambda}^{-1}_{t-1}\left[\sum_{i=1}^{N} z_i\hat{k}^{-1}_{ii}(y_i - \hat{\mu}_i)\hat{k}_{ii}\right]_{t-1} \\ &= \hat{\underline{\alpha}}_{t-1} + \hat{\Lambda}^{-1}_{t-1}Z'(\underline{y} - \hat{\underline{y}}_{t-1}), \end{aligned} \tag{5.3.12}$$

since $\partial\eta_i / \partial\mu_i = k_{ii}$. The result of equation (5.3.10) is precisely the result derived for principal component estimation of the the logit model in equation (3.12.4).

For a third example, consider the unit gamma $Y_i \sim \Gamma(1, \lambda_i)$ where $r_i = 1$ is a known nuisance parameter. Thus

$$g(\mu_i) = \theta_i = -\mu_i^{-1} = -\lambda_i = \underline{x}'_i\underline{\beta} = \underline{z}'_i\underline{\alpha} = \eta_i$$

and $\quad \mu_i = E(Y_i) = -\eta_i^{-1} = h(\eta_i).$

Hence $h'(\eta_i) = \eta_i^{-2}$ and $k_{ii}^{-1} = \eta_i^{-2}$ giving $y_i^* = \eta_i + [y_i + (\eta_i^{-1})]\,\eta_i^2 = 2\eta_i + y_i\,\eta_i^2$ evaluated at $\hat{\underline{\alpha}}_{t-1}$. The iterative equation is then given by

$$\hat{\underline{\alpha}}_t = [\hat{\Lambda}^{-1}Z'\hat{K}^{-1}\underline{y}^*]_{t-1}.$$

Lastly consider a Poisson response. That is $Y_i \sim \text{Poisson}(\lambda_i)$. $\mu_i = \lambda_i = e^{\eta_i} = k_{ii}^{-1} = h'(\eta_i)$. $y_i^* = (y_i / e^{\eta_i}) - 1 + \eta_i$. Thus

$$\hat{\underline{\alpha}}_t = [\hat{\Lambda}^{-1}Z'\hat{K}^{-1}\underline{y}^*]_{t-1}.$$

5.4 AN ALTERNATE PRINCIPAL COMPONENT ESTIMATOR IN THE GLM

Extending Schaefer's (1986) one step logistic principal component estimator to the generalized linear model, consider the maximum likelihood estimator in equation (2.4.7).

$$\hat{\underline{\beta}}_t = \left[(X'\hat{K}^{-1}X)^{-1}X'\hat{K}^{-1}\underline{y}^*\right]_{t-1},$$

where $\Phi = X'K^{-1}X$, $K^{-1} = \text{diag}\{k_{ii}^1\}$, and $k_{ii}^1 = [h'(\eta_i)]^2 / \text{Var}(Y_i)$. K^{-1} and $\underline{y}^*$ are re-estimated at each step. Thus, if the initial estimate is at the origin ($\underline{0}$), then $\underline{\beta}_{ML}$ can be expressed as

$$\hat{\underline{\beta}}_{ML} = \sum_{t=1}^{L}(X'\hat{K}_{t-1}^{-1}X)^{-1}X'\hat{K}_{t-1}^{-1}\left[\sum_{i=1}^{N}\underline{x}_i\,\hat{k}_{ii}^{-1}(y_i - \hat{\mu}_i)\frac{\partial\eta_i}{\partial\mu_i}\right]_{t-1},$$

where L is the iteration of convergence. Define

$$(X'K_t^{-1}X)^+ = \sum_{j=r}^{p} \lambda_{jt}^{-1} m_{jt} m'_{jt},$$

$$\text{where } (X'K_t^{-1}X)^{-1} = \sum_{j=0}^{p} \lambda_{jt}^{-1} m_{jt} m'_{jt}.$$

r is the number of components deleted.

In circumstances, such as the logistic model, when K_t is estimated well by the maximum likelihood estimate K_{ML}, then $\hat{\beta}_{pc}$ can be estimated by the one step solution

$$\hat{\beta}^*_{pc} = (X'\hat{K}_{ML}^{-1}X)^+ (X'\hat{K}_{ML}^{-1}X)\hat{\beta}_{ML} \tag{5.4.1}$$

when $(X'\hat{K}_t^{-1}X) \cong (X'\hat{K}_{ML}^{-1}X)$ and $(X'\hat{K}_t^{-1}X)^+ \cong (X'\hat{K}_{ML}^{-1}X)^+$. The development of $\hat{\beta}^*_{pc}$ follows naturally from equation (3.11.6). Note the following first order Taylor series approximation of a diagonal element of K^{-1}, $\hat{k}^{-1}$, about the true corresponding η. Let $\hat{\text{Var}}(Y) = q(\hat{\eta})$.

$$\hat{k}^{-1} \cong \frac{[h'_o(\eta)]^2}{q_o(\eta)} + \frac{2q_o(\eta)h''_o(\eta)h'_o(\eta) - (h'_o(\eta))^2 q'_o(\eta)}{(q_o(\eta))^2}(\hat{\eta} - \eta).$$

Therefore,

$$\text{Var}(\hat{k}^{-1}) \cong \frac{[2q_o(\eta)h''_o(\eta)h'_o(\eta) - (h'_o(\eta))^2 q'_o(\eta)]^2}{(q_o(\eta))^4} \sum_{j=0}^{p} z_j^2 \hat{\lambda}_j^{-1} = K^*(\eta_o) \sum_{j=0}^{p} z_j^2 \hat{\lambda}_j^{-1} \tag{5.4.2}$$

from equation (3.8.1). The subscript i is suppressed in equation (5.4.2). K^* is a constant. The variance, given in equation (5.4.2), will not be as affected for observations in the original data set as for the observations outside the mainstream of collinearity since these points, in general, do not deviate much in the z_j direction corresponding to "small" λ_j. This suggests K_{ML} will estimate K relatively well for the original data.

The asymptotic variance and bias of $\underline{\beta}^*_{pc}$ is equivalent to that for the iterative method given in equations (5.3.7) and (5.3.8). That is asymptotically,

$$\begin{aligned} \mathrm{Var}(\underline{\beta}^*_{pc}) &= (X'K^{-1}X)^+ X'K^{-1}X(X'K^{-1}X)^+ \\ &= M_s\Lambda_s^{-1}M'_s. \end{aligned} \tag{5.4.3}$$

The asymptotic distribution of $\underline{\beta}^*_{pc}$ is identical to equation (5.5.9).

5.5 INFERENCES CONCERNING THE PRINCIPAL COMPONENTS

The log-likelihood function follows directly from equation (2.2.2),

$$l = [\, yb(\theta) + c(\theta)] \,/\, q(\phi) + d(y, \phi). \tag{5.5.1}$$

The inferences regarding the principal components for GLM follows directly from section 2.5. For the principal component generalized linear model, define the score with respect to α_j to be

$$U_j^* = \frac{\partial l}{\partial \alpha_j} \qquad j = 0, 1, \ldots, p. \tag{5.5.2}$$

In obtaining the principal component maximum likelihood parameter estimates, $\underline{U}^* = (U_0^*, U_1^*, \ldots, U_p^*)'$ is set to zero, where

$$E(\underline{U}^*) = \underline{0} \quad \text{and} \quad E(\underline{U}^*\underline{U}^{*\prime}) = \Phi = \Lambda\,. \tag{5.5.3}$$

By an extension of the Central Limit Theorem (Feller (1966)), the asymptotic distribution of $\underline{U}^*$ is multivariate $N(\underline{0}, \Phi = \Lambda)$; hence

$$\underline{U}^{*\prime}\Phi^{-1}\underline{U}^* = \underline{U}^{*\prime}\Lambda^{-1}\underline{U}^* \sim \chi^2_{p+1,\,0}. \tag{5.5.4}$$

When convergence is obtained using the iterative equation (5.3.4), consider the unique maximum likelihood estimate, $\hat{\underline{\alpha}}$. Define $\underline{\alpha}$ to be the true parameter vector. The Taylor series expansion of $U^*(\underline{\alpha})$ about $\hat{\underline{\alpha}}$ (Dobson (1983)) is

$$U^*(\underline{\alpha}) \cong U^*(\hat{\underline{\alpha}}) + H^*(\hat{\underline{\alpha}})(\underline{\alpha} - \hat{\underline{\alpha}}), \tag{5.5.5}$$

where H^* is the Hessian matrix evaluated at the maximum likelihood estimates, $\hat{\underline{\alpha}}$. Thus,

$$U^*(\underline{\alpha}) \cong U^*(\hat{\underline{\alpha}}) - \Lambda(\underline{\alpha} - \hat{\underline{\alpha}}), \tag{5.5.6}$$

since $\Phi = \Lambda = E(-H^*)$. This implies that

$$(\hat{\underline{\alpha}} - \underline{\alpha}) \cong \Lambda^{-1}\underline{U}^*, \tag{5.5.7}$$

since $U(\hat{\underline{\alpha}}) = 0$ by definition. By taking expectations of both sides of equation (5.5.7),

$$E(\hat{\underline{\alpha}}) = \underline{\alpha} \quad \text{asymptotically,}$$

since $E(\underline{U}^*) = 0$. Similarly

$$E[(\hat{\underline{\alpha}} - \underline{\alpha})(\hat{\underline{\alpha}} - \underline{\alpha})'] = \Lambda^{-1}E(\underline{U}^*\underline{U}^{*\prime})\Lambda^{-1} = \Lambda^{-1},$$

for Λ nonsingular. Thus for large samples

$$\begin{aligned} \hat{\underline{\alpha}} &\dot{\sim} N(\underline{\alpha}, \Lambda^{-1}) \\ (\hat{\underline{\alpha}} - \underline{\alpha})'\Lambda(\hat{\underline{\alpha}} - \underline{\alpha}) &\dot{\sim} \chi^2_{p+1,0}. \end{aligned} \tag{5.5.8}$$

It follows that

$$\underline{b}_s^{pc} \dot{\sim} N(M_s\underline{a}_s, M_s\Lambda_s^{-1}M'_s). \tag{5.5.9}$$

5.6 HYPOTHESIS TESTING AND DELETION OF COMPONENTS

There is an assortment of rules for choosing the proper principal components to delete. A selection is given in Lee's dissertation (1986). Perhaps the most common rule is the one which deletes the principal components associated with the smallest eigenvalues. This method can be criticized since it does not take into account any of the Y data information. The stepwise method mentioned in section 3.15 using a t-like statistic takes into account the slope of the data in the direction of the principal component in question. Moreover, for the generalized linear model, a χ^2 statistic will also be suggested to determine the goodness-of-fit based on a subset of components.

The theory from section 2.6 naturally extends to the principal components. Consider the principal components, $Z = XM$, where M is the orthogonal matrix that diagonalizes the information matrix. Recall the overspecified or maximal model which has as many parameters as the N observations. Thus the maximal model has the parameter vector

$$\underset{\sim}{\alpha}_{\max} = [\alpha_1, \alpha_2, \ldots, \alpha_N]'.$$

To determine whether another model with $(p+1 < N)$ parameters $\underset{\sim}{\alpha} = [\alpha_0, \alpha_1, \ldots, \alpha_p]'$ is adequate relative to the maximal model, compare their likelihood functions (in keeping $p+1$ as small as possible). If $L(\underset{\sim}{\alpha}; \underset{\sim}{y}) \cong L(\underset{\sim}{\alpha}_{\max}; \underset{\sim}{y})$, then the model describes the data well. However, if $L(\underset{\sim}{\alpha}; \underset{\sim}{y}) << L(\underset{\sim}{\alpha}_{\max}; \underset{\sim}{y})$, then the model is poor relative to the maximal model. This suggests the likelihood ratio test using the statistic

$$\begin{aligned} \lambda &= L(\hat{\underset{\sim}{\alpha}}_{\max}; \underset{\sim}{y}) \,/\, L(\hat{\underset{\sim}{\alpha}}; \underset{\sim}{y}) \\ \text{or} \quad \ln \lambda &= l(\hat{\underset{\sim}{\alpha}}_{\max}; \underset{\sim}{y}) - l(\hat{\underset{\sim}{\alpha}}; \underset{\sim}{y}) \end{aligned} \tag{5.6.1}$$

If λ is large, then claim $\underset{\sim}{\alpha}$ is a poor model.

The sampling distribution of $\ln \lambda$ can be approximated by the following Taylor series expansion of $l(\underline{\alpha}; \underline{y})$ about the maximum likelihood estimator $\hat{\underline{\alpha}}$.

$$l(\underline{\alpha}; \underline{y}) \cong l(\hat{\underline{\alpha}}; \underline{y}) + (\underline{\alpha} - \hat{\underline{\alpha}})' U^*(\hat{\underline{\alpha}}) + (1/2)(\underline{\alpha} - \hat{\underline{\alpha}})' H(\hat{\underline{\alpha}}) (\underline{\alpha} - \hat{\underline{\alpha}}), \tag{5.6.2}$$

where $H(\hat{\underline{\alpha}})$ is the Hessian matrix evaluated at the maximum likelihood estimate. Recall that $U^*(\hat{\underline{\alpha}}) = \underline{0}$ by definition and $\Lambda = \Phi = -E(H)$ for large samples. Thus equation (5.6.2) can be rewritten as

$$2[l(\hat{\underline{\alpha}}; \underline{y}) - l(\underline{\alpha}; \underline{y})] = (\underline{\alpha} - \hat{\underline{\alpha}})' \Lambda (\underline{\alpha} - \hat{\underline{\alpha}}) \dot{\sim} \chi^2_{p+1,0}, \tag{5.6.3}$$

from (5.5.8).

The counterpart of the scaled deviance is

$$S = 2 \ln \lambda = 2[l(\hat{\underline{\alpha}}_{\max}; \underline{y}) - l(\hat{\underline{\alpha}}; \underline{y})]. \tag{5.6.4}$$

The scaled deviance can be broken down into the following components

$$\begin{aligned} S = 2\{&[l(\hat{\underline{\alpha}}_{\max}; \underline{y}) - l(\underline{\alpha}_{\max}; \underline{y})] - [l(\hat{\underline{\alpha}}; \underline{y}) - l(\underline{\alpha}; \underline{y})] + [l(\underline{\alpha}_{\max}; \underline{y}) - l(\underline{\alpha}; \underline{y})]\} \\ &\dot{\sim} \chi^2_{N-p-1,0}, \end{aligned} \tag{5.6.5}$$

when $l(\underline{\alpha}_{\max}; \underline{y}) \cong l(\underline{\alpha}; \underline{y})$; otherwise, equation (5.6.5) has an asymptotic noncentral χ^2 distribution.

In deciding which principal components should be deleted, perhaps the most useful hypothesis test is of the form

$$\begin{aligned} H_0&: \underline{\alpha} = \underline{\alpha}_0 \quad (q+1) \\ H_1&: \underline{\alpha} = \underline{\alpha}_1 \quad (p+1), \end{aligned} \tag{5.6.6}$$

where $q < p < N$ and H_0 is nested in H_1. H_0 is tested against the alternative by using the difference in the log-likelihood statistics,

$$S^* = S_0 - S_1 = 2\,[l(\hat{\underline{\alpha}}_1; \underline{y}) - l(\hat{\underline{\alpha}}_0; \underline{y})]. \tag{5.6.7}$$

If both H_0 and H_1 describe the data adequately relative to the maximal model, then

$$\begin{aligned} & S_0 \dot{\sim} \chi^2_{N-q-1,0} \\ \text{and} \quad & S_1 \dot{\sim} \chi^2_{N-p-1,0}\,. \\ \text{Thus} \quad & S^* \dot{\sim} \chi^2_{p-q,0}\,. \end{aligned} \tag{5.6.8}$$

Notice that if $q + 1 = p$, then $S^* \dot{\sim} \chi^2_{1,0}$.

Consider the example when $Y_i \sim \text{binomial}(n, \pi_i)$. To test the hypothesis in equation (5.6.6), use

$$S^* = 2\left\{\sum_{i=1}^{N} y_i(\underline{z}'_{i1}\hat{\underline{\alpha}}_1 - \underline{z}'_{i0}\hat{\underline{\alpha}}_0) + n\sum_{i=1}^{N} [\ln(1 + e^{\underline{z}'_{i1}\hat{\underline{\alpha}}_1}) - \ln(1 + e^{\underline{z}'_{i0}\hat{\underline{\alpha}}_0})]\right\},$$

which has an asymptotic χ^2_{p-q} distribution under H_o .

For normally distributed data with unknown variance, the likelihood ratio test can be put into the form of a F test.

$$[SSE_o - SSE_1] \,/\, [(p-q)\hat{\sigma}^2_1] \sim F_{p-q,N-p,0},$$

where the full and reduced models use the least squares estimates.

Perhaps a more common test in practice would be the one of the form

$$H_o\colon C\underline{\alpha} = \underline{0}, \tag{5.6.9}$$

where C is a $q \times (p+1)$ matrix of constants. In particular, the test for the deletion of a single principal components would yield the choice of $C = (0,...,0, 1, 0, ...,0)$. In the case where all the

principal components are kept in the canonical form model, $\hat{\underline{\alpha}}$'s are indeed maximum likelihood estimates. Relying heavily on this fact, $\hat{\underline{\alpha}}$ has an asymptotic limiting normal distribution

$$\hat{\underline{\alpha}} \dot{\sim} N(\underline{\alpha}, \Lambda^{-1}). \tag{5.6.10}$$

It follows under H_o,

$$\hat{\underline{\alpha}}' C'(C\Lambda^{-1}C')^{-1}C\hat{\underline{\alpha}} \dot{\sim} \chi^2_q. \tag{5.6.12}$$

Hence, the test for a single component simplifies to

$$\hat{\alpha}_j^2 \lambda_j \dot{\sim} \chi^2_1. \tag{5.6.13}$$

The above statistic is compared to the appropriate percentage point of the asymptotic chi-square distribution. Of course λ_i is usually unknown; therefore the test

$$t_i^* = \hat{\alpha}_i \hat{\lambda}_i^{1/2} \tag{5.6.14}$$

is a common test for a single component using $N - p - 1$ degrees of freedom for the t-distribution.

Jolliffe (1986) develops several strategies for the selection of components in principal component standard multiple regression. One such strategy is to simply delete all the components associated with small eigenvalues below a specified cutoff. A useful upper limit in practice is between .01 and .1. This procedure is certainly useful in the GLM.

A different approach from deleting small eigenvalues is one which incorporates the t-test given in equation (5.6.14). Hence a procedure could be used which deletes components based on its contribution to the regression via a t-test. However, Jolliffe warns, for standard PC regression, that usually more components will be retained than are really necessary if components are deleted in succession until a significant t -statistic is reached.

A natural extension considers the VIF's developed in section 4.4. Delete components successively until all the VIF's are below a specified value. Recall that $\text{VIF}_i = (1 - R_i^2)^{-1}$, where R_i^2 is the coefficient of determination for the regression of the standardized $\hat{K}^{-1/2}\underset{\sim}{X}_i$ on $\hat{K}^{-1/2}X_{-i}$. Values of $R_i^2 > .90$ yield $\text{VIF}_i > 10$ whereas values of $R_i^2 > .75$ yield $\text{VIF}_i > 4$. In standard multiple regression, Jolliffe points out that although this procedure appears to be more sophisticated, it is almost as arbitrary as the eigenvalue cutoff value given above.

Hill et al. (1977) considers a more sophisticated approach to deletion of components. The weak criterion is one where the objective is to get $\underset{\sim}{b}_z^{pc}$ close to $\underset{\sim}{\beta}$. That is $\underset{\sim}{b}_z^{pc}$ is preferred over $\hat{\underset{\sim}{\beta}}$ if

$$\text{tr}[MSE(\underset{\sim}{b}_z^{pc})] \leq \text{tr}[MSE(\hat{\underset{\sim}{\beta}})], \tag{5.6.15}$$

where $MSE(\underset{\sim}{b}_z^{pc}) = E[(\underset{\sim}{b}_z^{pc} - \underset{\sim}{\beta})(\underset{\sim}{b}_z^{pc} - \underset{\sim}{\beta})']$. Notice that equation (5.6.15) is equivalent to

$$\|\underset{\sim}{b}_z^{pc} - \underset{\sim}{\beta}\| \leq \|\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}\|.$$

A stronger criterion is more oriented toward prediction of $g(y)$ rather than estimation of the coefficients. The requirement is now

$$MSE(\underset{\sim}{c}'\underset{\sim}{b}_z^{pc}) \leq MSE(\underset{\sim}{c}'\hat{\underset{\sim}{\beta}}),$$

for all nonnull $\underset{\sim}{c}$ of proper dimension. Notice for $\underset{\sim}{c}'$ in the X space of interest, this is a prediction oriented criterion.

5.7 A VARIETY OF APPLICATIONS OF PCA TO BINARY RESPONSES

Consider the class of generalized linear models where the outcome is binary in nature. Suppose that at each of N various combinations of the covariates, there are n_i binary responses. Define

$$Y_{ij} = \begin{cases} 1 & \text{if the outcome j at covariate combination i is a success} \\ 0 & \text{otherwise} \end{cases}$$

and π_i is the probability of success at covariate combination i. Thus $Y_{i.} \sim$ binomial(n_i, π_i) and is a member of the exponential family as given in equation (3.2.1).

The proportion of success is then given by $\hat{p}_i = Y_{i.} / n_i$ for $i = 1, 2, ..., N$. For $n_i\pi_i$ sufficiently large

$$\hat{p}_i \dot{\sim} N(\pi_i, \pi_i(1-\pi_i) / n_i) \qquad (5.7.1)$$

To model π_i as a function of the continuous covariates, as in logistic regression, recall the generalized linear model

$$g(\pi_i) = \underline{x}'_i\underline{\beta},$$

as in equation (2.1.3). g is the link function between the mean, $n_i\pi_i$, and the systematic component, $\underline{x}'_i\underline{\beta}$. Perhaps the most obvious link is the linear probability model of the form

$$\pi_i = \underline{x}'_i\underline{\beta} = \underline{z}'_i\underline{\alpha}. \qquad (5.7.2)$$

Despite the attractiveness of equation (5.7.2) which allows the assumption of the additive error term with normality, the linear probability model has some serious drawbacks such as predicted probabilities falling outside the unit interval.

Another model proposed to link π to $\underline{x}'\underline{\beta}$ is the angular model. Let

$$\pi = \sin^2(\underline{x}'\underline{\beta}).$$

To ensure the predicted probability is contained in the unit interval, a cumulative probability distribution is often modelled (Dobson (1983)). Consider

$$\pi = g^{-1}(\underline{x}'\underline{\beta}) = h(\eta) = \int_{-\infty}^{t} f(v)\, dv \ , \tag{5.7.3}$$

where $f(v)$ is a probability density function and $0 < \pi < 1$. t is related to $\eta_i = \underline{x}'_i\underline{\beta} = \underline{z}'_i\underline{\alpha}$, with all the $(p+1)$ principal components. Notice π is a nondecreasing function of t. Table 7 contains various cumulative probability distributions used to model π. Note that the Probit model has particular use for the median lethal dose (LD50) when $\mu = t$ (see Finney (1971)).

Recall the principal component maximum likelihood iterative scheme as given in section 5.3.

$$\hat{\pi}_s^{pc} = h_s(\underline{x}'_i\underline{b}_s^{pc}) \ ,$$

where $\underline{b}_s^{pc} = M_s\hat{\underline{\alpha}}_s^{pc}$ are the maximum likelihood estimates using s principal components.

Notice that the log-likelihood equation given in equation (2.2.2) can also be written as

$$\begin{aligned} l(\underline{\pi}, \underline{y}) &= \ln \prod_{i=1}^{N} \binom{n_i}{Y_{i\cdot}} \pi_i^{Y_{i\cdot}} (1-\pi_i)^{n_i - Y_{i\cdot}} \\ &= \sum_{i=1}^{N} \left[Y_{i\cdot} \ln(\pi_i) + (n_i - Y_{i\cdot}) \ln(1-\pi_i) + \ln\binom{n_i}{Y_{i\cdot}} \right]. \end{aligned} \tag{5.7.4}$$

Since, the maximal model as defined in section 5.6 has as many parameters as observations, equation (5.7.4) above can be maximized with respect to π_i as well as to α_i for $i = 1, 2, \ldots, N$. Thus

Table 7. VARIOUS MODELS FOR BINARY RESPONSES

Model	Density
Logistic	$f(v) = \dfrac{e^{(v-\mu)/k}}{k[1+e^{(v-\mu)/k}]^2}$
Probit	$f(v) = \dfrac{\exp\{(-2\sigma^2)^{-1}(v-\mu)^2\}}{\sigma\sqrt{2\pi}}$
Linear	$f(v) = (b-a)^{-1},\ b > a$
Extreme	$f(v) = \beta^{-1}\exp[\beta^{-1}(v-\alpha) - \exp(\beta^{-1}(v-\alpha))]$

Model	Probability $= \int_{v=-\infty}^{t} f(v)dv$	Link to η
Logistic	$\pi = [1+\exp(-(t-\mu))/k]^{-1}$	$\ln\left(\dfrac{\pi}{1-\pi}\right) = \eta = \dfrac{t-\mu}{k}$
Probit	$\pi = \Phi\left(\dfrac{t-\mu}{\sigma}\right)$	$\Phi^{-1}(\pi) = \eta = \dfrac{t-\mu}{\sigma}$
Linear	$\pi = \dfrac{t-a}{b-a},\ a \le t \le b$	$\pi = \eta = \dfrac{t-a}{b-a}$
Extreme	$\pi = 1 - \exp[-\exp(\beta^{-1}(t-\alpha))]$	$\ln(-\ln(1-\pi)) = \eta$ $= \beta^{-1}(t-\alpha)$

$$0 = \frac{\partial l}{\partial \pi_i} = \frac{Y_{i.}}{\pi_i} - \frac{n_i - Y_{i.}}{1 - \pi_i} \tag{5.7.5}$$

and $\hat{\pi}_{i,max} = Y_{i.} / n_i$. It follows then that

$$\hat{\pi}_{i,p}^{pc} = h_s(x'_i b_s^{pc}), \tag{5.7.6}$$

where p is the number of continuous covariates to model π in the nonmaximal model.

The scaled deviance can be thought of as

$$\begin{aligned} S &= 2\,[l(\hat{\pi}_{\max}; Y_{i.}) - l(\hat{\pi}_{p,s}^{pc}; Y_{i.})] \\ &= 2\sum_{i=1}^{N}\left[Y_{i.} \ln \frac{Y_{i.}}{n_i h_s(x'_i b_s^{pc})} + (n_i - Y_{i.}) \ln \frac{n_i - Y_{i.}}{n_i - n_i h_s(x'_i b_s^{pc})} \right] \\ &= 2\sum_{i=1}^{2N} o_i \ln (o_i / e_i) \\ &\dot{\sim} \chi^2_{N-p-1}, \end{aligned} \tag{5.7.7}$$

where o_i are the observed frequencies and e_i are the expected frequencies of the $2N$ cells given in Table 8.

Table 8. N INDEPENDENT BINOMIAL RANDOM VARIABLES

	Binomial Trials					
	1	2	-	-	-	N
No. Successes	$Y_{1\bullet}$	$Y_{2\bullet}$	-	-	-	Y_N
No. Failures	$n_1 - Y_{1\bullet}$	$n_2 - Y_{2\bullet}$	-	-	-	$n_N - Y_{N\bullet}$
No. Trials	n_1	n_2				n_N

Chapter VI

RIDGE ESTIMATORS IN THE GLM

6.1 INTRODUCTION

Schaefer (1979) has developed a ridge estimator for logistic regression when an alternate estimation technique is desired. See section 3.20. The idea of a ridge estimator can be extended to the GLM. Recall section 5.1 which discusses several effects of an ill-conditioned information matrix to the GLM. From equation (2.4.7), $\hat{\underline{\beta}}$ is an iterative reweighted least squares (IWLS) estimate of β. Walker and Duncan (1967) demonstrate for the logit model in equation (3.4.1) that $\hat{\underline{\beta}}$ minimizes the weighted sum of squares error ($WSSE$) and thus is the best estimator based on $WSSE$ criterion. However, $\|\hat{\underline{\beta}}\|$ may be too long on the average. Recall

$$\begin{aligned} \text{Var}(\hat{\underline{\beta}}) &= E\{(\hat{\underline{\beta}} - E(\hat{\underline{\beta}}))(\hat{\underline{\beta}} - E(\hat{\underline{\beta}}))'\} \\ &\cong (X'K^{-1}X)^{-1} = \Phi^{-1}. \end{aligned} \tag{6.1.1}$$

It follows then that

$$\sum_{i=0}^{p} \mathrm{Var}(\hat{\beta}_i) \cong \mathrm{tr}(X'K^{-1}X)^{-1} = \sum_{i=0}^{p} \lambda_i^{-1}. \tag{6.1.2}$$

Equation (6.1.2) requires switching expected value with trace. If the information matrix Φ for $\hat{\underline{\beta}}$ is near singular, based on a condition index (equation (4.3.1)), then the norm of the estimated parameter vector maybe too long.

6.2 RIDGE ESTIMATORS IN THE GLM

Clearly, an alternate estimator, $\underline{\beta}^R$, for $\underline{\beta}$ should have a norm smaller than that of the maximum likelihood estimator, $\hat{\underline{\beta}}$. On the other hand, the trivial estimator $\underline{\beta}^R = \underline{0}$ is not acceptable for obvious reasons. Thus $\underline{\beta}^R$ should be reasonably close to $\hat{\underline{\beta}}$. Define closeness in terms similar to Hoerl and Kennard (1970) and Schaefer (1979) as

$$WSSE(\underline{\beta}^R) = WSSE(\hat{\underline{\beta}}) + \delta\,, \tag{6.2.1}$$

for $\delta > 0$. Solving for δ in equation (6.2.1),

$$\begin{aligned} \delta &= WSSE(\underline{\beta}^R) - WSSE(\hat{\underline{\beta}}) \\ &= (\underline{y} - h^R(\underline{\eta}))'T_o^{-1}(\underline{y} - h^R(\underline{\eta})) - (\underline{y} - h(\hat{\underline{\eta}}))'T_o^{-1}(\underline{y} - h(\hat{\underline{\eta}})) \\ &= (h(\hat{\underline{\eta}}) - h^R(\underline{\eta}))'T_o^{-1}(h(\hat{\underline{\eta}}) - h^R(\underline{\eta})) + 2(h(\hat{\underline{\eta}}) - h^R(\underline{\eta}))'T_o^{-1}(\underline{y} - h(\hat{\underline{\eta}})), \end{aligned} \tag{6,2.2}$$

where $T_o^{-1} = \mathrm{diag}\{(k_{ii}^{-1} / [h'(\eta_i)]^2)_o\} = \mathrm{diag}\{1 / \mathrm{Var}(Y_i)\}$.

In requiring that the ridge estimator, $\underline{\beta}^R$, is consistent for $\underline{\beta}$ ($\hat{\underline{\beta}}$ is already consistent), approximate $h(\hat{\underline{\eta}})$ and $h^R(\underline{\eta})$ as follows using the first order Taylor series expansion about $\underline{\beta}$.

$$h(\hat{\underline{\eta}}) = h_o(\underline{\eta}) + D_o X(\hat{\underline{\beta}} - \underline{\beta})$$
$$\text{and} \quad h^R(\underline{\eta}) = h_o(\underline{\eta}) + D_o X(\underline{\beta}^R - \underline{\beta}) \tag{6.2.3}$$

for large N, where $D_o = \text{diag}\{h'_o(\eta)\}$ evaluated at $\underline{\beta}$. Equation (6.2.2) can now be re-expressed as

$$\begin{aligned}\delta &\cong (\hat{\underline{\beta}} - \underline{\beta}^R)' X' D_o T_o^{-1} D_o X(\hat{\underline{\beta}} - \underline{\beta}^R) + 2(\hat{\underline{\beta}} - \underline{\beta}^R)' X' D_o T_o^{-1}(\underline{y} - h(\hat{\underline{\eta}})) \\ &= (\hat{\underline{\beta}} - \underline{\beta}^R)' \Phi (\hat{\underline{\beta}} - \underline{\beta}^R),\end{aligned} \tag{6.2.4}$$

since $X' D_o T_o^{-1}(\underline{y} - h(\hat{\underline{\eta}})) = \underline{0}$ are the analog of the ML "normal" equations given in equation (2.4.2) for the generalized linear model. $X' K^{-1} X = \Phi$ is the information matrix.

Thus the GLM ridge estimator, $\underline{\beta}^R(\delta)$, is the estimator that has a minimum norm for δ fixed. Notice the similarity to Schaefer's (1979) work outlined in section 3.20. Consider the Lagrange minimization of

$$Q = (\underline{\beta}^R)' \underline{\beta}^R + \{d^{-1} [(\hat{\underline{\beta}} - \underline{\beta}^R)' \Phi (\hat{\underline{\beta}} - \underline{\beta}^R) - \delta]\}, \tag{6.2.5}$$

where d^{-1} is the Lagrange multiplier. The solution of equation (6.2.5) follows as

$$\frac{\partial Q}{\partial \underline{\beta}^R} = 0 \quad \text{implies} \quad \underline{\beta}^R(\delta) = (\Phi + dI)^{-1} \Phi \hat{\underline{\beta}}. \tag{6.2.6}$$

Letting $\Phi_d = X' K^{-1} X + dI$, the connection between d and δ is

$$\begin{aligned}\delta &= (\hat{\underline{\beta}} - \underline{\beta}^R)' \Phi (\hat{\underline{\beta}} - \underline{\beta}^R) \\ &= \hat{\underline{\beta}}' (I - \Phi \Phi_d^{-1}) \Phi (I - \Phi_d^{-1} \Phi) \hat{\underline{\beta}} \\ &= d^2 \hat{\underline{\beta}}' (\Phi + dI)^{-1} \Phi (\Phi + dI)^{-1} \hat{\underline{\beta}},\end{aligned} \tag{6.2.7}$$

since $d(\Phi + dI)^{-1} = I - (\Phi + dI)^{-1} \Phi$. Thus $\underline{\beta}^R$ is a function of d and can be expressed as

$$\underline{\beta}^R(d) = (\Phi + dI)^{-1}\,\Phi\,\hat{\underline{\beta}} \tag{6.2.8}$$

The asymptotic variance of $\underline{\beta}^R(d)$ is

$$\mathrm{Var}(\underline{\beta}^R(d)) = \Phi_d^{-1}\,\Phi\,\Phi_d^{-1}. \tag{6.2.9}$$

The corresponding bias can be quantified as

$$\mathrm{Bias}(\underline{\beta}^R(d)) = -\,d\Phi_d^{-1}\,\underline{\beta}. \tag{6.2.10}$$

The asymptotic distribution of $\underline{\beta}^R(d)$ is

$$\underline{\beta}^R(d) \dot{\sim}\ N(\Phi_d^{-1}\,\Phi\,\underline{\beta},\ \Phi_d^{-1}\,\Phi\,\Phi_d^{-1}) \tag{6.2.11}$$

The logit link with Bernoulli data yield results given in section 3.20. The examples below illustrate the generalization of Schaefer's (1979) result.

Consider the example when $y_i \sim N(\mu_i, \sigma^2)$. $\Phi = \sigma^{-2}X'X$ and

$$\begin{aligned}\underline{\beta}^R(d) &= (\sigma^{-2}X'X + d^*I)^{-1}\sigma^{-2}X'X\hat{\underline{\beta}}_{OLS}\\ &= (\sigma^{-2}X'X + \sigma^{-2}dI)^{-1}\sigma^{-2}X'\underline{y}\\ &= (X'X + dI)^{-1}X'\underline{y},\end{aligned}$$

which is precisely the ridge estimator given by Hoerl and Kennard (1970) when $d^* = d\sigma^{-2}$.

As another example, consider $Y_i \sim \mathrm{Poisson}(\lambda_i)$. $\Phi_p = X'K^{-1}X$ where $K^{-1} = \mathrm{diag}\{e^{\eta_i}\}$. Thus $\underline{\beta}^R(d) = (\hat{\Phi}_p + dI)^{-1}\,\hat{\Phi}_p\hat{\underline{\beta}}_{ML}$.

6.3 METHODS OF CHOOSING THE SHRINKAGE PARAMETER d

At present, there exist scores of methods to choose d for standard ridge multiple regression, assuming normal response and the identity link. In Schaefer's 1979 dissertation, the results for ridge multiple regression are relied on heavily in developing methods for choosing d for a Bernoulli response and a logit link function. Schaefer (1979) presents three methods of choosing d.

$$\begin{aligned} d_1 &= (\hat{\underline{\beta}}'_{ML}\hat{\underline{\beta}}_{ML})^{-1} \\ d_2 &= \{\max_i |\hat{\alpha}_i|\}^{-2} \\ d_3 &= (p+1)d_1. \end{aligned} \tag{6.3.1}$$

Notice that $d_1 < d_2 < d_3$. The value d_3 appears to represent the harmonic mean method of choosing the shrinkage parameter. Further, there is a similarity between d_2 and the maximum value of d for which the mean squared error (MSE) of the estimated coefficients in standard multiple regression (Tripp (1983)) is less than or equal to that of least squares. Schaefer admits that d_1 is considered as a possible candidate is mainly because of its ease in computation. In standard multiple ridge regression, the harmonic mean method is considered to be very conservative. Observe that d_1 and d_2 are even more conservative than d_3. It then is no surprise that, in his summary, Schaefer recommends d_3 as the best method of choosing a shrinkage parameter in the presence of an extremely ill-conditioned information matrix.

Schaefer's developments for a shrinkage parameter in logistic regression can quite naturally be extended into the ridge setting of generalized linear models. Various other techniques to choose d are available. Suggestions for choosing d based on trying to optimize the predictive capabilities of the generalized linear model are given. This dissertation considers the C_p criterion and the DF-trace criterion.

6.4 PREDICTION CRITERION FOR SHRINKAGE

In section 2.9, a C_p^* statistic was developed to assist in identifying a p parameter subset model, where $1 \leq p \leq k$ = maximal number of explanatory variables of interest. The candidate model chosen, using C_p^*, represented the model with a minimal blend of variance and bias of the predicted values, $\hat{y}$. This diagnostic has great implication for the GLM. Not only is it important to find an interval of d where there is improvement in the estimation of the parameters, but quite often the researcher wants good predictive capabilities. The notion of C_p^* can be developed into a prediction oriented method of choosing the shrinkage parameter, d, when using ridge regression in the GLM. An argument for a C_p^* as a method of choosing d will be outlined in this section.

Myers (1986) shows the development of the C_p statistic used as a prediction criterion for choice of d in usual ridge multiple regression. The statistic, C_p^R in this setting is given as

$$C_p^R = \frac{SSE_d}{\hat{\sigma}^2} - N + 2\,[1 + \mathrm{tr}(H_d)], \tag{6.4.1}$$

where SSE_d is the sum of squares error using the ridge parameter estimates. The matrix, $H_d = X(X'X + dI)^{-1}X'$, is the corresponding projection matrix or hat matrix in the ridge setting. The C_p^R given above uses the centered and scaled explanatory variables as the X matrix, ignoring the constant column of ones. The $1 + \mathrm{tr}(H_d)$ accounts for the constant term.

Recall that C_p^R denotes the C_p statistic for ridge regression. Consider the following development for a similar statistic in the GLM,

$$C_p^R = \sum_{i=1}^{N} \frac{\mathrm{Var}(\hat{y}_i) + \mathrm{Bias}^2(\hat{y}_i)}{\mathrm{Var}(Y_i)}. \tag{6.4.2}$$

First construct the variance portion.

$$\begin{aligned}
\sum_{i=1}^{N} \frac{\mathrm{Var}(\hat{y}_i)}{\mathrm{Var}(Y_i)} &= \sum_{i=1}^{N} \frac{1}{\mathrm{Var}(Y_i)} \mathrm{Var}[h(\eta_i^R)] \\
&\cong \sum_{i=1}^{N} \frac{1}{\mathrm{Var}(Y_i)} \mathrm{Var}[h(\eta_i) + h'(\eta_i)(\eta_i^R - \eta_i)] \\
&= \sum_{i=1}^{N} \frac{[h'(\eta_i)]^2}{\mathrm{Var}(Y_i)} \mathrm{Var}(\eta_i^R) \qquad (6.4.3) \\
&= \sum_{i=1}^{N} k_{ii}^{-1} \mathrm{Var}(\underline{x}_i' \Phi_d^{-1} X' K^{-1} X \hat{\underline{\beta}}) \\
&= \mathrm{tr}(K^{-1} X \Phi_d^{-1} \Phi \Phi^{-1} \Phi \Phi_d^{-1} X') \\
&= \mathrm{tr}(\Phi \Phi_d^{-1} \Phi \Phi_d^{-1}),
\end{aligned}$$

where $\Phi = X'K^{-1}X$ and $\Phi_d = X'K^{-1}X + dI$. Notice when $d = 0$ that $\sum_{i=1}^{N} \frac{\mathrm{Var}(\hat{y}_i)}{\mathrm{Var}(Y_i)} = p + 1$ which is completely consistent with ordinary least squares.

The bias portion of equation (6.4.2) is somewhat more difficult to develop. Define

$$\begin{aligned}
B^2 &= \sum_{i=1}^{N} \frac{\mathrm{Bias}^2(\hat{y}_i)}{\mathrm{Var}(Y_i)} \qquad (6.4.4) \\
&= (h(\underline{\eta}) - E[h(\underline{\eta}^R)])' T^{-1} (h(\underline{\eta}) - E[h(\underline{\eta}^R)]),
\end{aligned}$$

where $T^{-1} = \mathrm{diag}\{1 / \mathrm{Var}(Y_i)\}$. Consider the ridge counterpart of the quadratic form, χ^2, given in equation (2.8.2).

$$\begin{aligned}
\chi_R^2 &= \sum_{i=1}^{N} \frac{(y_i - \mu_i^R)^2}{\sigma_i^2} \qquad (6.4.5) \\
&= (\underline{y} - h(\underline{\eta}^R))' T^{-1} (\underline{y} - h(\underline{\eta}^R)).
\end{aligned}$$

Recall the following theorem from Graybill (1976). Let $\underline{W}$ be a $N \times 1$ random vector and let $E(\underline{W}) = \underline{\mu}$, $\mathrm{Cov}(\underline{W}) = \Sigma$. Then

$$E(\underline{W}' A \underline{W}) = \mathrm{tr}(A\Sigma) + \underline{\mu}' A \underline{\mu},$$

for A symmetric. Set

$$\begin{aligned}
\underline{W} &= \underline{Y} - h(\underline{\eta}^R) \\
\Sigma &= \mathrm{Var}[\underline{Y} - h(\underline{\eta}^R)] \\
A &= T^{-1} \\
\underline{\mu} &= h(\underline{\eta}) - E[h(\underline{\eta}^R)].
\end{aligned}$$

The expected value of χ^2_k follows.

$$\begin{aligned}
E(\chi^2_R) &= \mathrm{tr}(T^{-1}\ \mathrm{Var}(\underline{Y} - h(\underline{\eta}^R))) + B^2 \\
&= \mathrm{tr}(T^{-1}\ [T + \mathrm{Var}[h(\underline{\eta}^R)] - 2\mathrm{Cov}(\underline{Y}, h(\underline{\eta}^R))]) + B^2,
\end{aligned} \tag{6.4.6}$$

where

$$\begin{aligned}
\mathrm{Var}[h(\underline{\eta}^R)] &\cong \mathrm{Var}(h'(\underline{\eta})\underline{\eta}^R) \\
&= \mathrm{diag}\{[h'(\eta_i)]^2\}\mathrm{Var}(\underline{\eta}^R) \\
&= \mathrm{diag}\{[h'(\eta_i)]^2\}X\Phi_d^{-1}\Phi\Phi_d^{-1}X',
\end{aligned} \tag{6.4.7}$$

from the Taylor Series expansion above in equation (6.4.3). The covariance term is computed below.

$$\begin{aligned}
\mathrm{Cov}(\underline{Y},\ h(\underline{\eta}^R)) &\cong \mathrm{Cov}(\underline{Y},\ h(\underline{\eta}) + h'(\underline{\eta})(\underline{\eta}^R - \underline{\eta})) \\
&= \mathrm{Cov}(\underline{Y},\ h'(\underline{\eta})\underline{\eta}^R) \\
&= \mathrm{Cov}(\underline{Y},\ h'(\underline{\eta})X\Phi_d^{-1}X'K^{-1}\hat{\underline{\beta}}) \\
&= \mathrm{Cov}\Big(\underline{Y},\ h'(\underline{\eta})X\Phi_d^{-1}X'K^{-1}X\Big(\hat{\underline{\beta}}_{t-1} + (X'K^{-1}X)^{-1}\sum_{i=1}^{N} x_i k_{ii}^{-1}(y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}\Big)\Big) \\
&= \mathrm{Cov}\Big(\underline{Y},\ h'(\underline{\eta})X\Phi_d^{-1}X'K^{-1}X(X'K^{-1}X)^{-1}\sum_{i=1}^{N} x_i k_{ii}^{-1} y_i \frac{\partial \eta_i}{\partial \mu_i}\Big) \\
&= \mathrm{Cov}\Big(\underline{Y},\ h'(\underline{\eta})X\Phi_d^{-1}\sum_{i=1}^{N} x_i k_{ii}^{-1} y_i \frac{\partial \eta_i}{\partial \mu_i}\Big) \\
&= h'(\underline{\eta})X\Phi_d^{-1}\sum_{i=1}^{N} x_i\, k_{ii}^{-1}\frac{\partial \eta_i}{\partial \mu_i}\ T,
\end{aligned} \tag{6.4.8}$$

since $\mathrm{Cov}(A\underline{Y},\ B\underline{Y}) = AB\,\mathrm{Var}(\underline{Y})$. This implies

$$\mathrm{tr}(2T^{-1}\,\mathrm{Cov}(\underline{Y}, h(\underline{\eta}^R))) \cong 2\mathrm{tr}(\Phi\Phi_d^{-1}),$$

from equation (6.4.8). Thus in combining equations (6.4.7) and (6.4.8) into equation (6.4.6),

$$E(\chi_R^2) = N + \mathrm{tr}(\Phi\Phi_d^{-1}\Phi\Phi_d^{-1}) - 2\mathrm{tr}(\Phi\Phi_d^{-1}) + B^2. \tag{6.4.9}$$

Notice however that

$$\begin{aligned}\chi_R^2 &= \sum_{i=1}^{N} \frac{(y_i - \mu_i^R)^2}{\sigma_i^2} \\ &\cong \frac{D_{N,p;d}}{\hat{\phi}_{p,d=0}},\end{aligned} \tag{6.4.10}$$

from equation (2.9.2). $D_{N,p} / \hat{\phi}$ in GLM is a reasonable alternative for $SSE_p / \hat{\sigma}^2$ in ordinary least squares (see Pregibon (1979)). $D_{N,p}$ also lends itself to Aikaike's Information Criterion. Note that $D_{N,p;d}$ is the deviance comparing a p parameter model using the shrinkage parameter d to the maximal model. $\hat{\phi}_{p,d=0} = D_{N,p} / (N-p)$ is an estimate of the scale parameter, where D is the deviance of the p parameter model with shrinkage parameter of zero relative to the maximal model. Recall that χ^2 and $D_{N,p} / \hat{\phi}$ have the same limiting distribution. See equations (2.7.1) and (2.7.2).

Hence an estimate for $B^2 = \sum_{i=1}^{N} \frac{\mathrm{Bias}^2(\hat{y}_i)}{\mathrm{Var}(Y_i)}$ is given by

$$\begin{aligned}B^2 &\cong \chi_R^2 - N - \mathrm{tr}(\Phi\Phi_d^{-1}\,\Phi\,\Phi_d^{-1}) + 2\mathrm{tr}(\Phi\Phi_d^{-1}) \\ &\cong \frac{D_{N,p;d}}{\hat{\phi}_{p,d=0}} - N - \mathrm{tr}(\Phi\Phi_d^{-1}\,\Phi\,\Phi_d^{-1}) + 2\mathrm{tr}(\Phi\Phi_d^{-1}).\end{aligned} \tag{6.4.11}$$

Referring to the original motivation of C_p^R in equation (6.4.2)

$$C_p^R = \sum_{i=1}^{N} \frac{\mathrm{Var}(\hat{y}_i) + \mathrm{Bias}^2(\hat{y}_i)}{\mathrm{Var}(Y_i)}$$
$$\cong \frac{D_{N,p;d}}{\hat{\phi}_{p,d=0}} - N + 2\mathrm{tr}(\Phi\Phi_d^{-1}). \tag{6.4.12}$$

Notice that when normal response data is used with common variance, equation (6.4.12) simplifies to the ridge C_p^R in equation (6.4.1). Also outside of the ridge setting, the shrinkage parameter with $d = 0$, C_p^R in equation (6.4.12) is precisely Pregibon's $C_p^* = D_{N,p} / \hat{\phi} - N + 2(p+1)$ given in equation (2.9.2). Furthermore, again for normal response with the identity link function and common variance (without the ridge setting),

$$C_p^R = C_p^* = C_p, \tag{6.4.13}$$

when $d = 0$. C_p is the least squares Mallow's C_p.

Perhaps the most straight forward techniques to implement C_p^R as a diagnostic tool would be to plot C_p^R as a function of d. Choose d to minimize C_p^R. Such a choice will usually yield good quality of prediction in the generalized linear model. An example is given in section 6.6.

6.5 THE DF-TRACE METHOD FOR SHRINKAGE

The methods of choosing the shrinkage parameter d mentioned thus far have all been stochastic methods. That is d_1, d_2, d_3 and C_p^R are methods of choosing d which rely on the random response variable $\underline{Y}$. Perhaps in the same vein as Tripp (1983), a nonstochastic method of choosing d should be investigated. Tripp developed a shrinkage parameter estimate which solely relies on the ill-conditioning of the explanatory variables of standard multiple ridge regression with common variance: Tripp coins his method DF-trace, which for the most part

evaluates the trace of the ridge hat matrix. The counterpart to DF-trace in the GLM will be shown to rely on the information matrix. Immediately notice that K^{-1} is usually unknown and is estimated stochastically. Hence it can be argued that there does not truly exist a nonstochastic method of choosing d in the GLM.

Other complications arise in trying to parallel the construction of Tripp's shrinkage parameter estimate. Tripp considers the quasi-projector, P_F or the ridge hat matrix. That is, in standard multiple ridge regression,

$$\begin{aligned}\hat{\underline{y}} &= X\hat{\underline{\beta}}_R \\ &= H_d \underline{y} \\ &= X(X'X + dI)^{-1}X'\underline{y} \\ &= UD^2(D^2 + dI)^{-1}U'\underline{y} \\ &= UFU'\underline{y} \\ &= P_F\underline{y},\end{aligned}$$

from the singular value decomposition in equation (4.3.2). U represents the nonzero eigenvectors of XX'. $D^2 = \text{diag}\{\mu_i{}^2\}$ is a diagonal matrix of the eigenvalues of $X'X$, $F = D^2(D^2 + dI)^{-1}$ and $P_F = UFU'$. The μ are defined the singular values. The matrix X is considered as centered and scaled and does not contain the column of ones corresponding to the intercept term. Hence define

$$\begin{aligned}\text{DF-trace} &= \text{tr}(P_F) \\ &= \sum_{i=1}^{p} f_i \\ &= \sum_{i=1}^{p} \mu_i^2 / (\mu_i^2 + d).\end{aligned} \tag{6.5.1}$$

The difficulty in directly extending the above procedure to the GLM is that there does not exist a counterpart projection matrix to that of standard multiple regression.

The notion of a hat or projection matrix does not neatly extend beyond the usual linear model to the class of generalized linear models. However, there is some hope to finding a reasonable candidate within the GLM which does in some way connect to the structure of H_d, particularly in trace as in equation (6.5.1). In the general development of C_p, one can match up pieces of C_p^R in usual multiple ridge regression to that of the one of ridge regression in the GLM. The $\mathrm{tr}(\Phi\Phi_d^{-1})$ in the GLM corresponds to the $\mathrm{tr}(P_F) = \mathrm{tr}(H_d)$ in standard multiple regression. See and compare equations (6.4.1) and (6.4.12), where $\Phi_d = X'K^{-1}X + dI$. Notice also that

$$\mathrm{tr}(\Phi\Phi_d^{-1}) = \sum_{i=0}^{p} \lambda_i / (\lambda_i + d), \tag{6.5.2}$$

where the λ_i are the $p+1$ eigenvalues of $\Phi = X'K^{-1}X$. Equation (6.5.2) neatly matches the motivation of equation (6.5.1). In fact, equation (6.5.2) will collapse to equation (6.5.1) when the identity link is used for normal responses having common variance. The matrix K^{-1} is estimated via maximum likelihood. See equation (5.4.2) for an evaluation of the variance of the diagonal elements. The sum in equation (6.5.1) has p terms corresponding to the p explanatory variables, whereas the sum in equation (6.5.2) has $p+1$ terms since the constant term is incorporated into the weighting structure of the information matrix.

Thus a reasonable construction of a DF-trace statistic in the GLM would be

$$\mathrm{DF}^* = \sum_{i=0}^{p} \lambda_i / (\lambda_i + d) = \sum_{i=0}^{p} f_i. \tag{6.5.3}$$

Notice that in the least squares setting with common variance

$$\mathrm{DF}^* = \mathrm{DF} + 1$$

Further when $d = 0$,

$$DF^* = p + 1$$
$$DF = p.$$

The idea of a DF-trace procedure stems from (other than Tripp (1983)) Marquardt (1970) and Vinod (1976). The objective is to find the effective rank of $X'X$, $X'K^{-1}X$ in the GLM framework. Suppose that the rank$(X) = p'$. The value p' is employed to determine the shrinkage parameter d. Not only is the notion of DF essentially to give the effective rank of $X'X$, but also to give a glimpse at the nontrivial degrees of freedom for regression. One can see immediately the connection to principal component regression, which is ultimately a procedure to reduce the dimensionality of $X'X$ to $p - r$, where r is the number of trivial dimensions determined by some rule. Notice, however, that DF is not restricted to integer values. In situations when p' is not an integer, often the collinearity is termed a diffuse collinearity.

The mathematics of DF*-trace follows directly from Tripp. That is $f_i = \lambda_i / (\lambda_i + d)$ is a convex decreasing function in d. The slope of f_i is

$$\frac{\partial f_i}{\partial d} = -\frac{\lambda_i}{(\lambda_i + d)^2} \quad \text{and}$$
$$\frac{\partial DF^*}{\partial d} = -\sum_{i=0}^{p} \frac{\lambda_i}{(\lambda_i + d)^2}, \tag{6.5.4}$$

which is always less than zero and thus decreasing. Notice that for $d = 0$ the slope $\partial DF^* / \partial d = \sum_{i=0}^{p} \lambda_i^{-1}$. For orthogonal columns of $K^{-1/2}X$ and $d = 0$, the slope is $-(p + 1)$. The second derivative of f_i with respect to d is

$$\frac{\partial^2 f_i}{\partial d^2} = \frac{2\lambda_i^2}{(\lambda_i + d)^2}, \tag{6.5.5}$$

which is always positive; thus f_i is convex in d. For $d \geq 0$, then $0 \leq f_i \leq 1$ for all i. Notice that $\lim_{d \to \infty} f_i = 0$. Further $f_i = 1$ for $d = 0$. Notice that by setting $d = \lambda_i$, an intermediate value of $f_i = .5$ is found.

Using the effective rank of Φ as a criterion to estimate $\underline{\beta}$ certainly is somewhat subjective. Similar problems arise in the ridge trace procedures. Graphs do help in determining a reasonable window for the shrinkage parameter. Usually d is chosen large enough so that DF* -trace has stabilized. Thus DF*-trace can be thought of as an estimate of p', lending itself to diffuse collinearities. The plot of DF* vs. d will be instructive in determining the value of DF* for which the slope is near to that of an orthogonal system. Tripp (1983) is careful not to overdamp the dominant components. Bounds can be imposed on d to protect against overdamping. For example if the researcher recognizes that the smallest important f_i should not be shrunk more than a specified fraction ρ $(0 \leq \rho \leq 1)$, then set

$$f_i = \rho = \lambda_i / (\lambda_i + d).$$

This implies

$$d_{\max} = \lambda_i(1 - \rho) / \rho. \tag{6.5.6}$$

In practice, d is chosen to be much less than $d_{\max}$.

Tripp also points out that controversy in choosing d, which is commonplace in ridge trace procedures, can be avoided by also graphing a line representing the orthogonal situation. Recall that

$$\mathrm{DF}^* = \sum_{i=0}^{p} \lambda_i / (\lambda_i + d) = (1 + p) / (1 + d),$$

when the columns of $K^{-1/2}X$ are orthogonal giving $\lambda_i = 1$ for all i. Since DF* vs. d can be compared directly to the orthogonal system, there is less of an impact in a varied scale of d. An example follows in section 6.6.

6.6 EXAMPLE USING VARIOUS DEGREES OF SHRINKAGE

Previously in section 4.6, an example was given comparing various biased estimation techniques to maximum likelihood. Table 6 in section 4.6 gives a summary. The logit link function was used on Bernoulli cancer remission data displayed in Appendix A.

In an attempt to choose a ridge shrinkage parameter that yields good predictions, the C_p^R method can be implemented. In Figures 4 and 5, a plot of C_p^R vs. d is displayed. The two plots differ by a varied scale of the d axis. The plots suggest choosing a shrinkage parameter of approximately, $d_{CP} = .0080$. Such a choice minimizes C_p^R . See equation (6.4.2).

The DF*-trace procedure was also used as a diagnostic tool to choose d. In this case, the orthogonal system is overlayed (i.e. all $\lambda_i = 1$). Figures 6 and 7 display DF*-trace as a function of d. Notice that there is less impact of a varied scale, as Tripp (1983) suggested when also graphing the orthogonal situation. The DF*- trace procedure is a subjective one. However, the graphs suggest choosing $d_{DF} = .0003$; hence implying that the effective rank of Φ is in the order of 5.0. Recall the notion of diffuse collinearities explained in section 6.5. Table 6 in section 4.6 includes the C_p^R and DF* -trace parameter estimates along with the asymptotic standard errors.

6.7 A STEIN ESTIMATOR IN THE GENERALIZED LINEAR MODEL

An estimation technique, which was originally suggested by Stein (1960) for least squares estimation, is defined as

$$\hat{\underline{\beta}}_S = c\hat{\underline{\beta}} \ , \tag{6.7.1}$$

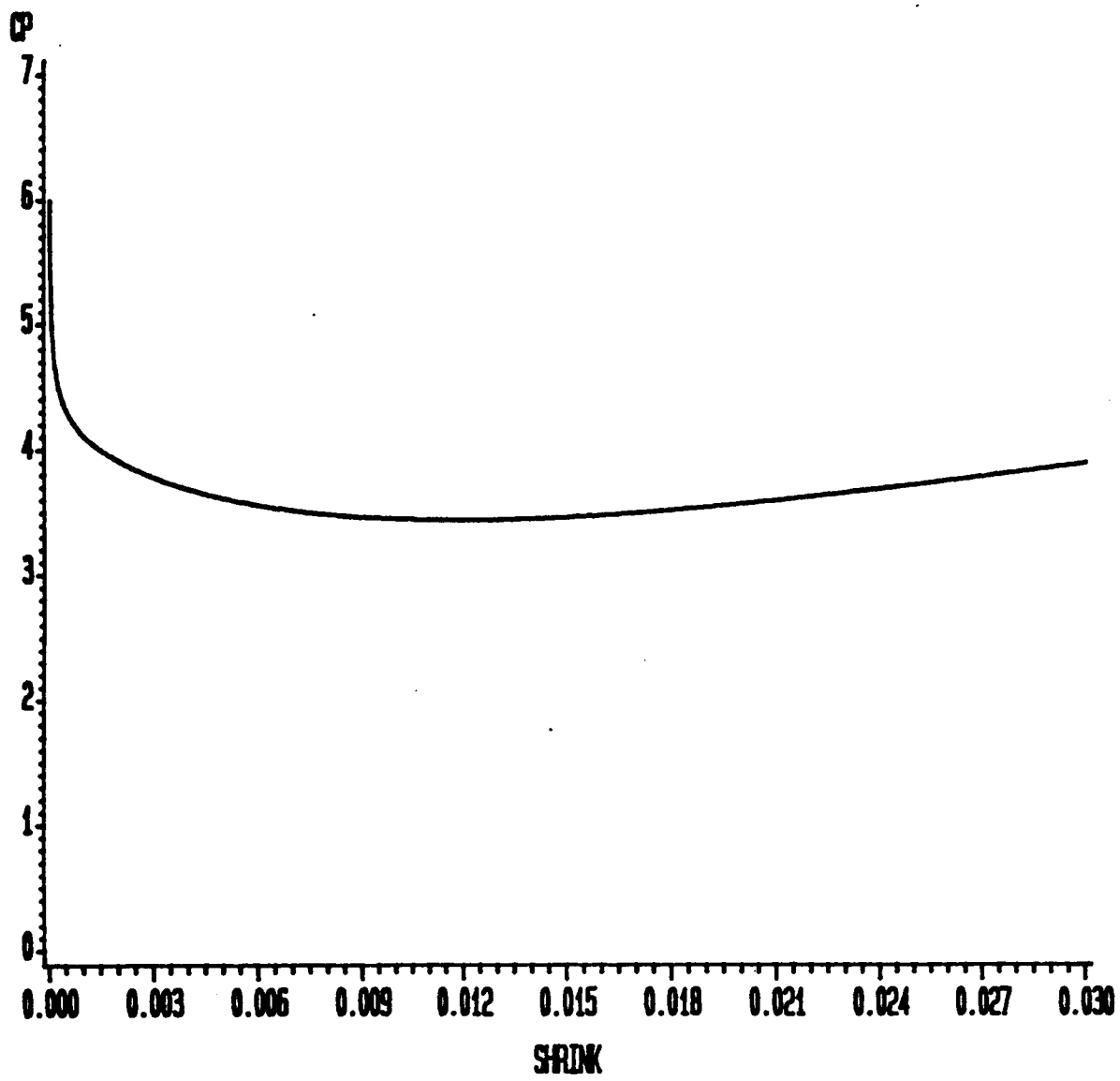

Figure 4. SHRINKAGE USING CP CRITERION FOR CANCER EXAMPLE

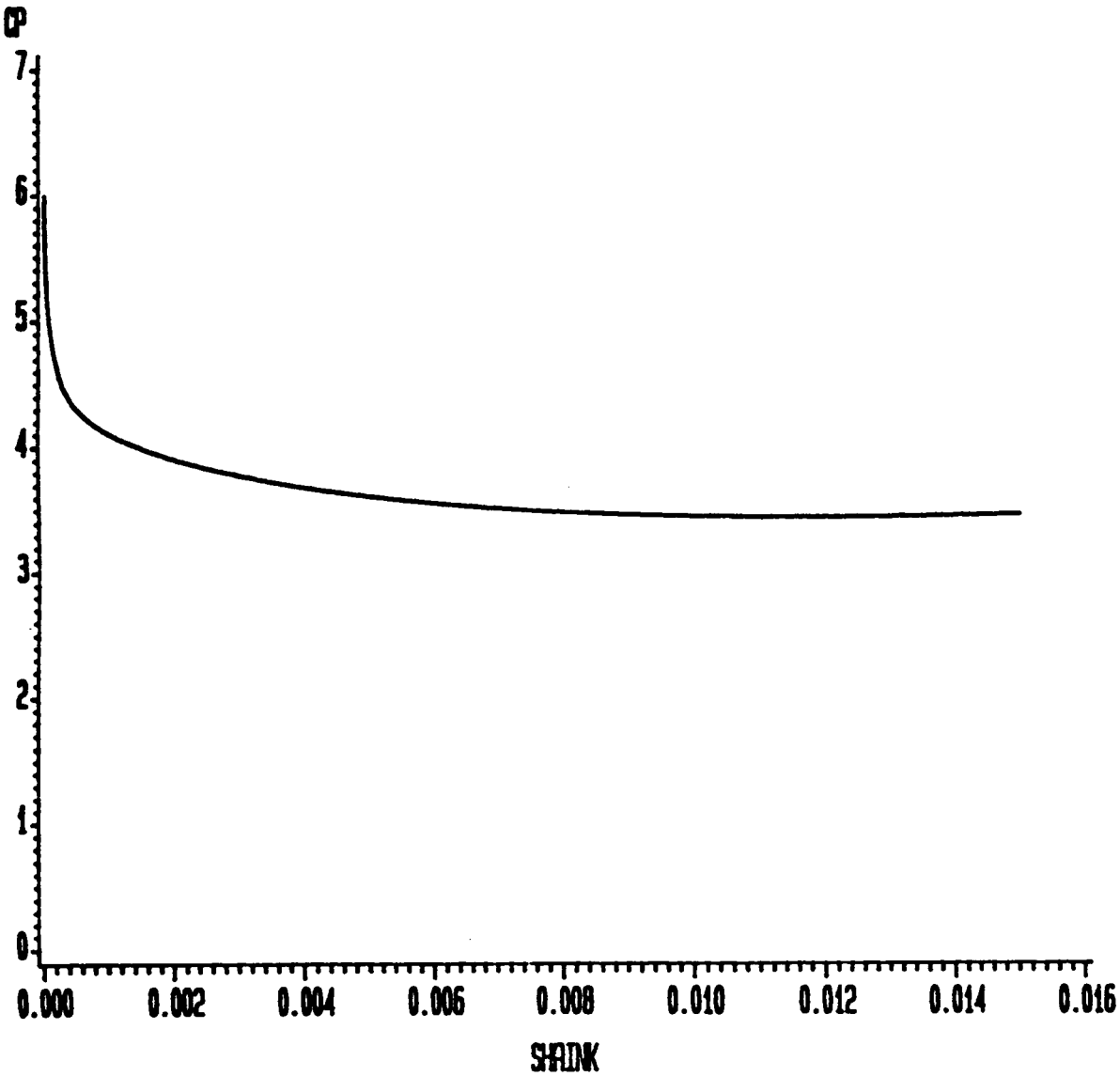

Figure 5. SHRINKAGE USING CP CRITERION FOR CANCER EXAMPLE

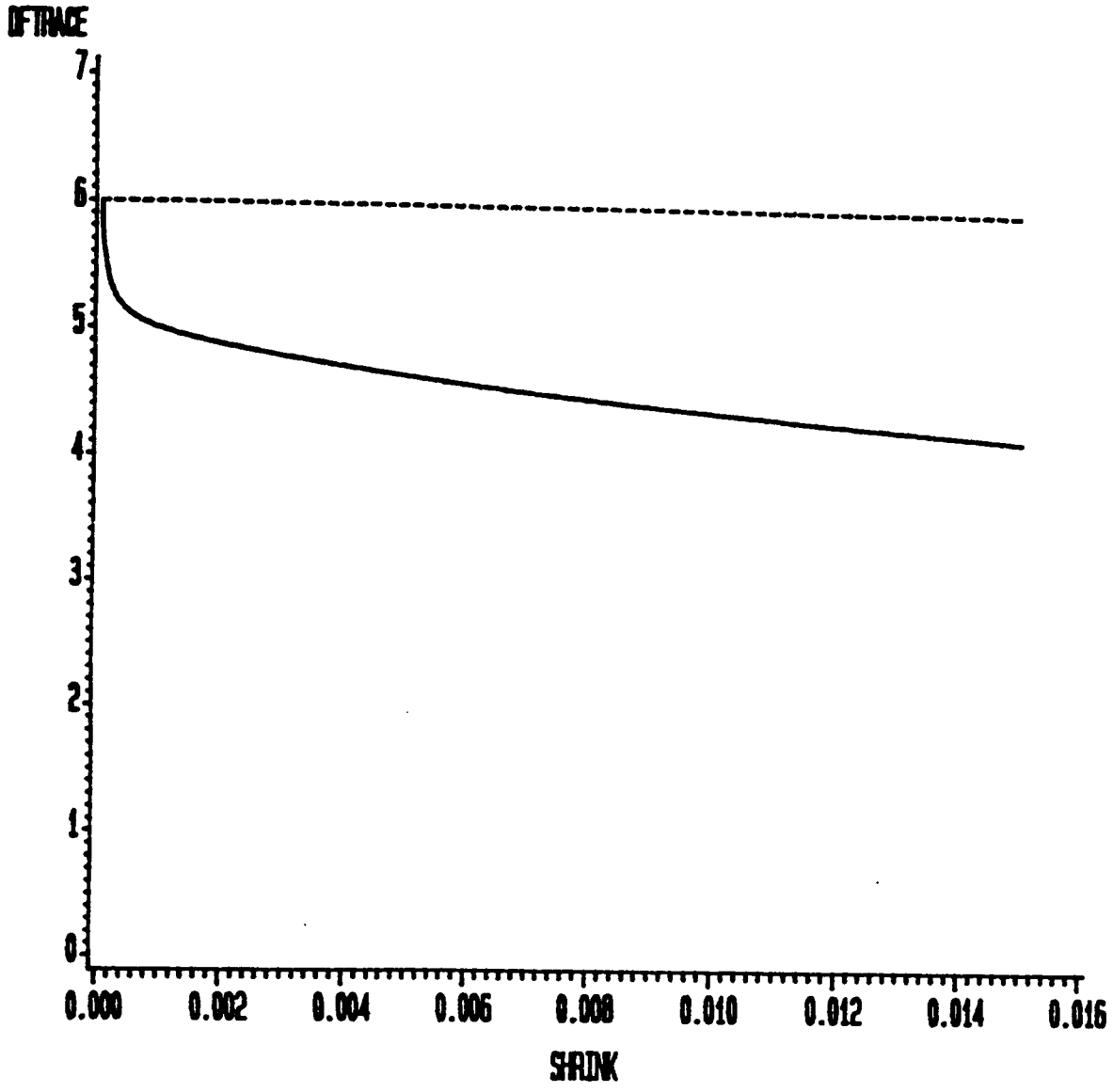

Figure 6. SHRINKAGE USING DF-TRACE FOR CANCER EXAMPLE

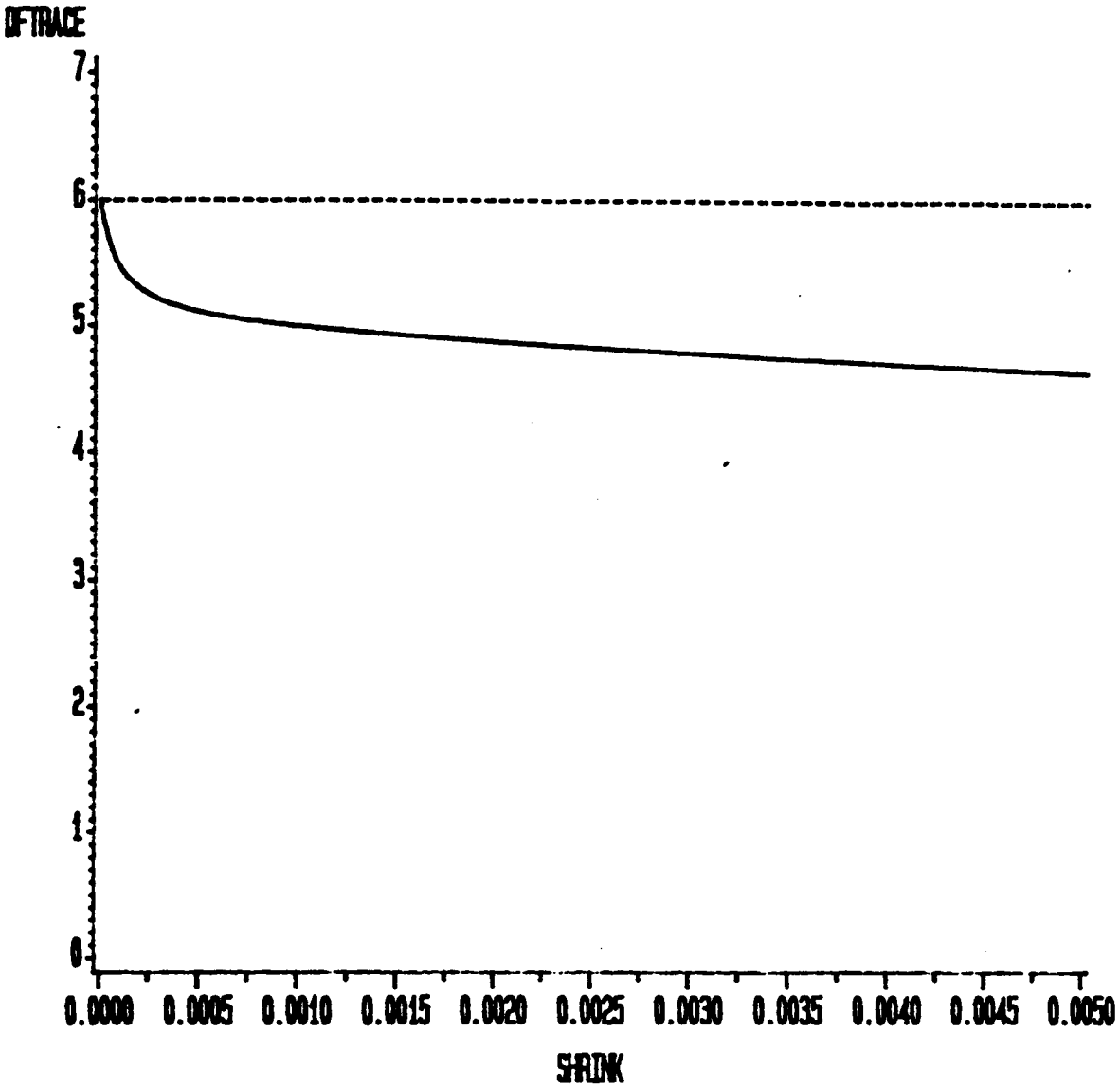

Figure 7. SHRINKAGE USING DF-TRACE FOR CANCER EXAMPLE

where $0 < c < 1$. Stein suggested the following choice for c

$$c = \hat{\underline{\beta}}'\hat{\underline{\beta}} \,/\, (\hat{\underline{\beta}}'\hat{\underline{\beta}} + \mathrm{tr}(X'X)^{-1}) \,. \tag{6.7.2}$$

The motivation for such an estimator is that in the presence of collinear explanatory variables (in least squares)

$$E(\hat{\underline{\beta}}'\hat{\underline{\beta}}) \geq \mathrm{tr}(\mathrm{Var}(\hat{\underline{\beta}})) = \mathrm{tr}(X'X)^{-1} = \sum_{i=0}^{p} \lambda_i^{-1} \geq \lambda_{\min}^{-1} \,,$$

which demonstrates that the estimate $\hat{\underline{\beta}}$ may be too long on the average.

Schaefer (1986) considers the natural extension for Stein estimation in a logistic regression setting. He simply presents a scalar shrinking of the maximum likelihood estimates by using

$$c = \hat{\underline{\beta}}'\hat{\underline{\beta}} \,/\, (\hat{\underline{\beta}}'\hat{\underline{\beta}} + \mathrm{tr}(X'\hat{V}X)^{-1}) \,, \tag{6.7.3}$$

where $\hat{V} = \mathrm{diag}\{\hat{\pi}_i(1 - \hat{\pi}_i)\}$ from maximum likelihood. Schaefer points out the ease of implementation of such an estimator in the existing logistic regression programs.

A wider application of Stein estimation can be addressed with the presence of of an ill-conditioned information matrix, Φ. For the generalized linear model, the corresponding choice of c is

$$c_1 = \hat{\underline{\beta}}'\hat{\underline{\beta}} \,/\, (\hat{\underline{\beta}}'\hat{\underline{\beta}} + \mathrm{tr}(\Phi^{-1})) \,, \tag{6.7.4}$$

which is a generalization of Schaefer's logistic Stein estimator.

The choice of c_1, given above, based on Schaefer's idea is one which minimizes the $E(L_1^2)$ criterion. See equation (7.2.5). Asymptotically the following holds true;

$$
\begin{aligned}
E(L_1^2) &= E(c_1\hat{\underline{\beta}} - \underline{\beta})'(c_1\hat{\underline{\beta}} - \underline{\beta}) \\
&= c_1^2 \operatorname{tr}(I\,\Phi^{-1}) + \sum_{i=0}^{p} (c_1 - 1)^2 \beta_i^2 \\
&= c_1^2 \sum_{i=0}^{p} \lambda_i^{-1} + (c_1 - 1)^2 \sum_{i=0}^{p} \beta_i^2 .
\end{aligned}
\tag{6.7.5}
$$

Notice that

$$
\frac{\partial E(L_1^2)}{\partial c_1} = 0 = 2c_1 \sum_{i=0}^{p} \lambda_i^{-1} + 2(c_1 - 1) \sum_{i=0}^{p} \beta_i^2 .
$$

Substituting $\hat{\beta}_i$ for β_i yields the the minimum c_1 for the $E(L_1^2)$ criterion in equation (6.7.4). Notice also that

$$
\frac{\partial^2 E(L_1^2)}{\partial c_1^2} = 2\left(\sum_{i=0}^{p} \lambda_i^{-1} + \sum_{i=0}^{p} \beta_i^2 \right) \geq 0 .
$$

Hence c_1 is in fact a minimum.

Perhaps a more appropriate choice for the scaling constant c in the logistic model, and for that matter all generalized linear models, is one based on the $E(L_2^2)$ criterion. See equation (7.2.6). The $E(L_2^2)$ criterion lends itself to the GLM since incorporates the asympotic variance-covariance matrix for the estimated parameter vector.

$$
E(L_2^2) = (c_2\hat{\underline{\beta}} - \underline{\beta})'\Phi(c_2\hat{\underline{\beta}} - \underline{\beta}). \tag{6.7.6}
$$

In taking the derivative of $E(L_2^2)$ with respect to c_2 and replacing α_i with $\hat{\alpha}_i$, the minimizing value of c_2 is

$$c_2 = \frac{\sum_{i=0}^{p} \hat{\alpha}_i^2 \lambda_i}{\sum_{i=0}^{p} \hat{\alpha}_i^2 \lambda_i + (p+1)}, \tag{6.7.7}$$

where again M is the orthogonal matrix to diagonalize Φ, $M'\Phi M = \text{diag}\{\lambda_i\}$, and $\hat{\underline{\alpha}} = M'\hat{\underline{\beta}}$. Quite clearly for any $0 < c < 1$, not only are the parameter estimates shrunken in magnitude, but also the variances of these estimators is reduced.

Chapter VII

GENERALIZED FRACTIONAL PRINCIPAL COMPONENT ANALYSIS

7.1 INTRODUCTION

The generalized principal component (GPC) estimator given in section 5.3 and the generalization of the ridge estimator given in section 6.2 can be shown to be members of a broader class of shrinkage estimators for the generalized linear model in the canonical form defined in equation (5.3.1). This broad class of shrinkage estimators will be referred to as generalized fractional principal component (GFPC) estimators. While, in the GLM principal component and ridge estimators shrink the estimated parameter vector toward length zero, GFPC estimation also accomplishes shrinking the parameter estimates by taking a general weighting of the canonical variable components (see Lee (1986)).

7.2 DEVELOPMENT OF GFPC

Consider the GLM in the canonical form.

$$g(\underline{\mu}) = Z\,\underline{\alpha}, \tag{7.2.1}$$

where $Z = XM$ and $\underline{\alpha} = M'\underline{\beta}$. The model given in equation (7.2.1) can be rewritten as

$$\begin{aligned} g^{*}(\underline{\mu}) &= ZF^{-}F\underline{\alpha} \\ &= W\underline{\gamma}, \end{aligned} \tag{7.2.2}$$

where $W = ZF^{-}$, $\underline{\gamma} = F\underline{\alpha}$ and $F = \mathrm{diag}\{f_{ii}\}$ is a diagonal matrix of weights. The weights in F are contained in the unit interval [0,1]. F^{-} is a generalized inverse of F since some of the diagonal elements may be zero. $g^{*}(\mu) = g(\mu)$ if and only if the $f_{ii} \neq 0$ for all i ; in this case use $F^{-1} = F^{-}$. The information matrix, corresponding to equation (7.2.2), is of the form

$$W'K^{-1}W = F^{-}M'\Phi MF^{-} = F^{-}\Lambda F^{-}.$$

Thus the maximum likelihood equations, given below, are similar to those in equation (5.3.4).

$$\begin{aligned} \hat{\underline{\gamma}}_t &= \hat{\underline{\gamma}}_{t-1} + (F^{-}\hat{\Lambda}_{t-1}F^{-}) - \left[\sum_{i=1}^{N} \underline{w}_i \hat{k}_{ii}^{-1}(y_i - \hat{\mu}_i)\frac{\partial \eta_i}{\partial \mu_i}\right]_{t-1} \\ &= (F^{-}\hat{\Lambda}_{t-1}F^{-}) - \left[\sum_{i=1}^{N} \underline{w}_i \hat{k}_{ii}^{-1}\left[\underline{w}'_i\hat{\underline{\gamma}}_{t-1} + (y_i - \hat{\mu}_i)\frac{\partial \eta_i}{\partial \mu_i}\right]\right]_{t-1} \\ &= [(F^{-}\hat{\Lambda}_{t-1}F^{-}) - W'\hat{K}^{-1}\underline{\gamma}^{*}]_{t-1} \\ &= [F\hat{\Lambda}^{-1}Z\hat{K}^{-1}\underline{\gamma}^{*}]_{t-1} \\ &= F\,\hat{\underline{\alpha}}_{t-1}, \end{aligned} \tag{7.2.3}$$

where $y_i^{*} = \eta_i + (y_i - \mu_i)\dfrac{\partial \eta_i}{\partial \mu_i}$ and K^{-1} are updated at each iteration step. If the maximum likelihood estimate for $\underline{\gamma}$ converges, then further estimation is not needed for K^{-1}. A Variance

argument supports the use of maximum likelihood estimates for the diagonal matrix K^{-1} when X data combinaiions are in the main stream (see equation (5.4.2)).

A reasonable approach to generalized fractional principal component parameter estimation, using canonical form models, is to first estimate the the full $p+1$ vector of maximum likelihood estimates, $\hat{\underline{\alpha}}$. Hence K^{-1} is also estimated via maximum likelihood. It follows that the class of GFPC estimators are

$$\hat{\underline{\gamma}} = F\hat{\underline{\alpha}}. \tag{7.2.4}$$

In the case when $F = I_{p+1}$, then the GFPC estimators reduce to maximum likelihood estimators. Further, extending equation (6.2.8) to the canonical form, a generalization for ridge estimators becomes

$$\begin{aligned} \underline{\alpha}^R(d) &= (\hat{\Lambda} + dI)^{-1}\hat{\Lambda}\hat{\underline{\alpha}} \\ &= F_R\hat{\underline{\alpha}}, \end{aligned}$$

where $F_R = \mathrm{diag}\{\lambda_i / (\lambda_i + d)\}$. Thus the generalizations to ridge estimators also fit equation (7.2.4) nicely. As another example, the generalized principal component (GPC) estimator simply chooses $f_{ii} = 1$ for the components chosen to stay in the model and $f_{ii} = 0$ otherwise. Table 9 lists choices of F for various generalized fractional principal component estimators.

As in section 6.7, measures of closeness between $\hat{\underline{\gamma}}$ and $\underline{\alpha}$ were developed by Stein (1960). To determine how the choice of F effects the asymptotic mean squared error (weighted or unweighted) for $\hat{\underline{\gamma}}$, consider the following criteria. Asymptotically, equations (7.2.5) and (7.2.6) hold true.

$$\begin{aligned} E(L_1^2) &= E(\hat{\underline{\gamma}} - \underline{\alpha})'(\hat{\underline{\gamma}} - \underline{\alpha}) \\ &= \mathrm{tr}(I\,F\Lambda^{-1}F) + \underline{\alpha}'(F - I)^2\underline{\alpha} \\ &= \sum_{i=0}^{p} f_{ii}^{\,2}\lambda_i^{-1} + \sum_{i=0}^{p} \alpha_i^2(f_{ii} - 1)^2, \end{aligned} \tag{7.2.5}$$

Table 9. GENERALIZED FRACTIONAL PC WEIGHTS

ESTIMATION TECHNIQUE	MATRIX F
Maximum Likelihood	$F = I_{p+1}$
Principal Component	$f_1 = \cdots = f_s = 1\,,\ f_{s+1} = \cdots = f_{p+1} = 0$
Ridge	$f_i = \lambda_i / (\lambda_i + k)$
Generalized Ridge	$f_i = \lambda_i / (\lambda_i + k_i)$
Stein: L_1^2	$f_i = \hat{\underline{\alpha}}'\hat{\underline{\alpha}} / (\hat{\underline{\alpha}}'\hat{\underline{\alpha}} + \sum_i \lambda_i^{-1})$ for all i
Stein: L_2^2	$f_i = \sum_i \lambda_i \alpha_i^2 / (\sum_i \lambda_i \alpha_i^2 + (p+1))$ all i
Fraction (ρ)	$f_1 = f_2 = \cdots f_{s-1} = 1,\quad 0 < f_s = \rho < 1$ $f_{s+1} = f_{s+2} = \cdots f_{p+1} = 0$

$$
\begin{aligned}
E(L_2^2) &= E(\hat{\underline{\gamma}} - \underline{\alpha})'\Lambda(\hat{\underline{\gamma}} - \underline{\alpha}) \\
&= \mathrm{tr}(\Lambda F \Lambda^{-1} F) + \underline{\alpha}'(F - I)^2 \Lambda \underline{\alpha} \\
&= \sum_{i=0}^{p} f_{ii}^{\,2} + \sum_{i=0}^{p} \alpha_i^2 \lambda_i (f_{ii} - 1)^2 .
\end{aligned}
\tag{7.2.6}
$$

7.3 COMPARISONS AMONG FRACTIONAL ESTIMATORS

Comparisons can be made between the various generalized fractional principal component estimators using the $E(L^2)$ criterion. For example, consider comparing maximum likelihood (ML) to principal component (GPC) via the asymptotic weighted mean squared error, $E(L_2^2)$. Recall that $r = p + 1 - s$ is the number of components deleted in a principal component setting. For simplicity in notation, let $\sum_r$ and $\sum_l = \sum_{p+1}$ be the respective sums over the deleted components and the entire set of components.

$$
\begin{aligned}
E_{PC}(L_2^2) &= \sum_l f_{ii}^{\,2} + \sum_l \alpha_i^2 \lambda_i (f_{ii} - 1)^2 \\
&= (p + 1) - r + \sum_r \alpha_i^2 \lambda_i .
\end{aligned}
\tag{7.3.1}
$$

For maximum likelihood

$$
E_{ML}(L_2^2) = p + 1, \tag{7.3.2}
$$

since each term of the sum on the right is identically equal to zero. The use of principal component estimators over maximum likelihood is only justified by the $E(L_2^2)$ criterion if

$$
\sum_r \lambda_i \alpha_i^2 \le r . \tag{7.3.3}
$$

Perhaps further developments in the area of GFPC will involve equation (7.3.3) as the null hypothesis in a testing scenario.

The fraction (ρ) technique yields

$$E_F(L_2^2) = p + \rho - r + \sum_r \alpha_l^2 \lambda_l + (1-\rho)^2 \alpha_r^2 \lambda_r. \tag{7.3.4}$$

Comparisons can be made to the principal component estimator via

$$E_{PC}(L_2^2) - E_F(L_2^2) = (1-\rho) - (1-\rho)\alpha_r^2 \lambda_r. \tag{7.3.5}$$

Notice when $\rho = 1$ that $E_{PC}(L_2^2) = E_F(L_2^2)$ as expected. Further $E_{PC}(L_2^2) \leq E_F(L_2^2)$ if and only if, for fixed ρ, $\alpha_r^2 \lambda_r \geq 1$.

To match the fractional (ρ) estimator against maximum likelihood notice that

$$E_{ML}(L_2^2) - E_F(L_2^2) = 1 - \rho + r - \sum_r \alpha_l^2 - (1-\rho)^2 \alpha_r^2 \lambda_r. \tag{7.3.6}$$

Hence the fractional (ρ) estimator is better in regard to $E(L_2^2)$ if

$$(1-\rho)[1 - (1-\rho)\alpha_r^2 \lambda_r] + r \geq \sum_r \alpha_l^2 \lambda_l. \tag{7.3.7}$$

Notice in the case when $\rho = 1$, the result from equation (7.3.3) holds.

It follows from equation (7.2.6) that the $E(L_2^2)$ for the Stein estimator is

$$E_S(L_2^2) = (p+1)c_2^2 + (c_2 - 1)^2 \sum_{p+1} \alpha_l^2 \lambda_l, \tag{7.3.8}$$

where $c_2 = \sum_i \alpha_i^2\lambda_i / (\sum_i \alpha_i^2\lambda_i + p + 1)$. In comparing the Stein estimator to the maximum likelihood estimator, consider

$$\begin{aligned} E_S(L_2^2) - E_{ML}(L_2^2) &= (p+1)(c_2^2 - 1) + (c_2 - 1)^2 \sum_{p+1} \alpha_i^2\lambda_i \\ &= (c_2 - 1)[(p+1)(c_2 + 1) + (c_2 - 1)\sum_{p+1} \alpha_i^2\lambda_i]. \end{aligned} \tag{7.3.9}$$

The Stein estimator will make

$$E_S(L_2^2) - E_{ML}(L_2^2) < 0,$$

if and only if

$$1 > c_2 > \frac{\sum_{p+1} \alpha_i^2\lambda_i - (p+1)}{\sum_{p+1} \alpha_i^2\lambda_i + (p+1)}$$

and $\sum_{p+1} \alpha_i^2\lambda_i > (p+1)$. Otherwise the maximum likelihood is a better than Stein in terms of $E(L_2^2)$.

In comparing generalized ridge to ridge, certainly there exists a set of k_i such that

$$\begin{aligned} E_{GR}(L_2^2) - E_R(L_2^2) &= \sum_{p+1} \left[\frac{\lambda_i^2}{(\lambda_i + k_i)^2} - \frac{\lambda_i^2}{(\lambda_i + k)^2} \right] \\ &\quad + \sum_{p+1} \alpha_i^2\lambda_i \left[\frac{k_i^2}{(\lambda_i + k_i)^2} - \frac{k^2}{(\lambda_i + k)^2} \right] \le 0. \end{aligned} \tag{7.3.10}$$

In theory, generalized ridge is trivially guaranteed to be at least as good as the ridge method which may or may not be the case in practice.

Chapter VIII

SIMULATION STUDY

8.1 INTRODUCTION

The relative merits of various asymptotically biased estimation techniques are investigated as reasonable alternatives to method of scoring maximum likelihood via a simulation study. The parameter estimation techniques are evaluated by variance, bias, and mean square error (MSE). Other factors of interest are the sample size, number of explanatory variables, severity of ill-conditioning of the information, and the distributional form of the response variable.

8.2 PROCEDURE FOR SIMULATION

Appendix B comprises the hub of the simulation study; a program which is written in *SAS* Proc Matrix. The program uses either $p = 3$ or $p = 5$ centered and scaled explanatory variables (augmented to a constant term) from a fixed data set of $N = 17$ or $N = 45$. The

smaller data set is a random subset of the larger data set. Furthermore, the ill-conditioning of the information matrix is deemed as moderate or severe. Hence, consider the resulting $2 \times 2 \times 2 = 8$ possible combinations of factors.

In using these combinations of factors, the main interest is the distributional form of the response variable and how well various techniques estimate the unknown parameter vector. Normal data with the identity link reduces maximum likelihood estimation to the one step least squares multiple regression parameter estimation and biased estimation for multiple regression is well documented. Consequently, normal data will not be incorporated into this simulation study. Response distributions of interest will include the Poisson and Bernoulli. The program is capable of changing the link function as well as the diagonal matrix of weights, K^{-1} . The method of scoring is used for maximum likelihood.

In this study, the parameter vector $\underline{\beta}$ is assumed to be known and is fixed within the program (labelled as TRUEB). When further given the fixed, nonstochastic, known explanatory variables X , then the linear combination $\underline{\eta} = X\underline{\beta}$ is trivially known. In using the natural link function of the exponential family, the following relationship is utilized

$$g(\underline{\mu}) = \underline{\eta}$$
$$\underline{\mu} = h(\underline{\eta}),$$

where h is a nonlinear function in the parameters. Once given the vector $\underline{\mu} = E(\underline{Y})$ of means and the other parameters of the distribution which are functions of the known $\underline{\eta}$, naturally the next step of action is to generate a vector of random response variates from a specified distributional form using functions given in *SAS* Basics Manual (1982). As a result, the $N \times 1$ vector of responses, $\underline{Y}$, are generated via known $\underline{\eta}$ and the natural link function.

Having generated $\underline{Y}$ and X in hand, estimation of the parameter vector is in order. In fact, six estimation techniques are implemented: (1) Method of scoring maximum likelihood, (2) Ridge estimation using the harmonic mean method for shrinkage, (3) Schaefer's principal

component technique deleting one dimension, (4) Schaefer's principal component technique deleting two dimensions, (5) Iterative principal component technique minus one dimension, (6) Iterative principal component technique minus two dimensions. Refer to these by number.

From the preset parameter vector $\underline{\beta}$ and the explanatory variables X of dimension $N \times 4$ or $N \times 6$, for $N = 17$ or $N = 45$, the response vector $\underline{Y}$ is generated 1000 times (repetitions). Consequently, the six estimation procedures outlined above are also computed 1000 times, corresponding to each response vector. Using the 1000 repetitions, the *SAS* program has the ability to compute a sample mean vector of parameter estimators for each of the six estimation techniques. The program also computes the sample variance of each component of each vector of each estimation technique. Moreover, since the true $\underline{\beta}$ is known, bias can be computed for each technique (r) via $(\overline{\underline{\beta}}^{(r)} - \underline{\beta})$ for $r = 1, 2, \ldots, 6$. Combining variance and bias above leads to a mean square error criterion.

Anomalies and other nuisances occur during the course of 1000 repetitions. In the event that convergence is not met during an iterative procedure or some estimation technique yields $\|\hat{\underline{\beta}}\|$ which is inflated over an upper bound (10^5) or shrunk below a lower bound (10^{-5}), then the estimate is set to zero and does not contribute to the summary statistics. Hence the results presented are conditional on convergence and parameter estimates within subjective bounds. For example, if a repetition of response vector does not provide convergence in maximum likelihood estimates, then the estimates (1), (2), (3), and (4) are all set to zero since they rely on maximum likelihood.

8.3 RESULTS OF SIMULATION

Tables 10, 11, 12, and 13 present results for Poisson responses using the larger sample size of $N = 45$. There is a world of information to summarize.

First notice the effects of severe ill-conditioning in Tables 10 and 12 when compared to moderate ill-conditioning in Tables 11 and 13. The detrimental effects of severe ill-conditioning are apparent in observing the large values of MSE for the maximum likelihood estimates. Any choice of biased estimation greatly reduces the MSE. MSE is not nearly as greatly inflated for moderate ill-conditioning.

Secondly, it appears, from these simulation results, that the number of explanatory variables has an impact on the number of repetitions dropped from the summary statistics. The number of anomalies increase for estimation with increased number of regressors. For example, in the case of $p+1=4$ in Table 10 and Table 11, 11 / 1000 and 18 / 1000 repetitions are dropped respectively from the analysis. However for $p+1=6$, then the Tables 12 and 13 display 101 / 1000 and 52 / 1000 repetitions are dropped respectively. Perhaps one explanation would be that in the presence of ill-conditioned information, the value $\|\hat{\underline{\beta}}\|$ is large on the average. With $p+1=6$, there is a greater likelihood for $\|\hat{\underline{\beta}}\|$ to inflate beyond the upper bound.

As a few other general observations, notice in Tables 10 and 12 that the ridge estimate is not doing as well as in Tables 11 and 13. A reasonable explanation is that, in Tables 10 and 12, the presence of severely ill-conditioned information yields maximum likelihood estimates large in magnitude. Hence, in choosing a shrinkage parameter via the harmonic mean, it follows that $(p+1)/\hat{\underline{\beta}}'\hat{\underline{\beta}}$ is quite small. Further notice how similar Schaefer's principal component estimators are to the iterative principal component estimators on the average even though they differ at each repetition.

Tables 14, 15, 16, and 17 display the results for Bernoulli responses again using the larger sample size of $N=45$. The summary for the logit link using Bernoulli data is very similar of that described above for the Poisson data. The similarities in the results between the characteristics of the logistic and Poisson regressions are very reassuring. It is interesting to note that in all cases of moderate ill-conditioning (Tables 11, 13, 15 and 17) the principal component estimators minus two dimension (methods (4) and (6)), as expected, are not doing well. The

overdamping of these principal component estimators gives estimators near zero, with the wrong sign or high *MSE*. The last item to address is the issue of decreased sample size.

Tables 18 through 25 comprise the results decreased sample size. The small sample results (Tables 18-25) are in the same order of presentation as the large sample results (Tables 10-17), of course with the exception that $N = 17$ instead of $N = 45$. From the simulation results given, it is difficult to assess general statements about the effects of decreased sample size on mean square error. The major consequence is the increased number of repetitions deleted from the analysis. Perhaps this is the expected consequence since there is less data to support the regression. In the worst case investigated, Poisson regression with $p + 1 = 6$ having severe ill-conditioning (Table 20) rejected over 800 of the 1000 repetitions mainly due to violations of maximum likelihood estimation beyond the set upper bound. But even by reducing $p + 1 = 4$ having moderate ill-conditioning (Table 19), maximum likelihood was still rejected over 300 of the 1000 times. This suggests some alternate method to that of the method of scoring should be used for small sample sizes having ill-conditioned information. Again very similar results held for the small sample logistic regression simulation, except not as extreme. Logistic regression is typically very nice to work with due to the boundedness property of the diagonal matrix K^{-1}, unlike most other members of the generalized linear model.

Table 10. POISSON SIMULATION RESULTS

Response = Poisson $N = 45$ $p + 1 = 4$ Severe Ill-conditioning

Conditional Analysis on 989 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.467$
$\lambda_1 = 1.512$
$\lambda_2 = .012$
$\lambda_3 = .006$

TRUEB $= (-.5, -2, 1, 1)$

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$
Scoring (ML)	-.575	-2.084	2.372	-.183
Ridge (HM)	-.521	-1.741	1.297	.522
Schaefer PC(-1)	-.535	-2.089	1.020	1.021
Schaefer PC(-2)	-.389	.271	.364	.367
Iterative PC(-1)	-.555	-2.092	1.057	1.057
Iterative PC(-2)	-.475	.286	.383	.385

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$
Scoring (ML)	0.046	2.137	75.780	74.497
Ridge (HM)	0.040	1.761	12.362	11.839
Schaefer PC(-1)	0.043	2.126	0.523	0.527
Schaefer PC(-2)	0.036	0.117	0.282	0.286
Iterative PC(-1)	0.044	2.112	0.531	0.536
Iterative PC(-2)	0.035	0.109	0.256	0.259

	MSE
Scoring (ML)	155.757
Ridge (HM)	26.388
Schaefer PC(-1)	3.229
Schaefer PC(-2)	6.698
Iterative PC(-1)	3.340
Iterative PC(-2)	6.648

Table 11. POISSON SIMULATION RESULTS

Response = Poisson $N = 45$ $p + 1 = 4$ Moderate Ill-conditioning

Conditional Analysis on 982 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.216$
$\lambda_1 = 1.254$
$\lambda_2 = .524$
$\lambda_3 = .006$

TRUEB $= (-.5, -2, 1, 1)$

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$
Scoring (ML)	-.576	-1.943	1.277	.982
Ridge (HM)	-.483	-1.254	.764	.529
Schaefer PC(-1)	-.525	-2.322	.459	.514
Schaefer PC(-2)	-.393	-.326	.396	-.407
Iterative PC(-1)	-.548	-2.315	.471	.477
Iterative PC(-2)	-.482	-.374	.506	-.531

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$
Scoring (ML)	0.048	2.324	2.183	1.906
Ridge (HM)	0.037	1.149	0.921	0.878
Schaefer PC(-1)	0.038	1.559	0.555	0.693
Schaefer PC(-2)	0.039	0.243	0.485	0.477
Iterative PC(-1)	0.039	1.592	0.554	0.694
Iterative PC(-2)	0.037	0.246	0.490	0.503

	MSE
Scoring (ML)	6.547
Ridge (HM)	4.219
Schaefer PC(-1)	3.478
Schaefer PC(-2)	6.404
Iterative PC(-1)	3.536
Iterative PC(-2)	6.509

Table 12. POISSON SIMULATION RESULTS

Response = Poisson $N = 45$ $p + 1 = 6$ Severe Ill-conditioning

Conditional Analysis on 899 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.662$
$\lambda_1 = 1.943$
$\lambda_2 = .865$
$\lambda_3 = .517$
$\lambda_4 = .010$
$\lambda_5 = .003$

TRUEB $= (-.5, -2, 1, 1, -1, 1)$

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$	$\bar{\beta}_4$	$\bar{\beta}_5$
Scoring (ML)	-.629	-2.215	1.373	.832	-.881	.646
Ridge (HM)	-.549	-1.783	1.009	.719	-.829	.553
Schaefer PC(-1)	-.586	-2.206	.933	1.183	-.895	.823
Schaefer PC(-2)	-.499	-1.367	.559	.701	-1.115	.245
Iterative PC(-1)	-.605	-2.202	.951	1.189	-.870	.699
Iterative PC(-2)	-.561	-1.389	.603	.747	-1.166	.196

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$	$s^2(\beta_4)$	$s^2(\beta_5)$
Scoring (ML)	0.056	2.316	119.793	108.679	3.414	4.912
Ridge (HM)	0.046	1.630	13.581	11.147	1.815	1.814
Schaefer PC(-1)	0.049	2.217	0.538	0.745	2.416	2.118
Schaefer PC(-2)	0.043	1.567	0.379	0.561	1.180	1.585
Iterative PC(-1)	0.051	2.237	0.548	0.764	2.454	2.241
Iterative PC(-2)	0.046	1.529	0.385	0.573	1.293	1.707

	MSE
Scoring (ML)	239.039
Ridge (HM)	29.791
Schaefer PC(-1)	8.514
Schaefer PC(-2)	6.285
Iterative PC(-1)	8.693
Iterative PC(-2)	6.705

Table 13. POISSON SIMULATION RESULTS

Response = Poisson $N = 45$ $p + 1 = 6$ Moderate Ill-conditioning

Conditional Analysis on 948 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.881$
$\lambda_1 = 1.373$
$\lambda_2 = .998$
$\lambda_3 = .839$
$\lambda_4 = .306$
$\lambda_5 = .003$

TRUEB = (−.5, −2, 1, 1, −1, 1)

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$	$\bar{\beta}_4$	$\bar{\beta}_5$
Scoring (ML)	-.621	-1.993	1.212	.996	-.974	.822
Ridge (HM)	-.592	-1.196	.712	.396	-.586	.367
Schaefer PC(-1)	-.569	-2.075	.993	.810	-.795	.828
Schaefer PC(-2)	-.561	-.791	.589	-.569	-.526	-.391
Iterative PC(-1)	-.594	-2.209	1.005	.811	-.797	.842
Iterative PC(-2)	-.539	-.840	.546	-.599	-.517	-.539

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$	$s^2(\beta_4)$	$s^2(\beta_5)$
Scoring (ML)	0.051	3.179	2.527	2.732	2.433	2.826
Ridge (HM)	0.036	1.091	0.889	0.999	0.804	0.863
Schaefer PC(-1)	0.045	1.335	1.515	1.700	1.101	1.719
Schaefer PC(-2)	0.043	0.683	0.913	0.597	0.624	0.993
Iterative PC(-1)	0.046	1.365	1.549	1.713	1.124	1.755
Iterative PC(-2)	0.041	0.716	0.974	0.663	0.665	1.115

	MSE
Scoring (ML)	14.139
Ridge (HM)	6.350
Schaefer PC(-1)	7.533
Schaefer PC(-2)	10.001
Iterative PC(-1)	7.571
Iterative PC(-2)	10.390

Table 14. LOGISTIC SIMULATION RESULTS

Response = Bernoulli $N = 45$ $p + 1 = 4$ Severe Ill-conditioning

Conditional Analysis on 995 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.805$
$\lambda_1 = 1.595$
$\lambda_2 = .088$
$\lambda_3 = .013$

TRUEB = (−.5, −2, 1, 1)

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$
Scoring (ML)	-.559	-2.548	1.507	.905
Ridge (HM)	-.520	-1.996	.901	1.055
Schaefer PC(-1)	-.527	-2.394	1.132	1.103
Schaefer PC(-2)	-.555	.335	.381	.387
Iterative PC(-1)	-.543	-2.871	1.178	1.151
Iterative PC(-2)	-.597	.352	.377	.383

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$
Scoring (ML)	0.122	8.520	292.730	295.356
Ridge (HM)	0.105	5.717	45.255	47.085
Schaefer PC(-1)	0.109	7.123	1.633	1.639
Schaefer PC(-2)	0.084	0.279	0.619	0.630
Iterative PC(-1)	0.118	7.647	1.709	1.729
Iterative PC(-2)	0.096	0.335	0.725	0.739

	MSE
Scoring (ML)	597.198
Ridge (HM)	98.176
Schaefer PC(-1)	10.688
Schaefer PC(-2)	7.827
Iterative PC(-1)	11.581
Iterative PC(-2)	8.196

Table 15. LOGISTIC SIMULATION RESULTS

Response = Bernoulli $N = 45$ $p + 1 = 4$ Moderate Ill-conditioning

Conditional Analysis on 994 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.184$
$\lambda_1 = 1.222$
$\lambda_2 = .521$
$\lambda_3 = .073$

TRUEB $= (-.5, -2, 1, 1)$

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$
Scoring (ML)	-.560	-2.259	1.207	1.001
Ridge (HM)	-.593	-1.338	.669	.589
Schaefer PC(-1)	-.516	-2.201	.669	.567
Schaefer PC(-2)	-.570	-.551	.606	-.575
Iterative PC(-1)	-.535	-2.240	.669	.548
Iterative PC(-2)	-.509	-.589	.682	-.654

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$
Scoring (ML)	0.134	8.521	8.842	7.371
Ridge (HM)	0.100	2.891	2.592	2.357
Schaefer PC(-1)	0.103	4.137	1.359	3.111
Schaefer PC(-2)	0.089	0.618	0.871	0.755
Iterative PC(-1)	0.113	3.769	1.598	3.541
Iterative PC(-2)	0.105	0.707	1.035	0.931

	MSE
Scoring (ML)	24.583
Ridge (HM)	8.749
Schaefer PC(-1)	8.547
Schaefer PC(-2)	7.369
Iterative PC(-1)	9.294
Iterative PC(-2)	7.899

Table 16. LOGISTIC SIMULATION RESULTS

Response = Bernoulli $N = 45$ $p + 1 = 6$ Severe Ill-conditioning

Conditional Analysis on 972 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.564$
$\lambda_1 = 1.722$
$\lambda_2 = 1.068$
$\lambda_3 = .582$
$\lambda_4 = .057$
$\lambda_5 = .008$

TRUEB = (−.5, −2, 1, 1, −1, 1)

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$	$\bar{\beta}_4$	$\bar{\beta}_5$
Scoring (ML)	-.622	-2.798	.507	2.239	-.876	.959
Ridge (HM)	-.558	-2.133	.954	1.097	-.830	.836
Schaefer PC(-1)	-.582	-2.421	1.143	1.342	-.918	1.188
Schaefer PC(-2)	-.580	-.896	.352	.383	-.956	-.223
Iterative PC(-1)	-.601	-2.704	1.182	1.378	-.915	1.064
Iterative PC(-2)	-.536	-.973	.515	.542	-.997	-.299

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$	$s^2(\beta_4)$	$s^2(\beta_5)$
Scoring (ML)	0.165	10.632	519.081	491.729	10.843	16.981
Ridge (HM)	0.127	6.383	47.865	45.102	5.880	7.159
Schaefer PC(-1)	0.138	9.211	2.077	2.584	8.528	9.329
Schaefer PC(-2)	0.103	2.486	0.958	0.988	2.856	1.818
Iterative PC(-1)	0.148	9.905	2.223	2.763	8.852	10.574
Iterative PC(-2)	0.123	3.578	1.121	1.147	3.253	2.106

MSE

Scoring (ML)	1051.990
Ridge (HM)	112.404
Schaefer PC(-1)	32.338
Schaefer PC(-2)	13.523
Iterative PC(-1)	35.158
Iterative PC(-2)	14.126

Table 17. LOGISTIC SIMULATION RESULTS

Response = Bernoulli $N = 45$ $p + 1 = 6$ Moderate Ill-conditioning

Conditional Analysis on 979 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.252$
$\lambda_1 = 1.516$
$\lambda_2 = 1.078$
$\lambda_3 = .900$
$\lambda_4 = .310$
$\lambda_5 = .044$

TRUEB = (−.5, −2, 1, 1, −1, 1)

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$	$\bar{\beta}_4$	$\bar{\beta}_5$
Scoring (ML)	-.602	-2.852	1.248	1.123	-1.089	1.071
Ridge (HM)	-.505	-1.351	.668	.535	-.611	.562
Schaefer PC(-1)	-.536	-2.059	.980	.806	-.872	.873
Schaefer PC(-2)	-.586	-.784	.520	-.510	-.529	-.529
Iterative PC(-1)	-.560	-2.170	1.039	.821	-.919	.926
Iterative PC(-2)	-.538	-.875	.599	-.579	-.541	-.519

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$	$s^2(\beta_4)$	$s^2(\beta_5)$
Scoring (ML)	0.157	11.910	11.065	10.509	9.914	10.779
Ridge (HM)	0.101	3.589	2.704	2.747	2.537	2.800
Schaefer PC(-1)	0.109	4.126	4.158	4.953	4.149	5.048
Schaefer PC(-2)	0.093	1.523	1.808	1.767	2.298	1.666
Iterative PC(-1)	0.121	4.915	3.757	5.258	4.060	5.619
Iterative PC(-2)	0.111	1.925	2.164	2.233	2.459	2.229

	MSE
Scoring (ML)	54.540
Ridge (HM)	15.271
Schaefer PC(-1)	21.619
Schaefer PC(-2)	15.203
Iterative PC(-1)	23.709
Iterative PC(-2)	17.855

Table 18. POISSON SIMULATION RESULTS

Response = Poisson $N = 17$ $p + 1 = 4$ Severe Ill-conditioning

Conditional Analysis on 503 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.331$
$\lambda_1 = 1.632$
$\lambda_2 = .031$
$\lambda_3 = .006$

TRUEB = (−.5, −2, 1, 1)

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$
Scoring (ML)	-.809	-1.617	2.384	-.559
Ridge (HM)	-.632	-1.195	1.079	.257
Schaefer PC(-1)	-.707	-1.669	.809	.963
Schaefer PC(-2)	-.374	.095	.063	.076
Iterative PC(-1)	-.719	-1.568	.789	.952
Iterative PC(-2)	-.571	.217	.124	.126

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$
Scoring (ML)	0.215	2.879	35.745	37.345
Ridge (HM)	0.125	1.204	4.749	5.277
Schaefer PC(-1)	0.156	1.747	0.801	0.675
Schaefer PC(-2)	0.078	0.125	0.185	0.154
Iterative PC(-1)	0.142	1.667	0.695	0.641
Iterative PC(-2)	0.092	0.161	0.207	0.171

	MSE
Scoring (ML)	80.376
Ridge (HM)	12.579
Schaefer PC(-1)	3.569
Schaefer PC(-2)	6.676
Iterative PC(-1)	3.522
Iterative PC(-2)	7.082

Table 19. POISSON SIMULATION RESULTS

Response = Poisson $N = 17$ $p + 1 = 4$ Moderate Ill-conditioning

Conditional Analysis on 689 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.039$
$\lambda_1 = 1.279$
$\lambda_2 = .676$
$\lambda_3 = .005$

TRUEB = (−.5, −2, 1, 1)

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$
Scoring (ML)	-.830	-2.097	1.109	.531
Ridge (HM)	-.533	-1.088	.561	.239
Schaefer PC(-1)	-.596	-1.119	.209	-.140
Schaefer PC(-2)	-.299	-.102	.174	-.198
Iterative PC(-1)	-.636	-1.265	.235	-.217
Iterative PC(-2)	-.518	-.081	.296	-.394

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$
Scoring (ML)	0.296	3.191	4.162	2.468
Ridge (HM)	0.114	1.163	0.979	0.739
Schaefer PC(-1)	0.117	1.562	0.740	0.797
Schaefer PC(-2)	0.076	0.167	0.209	0.323
Iterative PC(-1)	0.125	1.763	0.918	0.994
Iterative PC(-2)	0.079	0.208	0.203	0.289

	MSE
Scoring (ML)	10.668
Ridge (HM)	4.999
Schaefer PC(-1)	5.816
Schaefer PC(-2)	6.536
Iterative PC(-1)	6.525
Iterative PC(-2)	6.899

Table 20. POISSON SIMULATION RESULTS

Response = Poisson $N = 17$ $p + 1 = 6$ Severe Ill-conditioning

Conditional Analysis on 142 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 3.224$
$\lambda_1 = 1.966$
$\lambda_2 = .533$
$\lambda_3 = .243$
$\lambda_4 = .032$
$\lambda_5 = .003$

TRUEB $= (-.5, -2, 1, 1, -1, 1)$

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$	$\bar{\beta}_4$	$\bar{\beta}_5$
Scoring (ML)	-1.073	-1.939	2.449	-.796	-1.567	-.229
Ridge (HM)	-.691	-1.161	1.219	-.021	-1.078	.076
Schaefer PC(-1)	-.761	-1.359	.580	.719	-1.250	.009
Schaefer PC(-2)	-.501	-.762	.512	.651	-.796	.541
Iterative PC(-1)	-.797	-1.297	.664	.807	-1.219	-.009
Iterative PC(-2)	-.669	-.776	.583	.757	-.971	.388

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$	$s^2(\beta_4)$	$s^2(\beta_5)$
Scoring (ML)	0.711	7.272	86.207	92.529	4.503	9.300
Ridge (HM)	0.181	1.971	5.205	6.629	1.522	1.630
Schaefer PC(-1)	0.188	3.523	0.804	0.988	2.336	2.865
Schaefer PC(-2)	0.076	.595	0.376	0.505	.962	.562
Iterative PC(-1)	0.173	3.519	0.714	1.018	2.279	2.285
Iterative PC(-2)	0.075	.592	0.505	0.536	1.003	.709

	MSE
Scoring (ML)	208.727
Ridge (HM)	20.129
Schaefer PC(-1)	11.582
Schaefer PC(-2)	5.348
Iterative PC(-1)	11.187
Iterative PC(-2)	5.246

Table 21. POISSON SIMULATION RESULTS

Response = Poisson $N = 17$ $p + 1 = 6$ Moderate Ill-conditioning

Conditional Analysis on 418 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.849$
$\lambda_1 = 1.539$
$\lambda_2 = 1.102$
$\lambda_3 = .706$
$\lambda_4 = .297$
$\lambda_5 = .008$

TRUEB = (−.5, −2, 1, 1, −1, 1)

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$	$\bar{\beta}_4$	$\bar{\beta}_5$
Scoring (ML)	-1.170	-2.396	1.518	.562	-1.165	1.189
Ridge (HM)	-.574	-1.200	.589	.139	-.512	.533
Schaefer PC(-1)	-.688	-1.784	.559	.066	-.124	.692
Schaefer PC(-2)	-.520	-1.094	.115	-.023	-.341	.372
Iterative PC(-1)	-.809	-1.955	.373	-.020	-.236	.692
Iterative PC(-2)	-.664	-1.201	.121	-.220	-.559	.569

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$	$s^2(\beta_4)$	$s^2(\beta_5)$
Scoring (ML)	1.511	5.011	15.218	7.347	8.311	5.614
Ridge (HM)	0.108	1.005	1.199	0.897	0.980	0.881
Schaefer PC(-1)	0.174	1.627	1.329	1.155	1.658	1.776
Schaefer PC(-2)	0.099	0.876	0.502	0.620	0.531	0.780
Iterative PC(-1)	0.168	1.533	1.513	1.149	1.566	1.844
Iterative PC(-2)	0.104	1.171	0.673	0.759	0.596	0.940

	MSE
Scoring (ML)	44.542
Ridge (HM)	7.281
Schaefer PC(-1)	9.828
Schaefer PC(-2)	6.895
Iterative PC(-1)	9.882
Iterative PC(-2)	7.744

Table 22. LOGISTIC SIMULATION RESULTS

Response = Bernoulli $N = 17$ $p + 1 = 4$ Severe Ill-conditioning

Conditional Analysis on 791 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.543$
$\lambda_1 = 1.317$
$\lambda_2 = .118$
$\lambda_3 = .023$

TRUEB = (−.5, −2, 1, 1)

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$
Scoring (ML)	-.695	-3.532	1.112	1.848
Ridge (HM)	-.549	-1.989	1.205	.794
Schaefer PC(-1)	-.579	-2.343	1.022	1.253
Schaefer PC(-2)	-.524	.225	.284	.279
Iterative PC(-1)	-.596	-2.561	1.118	1.366
Iterative PC(-2)	-.588	.268	.292	.285

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$
Scoring (ML)	0.792	21.885	393.988	352.108
Ridge (HM)	0.339	7.573	38.239	33.921
Schaefer PC(-1)	0.516	9.563	2.532	3.581
Schaefer PC(-2)	0.204	0.315	0.539	0.549
Iterative PC(-1)	0.579	12.129	3.509	3.433
Iterative PC(-2)	0.275	0.335	0.744	0.770

	MSE
Scoring (ML)	770.607
Ridge (HM)	80.159
Schaefer PC(-1)	15.680
Schaefer PC(-2)	7.597
Iterative PC(-1)	19.322
Iterative PC(-2)	8.281

Table 23. LOGISTIC SIMULATION RESULTS

Response = Bernoulli $N = 17$ $p + 1 = 4$ Moderate Ill-conditioning

Conditional Analysis on 962 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.307$
$\lambda_1 = .868$
$\lambda_2 = .743$
$\lambda_3 = .082$

TRUEB $= (-.5, -2, 1, 1)$

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$
Scoring (ML)	-.783	-3.456	1.617	1.252
Ridge (HM)	-.569	-1.396	.687	.508
Schaefer PC(-1)	-.388	- .673	.253	-.130
Schaefer PC(-2)	-.352	-.244	.261	-.333
Iterative PC(-1)	-.577	- .790	.372	-.265
Iterative PC(-2)	-.561	-.301	.385	-.538

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$
Scoring (ML)	1.225	30.639	21.749	19.292
Ridge (HM)	0.258	3.929	3.488	3.549
Schaefer PC(-1)	0.197	1.677	1.337	1.512
Schaefer PC(-2)	0.160	0.506	0.509	0.617
Iterative PC(-1)	0.287	2.049	1.282	1.525
Iterative PC(-2)	0.248	0.531	0.630	0.723

	MSE
Scoring (ML)	75.268
Ridge (HM)	11.230
Schaefer PC(-1)	8.331
Schaefer PC(-2)	7.121
Iterative PC(-1)	8.601
Iterative PC(-2)	7.766

Table 24. LOGISTIC SIMULATION RESULTS

Response = Bernoulli $N = 17$ $p + 1 = 6$ Severe Ill-conditioning

Conditional Analysis on 450 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 3.511$
$\lambda_1 = 1.530$
$\lambda_2 = .785$
$\lambda_3 = .399$
$\lambda_4 = .265$
$\lambda_5 = .009$

TRUEB = (−.5, −2, 1, 1, −1, 1)

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$	$\bar{\beta}_4$	$\bar{\beta}_5$
Scoring (ML)	-.761	-3.520	-2.325	5.392	-1.184	1.364
Ridge (HM)	-.506	-1.809	.588	1.345	-.874	.905
Schaefer PC(-1)	-.608	-2.564	1.166	1.369	-1.040	1.169
Schaefer PC(-2)	-.504	-1.256	.643	.763	-.819	.767
Iterative PC(-1)	-.619	-3.541	1.228	1.549	-1.053	1.151
Iterative PC(-2)	-.568	-1.556	.724	.864	-1.225	.864

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$	$s^2(\beta_4)$	$s^2(\beta_5)$
Scoring (ML)	1.039	38.731	490.073	469.371	37.203	40.695
Ridge (HM)	0.298	8.065	20.308	19.798	7.354	7.126
Schaefer PC(-1)	0.529	21.311	4.074	6.799	20.837	19.570
Schaefer PC(-2)	0.232	3.594	1.506	1.517	2.816	3.582
Iterative PC(-1)	0.571	18.256	3.946	6.226	18.517	23.997
Iterative PC(-2)	0.337	4.911	1.922	2.000	3.579	4.019

	MSE
Scoring (ML)	1109.710
Ridge (HM)	63.292
Schaefer PC(-1)	73.776
Schaefer PC(-2)	13.579
Iterative PC(-1)	72.491
Iterative PC(-2)	16.932

Table 25. LOGISTIC SIMULATION RESULTS

Response = Bernoulli $N = 17$ $p + 1 = 6$ Moderate Ill-conditioning

Conditional Analysis on 789 / 1000 repetitions

Eigenvalue Structure

$\lambda_0 = 2.857$
$\lambda_1 = 1.570$
$\lambda_2 = .846$
$\lambda_3 = .577$
$\lambda_4 = .547$
$\lambda_5 = .130$

TRUEB = (−.5, −2, 1, 1, −1, 1)

Average	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$	$\bar{\beta}_4$	$\bar{\beta}_5$
Scoring (ML)	-.897	-3.719	1.582	1.860	-1.512	1.670
Ridge (HM)	-.515	-1.282	.554	.516	-.655	.538
Schaefer PC(-1)	-.383	-1.376	.082	-.012	-.519	.293
Schaefer PC(-2)	-.307	-.717	.067	-.297	-.582	.219
Iterative PC(-1)	-.569	-1.824	-.068	-.138	-.704	.294
Iterative PC(-2)	-.577	-.953	.138	-.517	-.602	.270

Variance	$s^2(\beta_0)$	$s^2(\beta_1)$	$s^2(\beta_2)$	$s^2(\beta_3)$	$s^2(\beta_4)$	$s^2(\beta_5)$
Scoring (ML)	1.699	47.120	46.141	51.585	31.299	49.014
Ridge (HM)	0.211	3.545	2.892	3.490	2.843	3.273
Schaefer PC(-1)	0.245	3.494	2.852	2.874	3.586	3.519
Schaefer PC(-2)	0.163	1.239	1.150	1.119	1.582	1.251
Iterative PC(-1)	0.317	3.790	3.197	2.894	4.147	3.994
Iterative PC(-2)	0.230	1.567	1.389	1.358	1.686	1.519

	MSE
Scoring (ML)	231.763
Ridge (HM)	17.048
Schaefer PC(-1)	19.272
Schaefer PC(-2)	11.519
Iterative PC(-1)	21.898
Iterative PC(-2)	12.088

Chapter IX

CONCLUSIONS, COMMENTS AND AREAS OF FUTURE RESEARCH

It is common practice among statisticians to impose various transformations to implement least squares. Least squares has become a classical and extremely popular method for solving statistical problems. In certain circumstances, perhaps least squares estimation has reached a point of overuse, particularly with noncontinuous or heavy tailed or nonsymmetric response distributions. Pregibon (1979) points out that because of the ease with which least squares can process data, statisticians will often transform data to a somewhat continuous, short tailed and symmetric distribution with stable variance. And certainly re-expressions can be a very effective method for analyses. For example, the square root transformation for count data and the arcsin transformation for proportion data are well documented. However, by incorporating a rich variety of distributional forms for the response variable, the structure of the generalized linear model can often provide a practical and elegant alternative to that of least squares.

As demonstrated in the preceding chapters, maximum likelihood estimation of the regression parameters maintains asymptotic properties of unbiasedness, efficiency, consistency and normality. On the other hand, maximum likelihood cannot withstand large variances and low noncentrality parameters of estimated coefficients, among many other undesirable properties, in the presence of an ill-conditioned information matrix. This dissertation has suggested several alternate estimation techniques in the framework of the generalized linear model. Chapters 5 and 6 put forth extensions to generalize Schaefer's logistic ridge estimator and one step adjustment principal component estimator. In addition, the author has developed an iterative principal component technique which can be used, if for nothing else, as a resort if in fact maximum likelihood does not converge.

In as much as the alternate parameter estimators, mentioned above, are adjustments to maximum likelihood, asymptotic unbiasedness no longer holds. However, variance of these estimators can be substantially reduced. As indicated by the simulation results in chapter 8, asymptotically biased estimators are clear winners in reference to mean square error when compared to the asymptotically unbiased maximum likelihood competitor, even with moderately ill-conditioned information.

As for prediction in the response, maximum likelihood is adequate for predicting at internal mainstream data combinations of the X space even with severely ill-conditioned information. When a researcher is interested in predicting outside the mainstream of internal data, then prediction can be atrocious. In the case that the researcher is not constrained to some theoretical model and prediction is of primary concern, then perhaps wary variable deletion via the diagnostics given in Chapter 4 is the best tactic. On the other hand, if given a theoretical model with ill-conditioned information, any of the asymptotically biased estimation approaches are a clear choice over maximum likelihood in terms of prediction variance.

Ground has been broken in terms of developments in choosing a shrinkage parameter for generalized ridge estimation. An extension to Schaefer's (1979) harmonic mean method has

been suggested. In addition, a C_p based method for shrinkage has been developed for prediction-oriented choices of d. Also, Tripp's (1983) DF-trace has been generalized. Future study will include further developments of shrinkage parameter choice.

The class of generalized fractional principal component (GFPC) estimators, outlined in Chapter 7, attempts to place a very broad class of estimators under one common umbrella. GFPC incorporates a general link function, an entire class of response distributions, and an array of estimation techniques. In fact if the identity link is used with normal response data, the GFPC collapses into the framework of fractional principal component (FPC) estimators given in Lee's (1986) dissertation. Further research will be devoted to this area of GFPC in much the same vein as Hocking, Speed, and Lynn (1976).

The author will continue research in the area of asymptotically biased estimators of the generalized linear model. One possible extension is observing the biased estimators from an iterative geometric point of view. The author also plans to continue developing new biased estimators. The first developments will be in the direction of a latent root estimator which incorporates the eigenvalues of $A'A$, where A is the matrix $K^{-1/2}X$ augmented with the response vector $g(\mu)$. Further simulations need to be done. Extensions to existing software, such as *GLIM*, should be pursued.

Chapter X

BIBLIOGRAPHY

Albert, A. and Anderson, J. A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. ***Biometrika*** , 71, 1, 1-10.

* Akaike, H. (1974). A New Look at the Statistical Model Identification. ***IEEE Trans. Automat. Control***, AC-19, 716-723.

Andrews, D. F. and Herzberg, A. M. (1985). ***Data.*** Springer-Verlag: New York.

Bartlett, M. S. (1953). Approximate Confidence Intervals. ***Biometrika***, 40, 12-19.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). ***Regression Diagnostics: Influential Data and Sources of Collinearity.*** Wiley: New York.

Bickel, P. and Doksum, K. (1976). ***Mathematical Statistics: Basic Ideas and Selected Topics.*** Holden-Day: Oakland.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1976). ***Discrete Multivariate Analysis: Theory and Practice*** MIT Press: Cambridge, Mass.

Bradley, R. A. and Gart, J. J. (1962). The Asymptotic Properties of Maximum Likelihood Estimators when Sampling from Associated Populations. ***Biometrika***, 49, 205-214.

Burdick, Donald S. (1987). Personal Communication.

Capps, Oral Jr. (1985). Class Notes.

Cox, D. R. (1970). ***The Analysis of Binary Data*** . Metheun: London.

Cox, D. R. and Hinkley, D. V. (1974). ***Theoretical Statistics.*** Chapman and Hall: London.

Dobson, A. J. (1983). ***An Introduction to Statistical Modelling*** . Chapman and Hall: London.

Feller, W. (1966). ***An Introduction to Probability and its Applications, Vol. 2.*** Wiley: New York.

Finney, D. J. (1947). The Estimation from Individual Records of the Relationship Between Dose and Quantal Response. ***Biometrika***, 34, 320-334.

Finney, D. J. (1971). ***Probit Analysis, 3rd edition.*** Cambridge Univ. Press: Cambridge.

Fisher, R. A. (1935). The Case of Zero Survivors (appendix to a paper by C.I. Bliss). ***Annals of Applied Biology***, 22, 164-165.

GLIM System Release 3.77 Manual (1986). Royal Statistical Society.

Good, I. J. (1969). Some Applications of the Singular Value Decomposition of a Matrix. ***Technometrics***, 21, 215-222.

Graybill, F. A. (1976). ***Theory and Application of the Linear Model*** Wadsworth: Pacific Grove, CA.

Greenacre, M. J. (1984). ***Theory and Applications of Correspondence Analysis.*** Academic Press: London.

Haberman, S. J. (1978). ***Analysis of Qualitative Data.*** Academic Press: New York.

Hartree, D. R. (1952). ***Numerical Analysis.*** Oxford University Press: London.

Hill, R. C., Fomby, T. B., Johnson, S. R. (1977). Component Selection Norms for Principal Component Regression. ***Commun. Stat.***, A6, 309-334.

Hocking, R. R., Speed, F. M., Lynn, M. J. (1976). A Class of Biased Estimators in Linear Regression. ***Technometrics***, 18, No. 4, 425.

Hoerl A. E. and Kennard R. W. (1970a). Ridge Regression: Biased Estimation for Nonorthogonal Problems. ***Technometrics***, 12, 55-67.

Hoerl A. E. and Kennard R. W. (1970b). Ridge Regression: Applications to Nonorthogonal Problems. ***Technometrics***, 12, 69-82.

Jennings, D. E. (1986). Outliers and Residual Distributions in Logistic Regression. ***JASA***, Vol. 81, No. 396, p.987.

Jolliffe, I. T. (1986). ***Principal Component Analysis.*** Springer-Verlag: New York.

Kendall, M. G. (1957). ***A Course in Multivariate Analysis.*** Griffin: London.

Kendall, M. and Stuart A. (1973). ***The Advanced Theory of Statistics, Vol. 2, 3rd ed.*** Hafner: New York.

Lawless, J. F. and Singhal, K. (1978). Efficient Screening of Nonnormal Regression Models. ***Biometrics***, 34, 318-327.

Lee, W. (1986). ***Fractional Principal Components Regression: A General Approach to Biased Estimators.*** Ph.D. Dissertation: Virginia Polytechnic Inst. and State Univ.

Mallows, C. L. (1973). Some comments on C_p. ***Technometrics***, Vol. 15, No. 4, 661-675.

Marquardt, D. W. (1970). Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. ***Technometrics***, 12, 591-612.

McCullagh, P. and Nelder, J. A. (1983). ***Generalized Linear Models.*** Chapman and Hall: New York.

McGilchrist, Charles A. (1987). Personal Correspondence and Class Notes.

Myers, Raymond H. (1985). Class Notes.

Myers, Raymond H. (1986). ***Classical and Modern Regression with Applications.*** Duxbury Press: Boston.

Myers, Sharon (1984). Personal Communication.

Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. ***Journal of the Royal Statistical Society***, Vol. 135, No. 3.

Pindyck, R. S. and Rubinfeld, D. L. (1981). ***Econometric Models and Economic Forecasts, 2nd ed.*** McGraw-Hill: New York.

Pregibon, D. (1979). ***Data Analytic Methods for Generalized Linear Models.*** Unpublished Ph.D. thesis: Univ. of Toronto.

Pregibon, D. (1981). Logistic Regression Diagnostics. ***The Annals of Statistics***, Vol. 9, No. 4, 705-724.

Rao, C. R. (1967). ***Linear Statistical Inference and Its Applications.*** Wiley: New York.

SAS Institute Inc. ***SAS User's Guide: Basics, 1982 Edition.*** Cary, NC: SAS Institute Inc., 1982. 923 pp.

SAS Institute Inc. **SAS/IML™** ***User's Guide, Version 5 Edition.*** Cary, NC: SAS Institute Inc., 1985. 300 pp.

Silvey, S. D. (1969). Multicollinearity and Imprecise Estimation. ***Journal of the Royal Statistical Society***, Series B, 31, 539-552.

Schaefer, R. L. (1979). ***Multicollinearity and Logistic Regression.*** Ph.D. Dissertation: Univ. of Michigan.

Schaefer, Roi, Wolfe (1984). A Ridge Logistic Estimator. ***Comm. St. - Th. Meth.***, 13(1), 99-114.

Schaefer, R. L. (1986). Alternative Estimators in Logistic Regression when the Data are Collinear. ***J. Statist. Comput. Simul.***, Vol. 25, 75-91.

Stein, C. M. (1960). Multiple Regression. ***Contributions to Probability and Statistics.*** Essays in Honor of Harold Hotelling, ed. I. Olkin, Stanford Univ. Press, 424-443.

SUGI Supplemental Library User's Guide, Version 5 Edition. Cary, NC: SAS Institute Inc., 1986. 662pp.

Tripp, R. E. (1983). ***Non-Stochastic Ridge Regression and the Effective Rank of the Regressors Matrix.*** Ph.D. Dissertation: Virginia Polytechnic Inst. and State Univ.

Vinod H. D. (1976). Applications of New Ridge Regression Methods to a Study of Bell System Scale Economics. ***J. Amer. Statistical Association***, 71, 356, 835-841.

Walker, S. H. and Duncan, D. B. (1967). Estimation of the Probability of an Event as a Function of Several Independent Variables. ***Biometrika***, 54, 1 and 2, 167.

Webster, J. T., Gunst, R. F., and Mason, R. L. (1974). Latent Root Regression Analysis, ***Technometrics***, 16, 513-522.

Appendix A

Data Set Used in Example

Determinants of Cancer Remission

The following data was taken from *SAS, SUGI* Supplementary Guide (1986). From the example below, data on cancer patients are analyzed to determine if the patient characteristics associate with cancer remission. Information was collected on the following variables.

Y	=	1 if cancer remission 0 if no cancer remission
X1	=	Cell index
X2	=	Smear index
X3	=	Infil index
X4	=	LI index
X5	=	Temperature

DATA

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
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

Appendix B
Simulation Program in SAS Proc Matrix

```
PROC MATRIX;
*
FIX PARAMETERS REGARDING VARIOUS CRITERIA;
*;

ITERAT=1000;SEED=959145; CONVERGE=.00001;
ITER=55;CONV=ITER-1;UPPER=10000;LOWER=.00001;MAX=15;MIN=-15;
*
TRUEB IS THE TRUE PARMETER VECTOR;
*;
TRUEB=-.5/-2 /1/1;
FETCH XY DATA=ONE;
N=NROW(XY);ONES=J(N,1,1);
*;
*NOTE: PULL OUT VARIABLES OF INTEREST FROM THE XY MATRIX;
*;
X1=XY(,1);
X2=XY(,2);
X3=XY(,3);
X4=XY(,4);
X5=XY(,5);
X6=XY(,6);
X7=XY(,7);
X8=(X1+X4)+X3;
X9=X1+X3+X7;
*
CONSTRUCT X OF INTEREST;
*;
X=X1||X9||X8     ;
*
CENTER AND SCALE THE X MATRIX;
*;
SUM=X(+,);MEAN=SUM#/N;
X=X-J(N,1)*MEAN;
SS=SQRT((X#X)(+,)); PRINT SS;
XCS=X#/(J(N,1)*SS);
*
AUGMENT THE CONSTANT TERM;
*;
X=ONES||XCS; PRINT X;
*
*NOTE: DEFINE THE DIMENSIONS;
*;
P=NCOL(X);P1=P-1;P2=P-2;P3=P-3;IDEN=I(P);
ONE=J(P,1,1); ONE1=J(P1,1,1); ONE2=J(P2,1,1);
*
STARTING VALUES FOR VARIOUS COUNTERS;
*;
BSUM=J(6,P,0);BSQ=BSUM;NOCONV1=0;NOCONV2=0;NOCONV3=0;
OLIERU1=0;OLIERL1=0;
OLIERU2=0;OLIERL2=0;
OLIERU3=0;OLIERL3=0;
COUNT1=0;COUNT2=0;COUNT3=0;
*
ETA IS THE LINEAR COMBINATION OF X AND PARAMETER VECTOR;
*;
ETA=X*TRUEB;PRINT ETA;
*
MAX AND MIN CONSTRAIN THE ARGUMENT FOR EXPONENTIATING;
*;
  DO LP1=1 TO N;
   IF ETA(LP1,1)<MIN THEN ETA(LP1,1)=MIN;
```

```
124           IF ETA(LP1,1)>MAX THEN ETA(LP1,1)=MAX;
125         END;
126       *
127       POISSON REGRESSION
128       CONSTRUCT THE MEAN VIA THE NATURAL LINK FUNCTION OF THE GLM;
129       *;
130       LAMBDA=EXP(ETA);
131       *
132       V IS THE DIAGONAL MATRIX OF WEIGHTS;
133       *;
134       V=DIAG(LAMBDA);
135       *
136       XVX IS THE INFORMATION MATRIX;
137       *;
138       XVX=X'*V*X;
139       IXVX=INV(XVX); PRINT IXVX;
140       EIGEN L M XVX;  PRINT L ;
141       L=L(1:2,*);M=M(,1:2);VPC=M*(INV(DIAG(L)))*M';
142       PRINT VPC;
143       *
144       CENTER AND SCALE SQRT(V)X MATRIX FOR DIAGNOSTICS;
145       *;
146       T=SQRT(V);T=T#X;
147       OUTPUT T  OUT=NEW;
148       SUM=T(+,);MEAN=SUM#/N;
149       T=T-J(N,1)*MEAN;
150       SS=SQRT((T#T)(+,));
151       T=T#/(J(N,1)*SS);
152       T=T'*T;PRINT T;
153       *
154       SPECTRAL DECOMPOSITION OF C-S INFORMATION FOR CONDITION INDEX;
155       *;
156       EIGEN LW MW T;PRINT LW;
157       *
158       START OF MAJOR DO LOOP FOR GENERATION OF DATA;
159       *;
160       DO LUPE=1 TO ITERAT;
161       Y=RANPOI(SEED,LAMBDA);
162       *;
163       *NOTE: SET STARTING VALUES;
164       *;
165       BETAML=0*ONE;
166       V=.5#I(N);
167       *;
168       *MAXIMUM LIKELIHOOD ESTIMATION VIA METHOD OF SCORING;
169       *NOTE: SET NUMBER OF ITERATIONS FOR ML;
170       *;
171       DO LP2=1 TO ITER;
172       INFORM=X'*V*X;IINFORM=INV(INFORM);
173       ETAML=X*BETAML;
174           DO LP3=1 TO N;
175            IF ETAML(LP3,1) > MAX THEN ETAML(LP3,1)=MAX;
176            IF ETAML(LP3,1) < MIN THEN ETAML(LP3,1)=MIN;
177           END;
178       MU=EXP(ETAML);
179       V=DIAG(MU);
180       PBETAML=BETAML;
```

```
181         BETAML=PBETAML+ IINFORM*X'*(Y-MU);
182         *;
183         *NOTE: SET DESIRED CONVERGENCE CRITERIA;
184         *;
185         IF ABS((BETAML-PBETAML)#/BETAML) < CONVERGE*ONE
186         THEN GO TO LABEL;
187         END;
188         LABEL:
189         *
190         END OF MAXIMUM LIKELIHOOD, CHECK ML BEHAVIOR;
191         *;
192         IF LP2 > CONV THEN BETAML=0*BETAML;
193         CUE1=BETAML'*BETAML;
194         IF CUE1 > UPPER THEN BETAML=0*BETAML;
195         IF CUE1 < LOWER THEN BETAML=0*BETAML;
196         IF CUE1 > UPPER THEN COUNT1=COUNT1+1;
197         IF CUE1 < LOWER THEN COUNT1=COUNT1+1;
198         IF LP2 > CONV THEN COUNT1=COUNT1+1;
199         IF (CUE1 > UPPER AND LP2 >CONV) THEN COUNT1=COUNT1-1;
200         IF (CUE1 < LOWER AND LP2 >CONV ) THEN COUNT1=COUNT1-1;
201         INFO=INFORM;
202         *;
203         *RIDGE ESTIMATION USING HARMONIC MEAN SHRINKAGE;
204         *;
205         SHRINK=P#/(BETAML'*BETAML);
206         RIDGE=INV(INFORM+( SHRINK#IDEN))*INFORM*BETAML;
207         *;
208         *SCHAEFER'S PC MINUS 1 DIMENSION;
209         *;
210         EIGEN L M INFORM;
211         L=L(1:P1,); L_=L; M=M(,1:P1); L=DIAG(L);L=INV(L);
212         IINFORM_=M*L*M';
213         BETASCH1= IINFORM_*INFORM*BETAML;
214         *;
215         *SCHAEFER'S PC MINUS 2 DIMENSION;
216         *;
217         L=L_(1:P2,); M=M(,1:P2); L=DIAG(L);L=INV(L);
218         IINFORM_=M*L*M';
219         BETASCH2= IINFORM_*INFORM*BETAML;
220         *;
221         *ITERATIVE PC APPROACH MINUS 1 DIMENSION;
222         *;
223         INFORM=INFO;
224         EIGEN L M INFORM;
225         L=L(1:P1,);M=M(,1:P1);
226         Z=X*M;
227         EGVL=DIAG(L);IEGVL=INV(EGVL);
228         ALPHA=0*M'*BETAML;
229         DO LP4=1 TO  ITER;
230         ETAPC=Z*ALPHA;
231              DO LP5=1 TO N;
232               IF ETAPC(LP5,1) > MAX THEN ETAPC(LP5,1)=MAX;
233               IF ETAPC(LP5,1) < MIN THEN ETAPC(LP5,1)=MIN;
234              END;
235         MUPC=EXP(ETAPC);
236         ALPHAP=ALPHA;
237         ALPHA=ALPHA+(IEGVL*Z'*(Y-MUPC));
```

```
238         IF ABS((ALPHA-ALPHAP)#/ALPHA) < CONVERGE*ONE1
239         THEN GO TO LABEL1;
240         END; LABEL1:
241         BETAPC1=M*ALPHA;
242         *
243         CHECK ITERATIVE PC BEHAVIOR;
244         *;
245         IF LP4 > CONV THEN BETAPC1=0*BETAML;
246         CUE2=BETAPC1'*BETAPC1;
247         IF CUE2 > UPPER THEN BETAPC1=0*BETAML;
248         IF CUE2 < LOWER THEN BETAPC1=0*BETAML;
249         IF CUE2 > UPPER THEN COUNT2=COUNT2+1;
250         IF CUE2 < LOWER THEN COUNT2=COUNT2+1;
251         IF LP4 > CONV THEN COUNT2=COUNT2+1;
252         IF (CUE2 > UPPER AND LP4 >CONV) THEN COUNT2=COUNT2-1;
253         IF (CUE2 < LOWER AND LP4 >CONV ) THEN COUNT2=COUNT2-1;
254         *;
255         *ITERATIVE PC APPROACH MINUS 2 DIMENSIONS;
256         *;
257         L=L(1:P2,);M=M(,1:P2);
258         Z=X*M;
259         ALPHA=0*M'*BETAML;
260         EGVAL=DIAG(L); IEGVAL=INV(EGVAL);
261         DO LP6=1 TO  ITER;
262         ETAPC=Z*ALPHA;
263             DO LP7=1 TO N;
264              IF ETAPC(LP7,1) > MAX THEN ETAPC(LP7,1)=MAX;
265              IF ETAPC(LP7,1) < MIN THEN ETAPC(LP7,1)=MIN;
266             END;
267         MUPC=EXP(ETAPC);
268         ALPHAP=ALPHA;
269         ALPHA=ALPHA+(IEGVAL*Z'*(Y-MUPC));
270         IF ABS((ALPHA-ALPHAP)#/ALPHA) < CONVERGE*ONE2
271         THEN GO TO LABEL2;
272         END; LABEL2:
273         BETAPC2=M*ALPHA;
274         *
275         CHECK ITERATIVE PC BEHAVIOR;
276         *;
277         IF LP6 > CONV THEN BETAPC2=0*BETAML;
278         CUE3=BETAPC2'*BETAPC2;
279         IF CUE3 > UPPER THEN BETAPC2=0*BETAML;
280         IF CUE3 < LOWER THEN BETAPC2=0*BETAML;
281         IF CUE3 > UPPER THEN COUNT3=COUNT3+1;
282         IF CUE3 < LOWER THEN COUNT3=COUNT3+1;
283         IF LP6 > CONV THEN COUNT3=COUNT3+1;
284         IF (CUE3 > UPPER AND LP6 >CONV) THEN COUNT3=COUNT3-1;
285         IF (CUE3 < LOWER AND LP6 >CONV) THEN COUNT3=COUNT3-1;
286         *
287         COUNTS MADE FOR CONDITIONAL ANALYSIS;
288         *;
289         IF CUE1 > UPPER THEN OLIERU1=OLIERU1+1;
290         IF CUE1 < LOWER THEN OLIERL1=OLIERL1+1;
291         IF CUE2 > UPPER THEN OLIERU2=OLIERU2+1;
292         IF CUE2 < LOWER THEN OLIERL2=OLIERL2+1;
293         IF CUE3 > UPPER THEN OLIERU3=OLIERU3+1;
294         IF CUE3 < LOWER THEN OLIERL3=OLIERL3+1;
```

```
295       IF LP2 > CONV THEN NOCONV1=NOCONV1+1;
296       IF LP4 > CONV THEN NOCONV2=NOCONV2+1;
297       IF LP6 > CONV THEN NOCONV3=NOCONV3+1;
298       *
299       CONSTRUCTION OF MEAN, VARIANCE, BIAS OF VARIOUS TECHNIQUES;
300       *;
301       BSUMP=BSUM;BSQP=BSQ;
302       BSUM(1,*)='BETAML';
303       BSUM(2,*)='RIDGE';
304       BSUM(3,*)='BETASCH1';
305       BSUM(4,*)='BETASCH2';
306       BSUM(5,*)='BETAPC1';
307       BSUM(6,*)='BETAPC2';
308       BSQ=BSUM#BSUM;
309       BSUM=BSUM+BSUMP;BSQ=BSQ+BSQP;
310       END;
311       PRINT NOCONV1 NOCONV2 NOCONV3  OLIERU1 OLIERL1 OLIERU2 OLIERL2
312       OLIERU3 OLIERL3 COUNT1 COUNT2 COUNT3;
313       BETABAR=J(6,P,0);
314       BETAVAR=J(6,P,0);
315       BETABAR(1,*)=BSUM(1,*)#/(ITERAT-COUNT1);
316       BETABAR(2,*)=BSUM(2,*)#/(ITERAT-COUNT1);
317       BETABAR(3,*)=BSUM(3,*)#/(ITERAT-COUNT1);
318       BETABAR(4,*)=BSUM(4,*)#/(ITERAT-COUNT1);
319       BETABAR(5,*)=BSUM(5,*)#/(ITERAT-COUNT2);
320       BETABAR(6,*)=BSUM(6,*)#/(ITERAT-COUNT3);
321       BETAVAR(1,*)=(1#/(ITERAT-COUNT1-1))#
322       (BSQ(1,*)-((ITERAT-COUNT1)#BETABAR(1,*)#BETABAR(1,*)));
323       BETAVAR(2,*)=(1#/(ITERAT-COUNT1-1))#
324       (BSQ(2,*)-((ITERAT-COUNT1)#BETABAR(2,*)#BETABAR(2,*)));
325       BETAVAR(3,*)=(1#/(ITERAT-COUNT1-1))#
326       (BSQ(3,*)-((ITERAT-COUNT1)#BETABAR(3,*)#BETABAR(3,*)));
327       BETAVAR(4,*)=(1#/(ITERAT-COUNT1-1))#
328       (BSQ(4,*)-((ITERAT-COUNT1)#BETABAR(4,*)#BETABAR(4,*)));
329       BETAVAR(5,*)=(1#/(ITERAT-COUNT2-1))#
330       (BSQ(5,*)-((ITERAT-COUNT2)#BETABAR(5,*)#BETABAR(5,*)));
331       BETAVAR(6,*)=(1#/(ITERAT-COUNT3-1))#
332       (BSQ(6,*)-((ITERAT-COUNT3)#BETABAR(6,*)#BETABAR(6,*)));
333       BIAS=BETABAR-(J(6,1,1)*TRUEB');
334       PRINT TRUEB;
335       PRINT BETABAR  BIAS;
336       PRINT BETAVAR;
337       BIAS2=BIAS*BIAS';BIAS2=DIAG(BIAS2);BIAS2=BIAS2*(J(6,1,1));
338       SUMMSE=(BETAVAR*(J(P,1,1)))+BIAS2;
339       PRINT SUMMSE;
340       *
341       VARIOUS PLOTS OF WEIGHTED VARIABLES;
342       *;

343       PROC PLOT DATA=NEW;
344       PLOT COL2*COL3;
```

The vita has been removed from the scanned document