COLLINEARITY IN GENERALIZED LINEAR MODELS

by

MURRAY J MACKINNON

M.Sc. University of Otago N.Z.

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

FACULTY OF GRADUATE STUDIES
Commerce and Business Administration

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April 1986

© Murray J Mackinnon, 1986

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the THE UNIVERSITY OF BRITISH COLUMBIA, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Commerce and Business Administration

THE UNIVERSITY OF BRITISH COLUMBIA
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date: April 1986

Abstract

The concept of collinearity for generalized linear models is introduced and compared to that for standard linear models. Two approaches for detecting collinearity are presented and shown to lead to the same diagnostic procedure. These are analysed for the Poisson, gamma, inverse Gaussian, p th order, binomial proportion and negative binomial models. A bound is derived for the degree of collinearity in a generalized linear model in terms of that of the standard linear model. Estimation methods based on ridge, prior likelihood and principal components are proposed, and briefly illustrated with a Monte Carlo simulation of a gamma model.

Table of Contents

<u>0.0 Introduction.</u>	1
<u>1.0 Collinearity in Standard Linear Models.</u>	3
1.1 Definition of Collinearity	3
1.2 Sources of Collinearity.	5
1.2.1 Large Pairwise Correlations	5
1.2.2 Data Collection	6
1.2.3 Model Specification	8
1.2.4 Overdefined Model	8
1.2.5 Outliers.	9
1.3 Effects of Collinearity.	10
1.3.1 Ill conditioning of X	10
1.3.2 Estimate Effects.	11
1.3.3 Inference Effects	13
1.3.4 Predictor Effects	14
<u>2.0 Collinearity in Generalized Linear Models</u>	16
2.1 Definition of a Generalized Linear Model	16
2.2 Definition of Collinearity in Generalized Linear Models	22
2.2.1 Linearisation of the Link Function.	23
2.2.2 Iteratively Reweighted Least Squares Approach.	25

2.2.3	Choice of Approach.	25
2.3	Relationship of the Standard Linear Model and the Generalized Model Collinearity Definitions . . .	26
2.4	Sources of Collinearity in a Generalized Linear Model.	29
2.5	Effects of Collinearity in a Generalized Linear Model.	32
2.5.1	Estimation Effects.	32
2.4.2	Inference and Predictor Effects	33
2.6	Appendix 2A Maximum Likelihood and the Generalized Linear Model	35
2.7	Appendix 2B Iteratively Reweighted Least Squares Algorithm.	40
3.0	<u>Diagnostics for Collinearity in Generalized Linear Models.</u>	42
3.1	Desirable Properties for Diagnostic Measures	42
3.2	Measures of Collinearity	43
3.3	Model Dependency in Collinear Systems.	46
4.0	<u>Estimation for Generalized Linear Models in the Presence of Collinearity</u>	50
4.1	Remedies for Collinear Sources	50
4.1.1	Large Pairwise Correlations in X.	51
4.1.2	Data Collection	51

LEAF IV OMITTED IN PAGE NUMBERING.

FEUILLET IV NON UNCLUS DANS LA PAGINATION.

4.1.3	Model Specification	51
4.1.4	Overdefined Model	52
4.1.5	Outliers.	52
4.2	Remedies for Collinear Effects	52
4.3	Estimation Methods	53
4.3.1	Ridge Estimation.	53
4.3.1.1	Standard Linear Case	54
4.3.1.2	General Linear Case.	55
4.3.2	Bayesian Estimation	60
4.3.2.1	Standard Linear Case	60
4.3.2.2	General Linear Case.	61
4.3.3	Principal Component Estimation.	62
4.3.3.1	Standard Linear Case	62
4.3.3.2	General Linear Case.	63
4.4	Appendix 4A Maximum Likelihood Estimates of the Posterior Distribution	65
5.0	<u>Illustrative Example.</u>	67
5.1	Introduction	67
5.2	Scope of the Simulation.	67
5.3	Generation of Collinear Data	68
5.3.1	Standard Linear Case.	68
5.3.2	General Linear Case	68
5.4	Simulation	71
5.4.1	Simulation Setup	71
5.4.2	Simulation Implementation and Difficulties	73

5.4.3 Simulation Results and Conclusions. . . .	74
<u>6.0 Summary and Conclusions</u>	81
<u>Bibliography</u>	83

0.0 Introduction

Collinearity has long been recognised as causing significant problems in the computation and estimation of the parameters in a standard linear model. However, the effects with a generalized linear model are relatively unexplored, with the exception of logistic regression [SCHAEFER,79]. It is the purpose of this thesis to seek a reasonable definition for collinearity in the generalized linear model, to examine its consequences and to propose estimation methods.

The first chapter identifies collinearity in the standard linear model and reviews its sources and effects, so as to provide an introduction and comparison for the general setting.

The second chapter seeks a reasonable definition for collinearity in the generalized linear model by considering two approaches. The first is based on a linearisation of the link function and the second, which is chosen, is motivated by the instability of matrices used in the iteratively reweighted least squares estimation. Using this definition, it is shown that collinearity is dependent on the model chosen.

The third chapter deals with collinearity diagnostics in the general setting. Desirable properties are proposed, and a criterion and identification scheme, extending the approach of [BELSEY,KUH,WELSCH,80], is used. To quantify the indeterminacy due to model dependency, weak bounds for the general criterion are developed in terms of the criterion for the standard linear model.

The fourth chapter investigates estimation methods. It opens by suggesting remedies for collinear sources and effects, while cautioning against the routine use of "mechanical" methods. Three methods are proposed. The ridge approach is shown to generalize, by following the logistic work of [SCHAEFER,79]. The second method is based on the prior likelihood philosophy of [EDWARDS,69], while the third generalizes the classical principal components solution.

The fifth chapter briefly illustrates the ideas of the previous chapters with a restricted Monte Carlo simulation of a gamma model. The problems and objectives associated with the construction of such a simulation are discussed. The simulation within its artificial setting demonstrates the advantages of a biased estimator in the presence of collinearity.

1.0 Collinearity in Standard Linear Models

Throughout the theoretical development of this thesis a constant parallel will be drawn between the standard linear model and the generalized linear model. Let the standard linear model of a response y in terms of predictor variables X be defined as follows.

Definition 1.1 Standard Linear Model

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

where y $n \times 1$ vector of responses

X $n \times p$ matrix of predictor variables

$$X = [x_1, \dots, x_p]$$

β $p \times 1$ vector of coefficients

ϵ $n \times 1$ vector of errors

where $\epsilon \sim N_n(0, \sigma^2 I)$

1.1 Definition of Collinearity

Collinearity, as considered in classical linear regression, is said to be present when :

- two or more predictor variables are highly correlated
- coefficient estimates are different in magnitude and/or sign from those hypothesised
- coefficient estimates are highly variable

- the sampled range of some predictors may be small
- matrices used in the computation of estimates are ill conditioned

But these are only possible symptoms or causes of the more basic problem of approximate linear dependencies amongst the columns of the predictor variables. This underlying problem is the foundation of the approach taken by [GUNST,83], and used here. With this approach it is possible to tie together diffuse causes such as deficient sampling range and outliers, and to give validity to symptoms such as highly variable estimates and sensitive row deletion statistics.

The following definition of collinearity is adapted from [GUNST,83].

Definition 1.2 Collinearity in the Standard Linear Model

$y = XB + \epsilon$ is present when :

For some suitably chosen $\delta \geq 0$ there exists a $p \times 1$ vector $c \neq 0$ such that

$$|Xc| \leq \delta |c|$$

It is to be noted that this definition only involves the data matrix X and makes no mention of the dependent variable

y, or the underlying model. So for the standard linear model, the detection of collinearity involves only the X matrix of predictors.

This definition is subjective because of the lack of guidance in the selection of δ . However it is no different from the others in the literature (eg. [BELSEY, KUH, WELSCH, 80] require subjective selection of the condition index of the matrix X). The [GUNST, 83] definition is preferred because the intuitive notions of collinearity listed above can be deduced more directly from it. The following sections based on [GUNST, 83] shows how these sources and effects can be explained in terms of the definition.

1.2 Sources of Collinearity

Collinearity can stem from any one of a number of sources, and all of these can be captured by Definition 1.2. Further, it is of considerable practical interest to have information on the source when trying to alleviate collinearity.

1.2.1 Large Pairwise Correlations

Suppose X has been scaled and centred so that $X^T X$ is in correlation form (ie. the correlation coefficient between x_j and x_k is r_{jk} , where x_j is the j th column of X). Let $X_{(j)}$ be the matrix with the j th column of X deleted and $\beta_{(j)}$ the

corresponding deleted predictor. Now from linear regression theory it is known (eg. [MONTGOMERY, PECK, 82] p430)

$$\beta_{(j)} = (X_{(j)}^T X_{(j)})^{-1} X_{(j)}^T x_j$$

Let c be defined as

$$X_c = x_j - X_{(j)} \beta_{(j)}$$

Now by the extra sum of squares principle

$$|X_c| = (1 - R_{j|})^{1/2}$$

where $R_{j|}$ is the coefficient of multiple correlation.

So if for some $\delta > 0$

$$(1 - R_{j|})^{1/2} \leq \delta (1 + \beta_{(j)}^T \beta_{(j)})$$

then we have $|X_c| \leq \delta |c|$ so a collinearity exists in terms of Definition 1.2.

Specifically if $p=2$ then $\beta_{(2)} = r_{12}$ and so c is given by

$$X_c = x_1 - r_{12} x_2 \quad \text{and}$$

$$|X_c| = (1 - r_{12}^2)^{1/2}$$

So if for some $\delta > 0$ we have

$$(1 - r_{12}^2)^{1/2} \leq \delta (1 + r_{12}^2)^{1/2} \quad \text{then}$$

$$|X_c| \leq \delta |c|$$

and so a collinearity exists as per Definition 1.2. In other words for r_{12} "large enough" a collinearity exists.

It is important to note though that only one dependency need be strong for all the correlations to tend towards unity. Specifically [BELSEY, KUH, WELSCH, 80] show when the singular value, connected with just one dependency, tends to zero all pairwise correlations tend to

unity. Hence, high correlations cannot be used to determine multiple collinearities.

1.2.2 Data Collection

If only a subset of the predictor variable space is sampled, collinearity can arise. For example, with the two variable case in the scatterplot below, the data is highly positively correlated and consequently collinear. This source of collinearity can be caused by extrinsic factors (eg. sampling deficiency due to chance, lack of available data) or more likely, as in an observational study, by intrinsic factors (eg. one variable is related to another, or one variable jointly influences two other variables).

1.2.3 Model Specification

Often in applied models there exist implicit constraints which manifest themselves in the data (eg. mixture models). However, they need not be exact dependencies. [BRADLEY, SRIVASTAVA, 79] consider the fitting of a low order polynomial in the case of a single predictor variable. They show that the correlation coefficient between the variables, corresponding to any two powers, can be extremely close to unity, if the differences in powers is even (even allowing for the improvement due to centring).

1.2.4 Overdefined Model

This is a special case of data collection, which occurs when there are more variables than observations (ie. $n < p$). The linear dependency of the collinearity can be seen as follows. Suppose the rank of X is r . Now $r \leq \min(n, p) = n$. By singular value decomposition there exist U orthogonal, V orthonormal and D diagonal such that

$$X = UDV^T$$

But U, V are orthogonal so $\text{rank}(D) = \text{rank}(X) = r$.

$$X = U \begin{bmatrix} D_{11} & 0 \\ 0 & 0 \end{bmatrix} V^T \quad \text{where } D_{11} : r \times r$$

$$X[V_1 \ V_2] = [U_1 \ U_2] \begin{bmatrix} D_{11} & 0 \\ 0 & 0 \end{bmatrix} \quad \begin{array}{l} \text{[postmultiply by } V] \\ \text{[} V \text{ orthonormal]} \end{array}$$

$$XV_2 = 0$$

So for any $\delta \geq 0$ there exists a vector c , namely $[V_2]$, such that $|Xc| = 0$ ie. $|Xc| \leq \delta |c|$.

1.2.5 Outliers

Collinearity can be induced by outliers (ie. observations which are atypical in the predictor variable space, as is shown in the following argument from [GUNST,83]. The linear dependency can be illustrated by considering the hypothetical example of $p=3$ predictor variables where x_{11} and x_{22} are atypically large in magnitude. Suppose k is fixed, $x_{11} = 0$, $x_{12} = k0$ and X is scaled so that the columns are of unit length. So

$$X = \begin{bmatrix} 1 & x_{11} & x_{1n} \\ 1 & x_{21} & x_{2n} \\ & \dots & \\ 1 & x_{n1} & x_{nn} \end{bmatrix} = \begin{bmatrix} 1 & \theta/\lambda_1 & k\theta/\lambda_n \\ 1 & x_{21}^*/\lambda_1 & x_{2n}^*/\lambda_n \\ & \dots & \\ 1 & x_{n1}^*/\lambda_1 & x_{nn}^*/\lambda_n \end{bmatrix}$$

where x_{11}^* , x_{1n}^* $i=2(1)n$ are the unscaled predictor variables and

$$\lambda_{1n} = \theta^2 + \sum_{i=1}^n x_{1i}^{*2} \quad \lambda_{nn} = k^2\theta^2 + \sum_{i=1}^n x_{1i}^{*2}$$

Consider $c^T = (0, 1, -1)$. Now

$$|Xc| = (\theta/\lambda_1 - k\theta/\lambda_n)^2 + \sum_{i=2}^n (x_{1i}^*/\lambda_1 - x_{1i}^*/\lambda_n)^2$$

Now for large θ , $\lambda_{1n} \approx \theta^2$ and $\lambda_{nn} \approx k^2\theta^2$. Hence the right hand side tends to 0. So for any $\delta > 0$ there is a value of θ such that $|Xc| \leq \delta|c|$. In other words the outlier, with θ sufficiently large, corresponds to a collinearity as in Definition 1.2.

1.3 Effects of Collinearity

This section illustrates, in the manner of [GUNST,83], the intuitive effects of collinearity using Definition 1.2.

1.3.1 Ill conditioning of X

It has been known for a long time that collinearity is intimately related to the conditioning of X. [KENDALL,57], [SILVEY,69] state that collinearities are associated with

small eigenvalues of $X^T X$, which is numerically equivalent to small singular values of X . This, when quantified more precisely, provides an alternative definition of collinearity, which [BELSEY, KUH, WELSCH, 80] advocate.

Specifically, suppose again that the columns of X are centred and scaled to unit length. Then by singular value decomposition

$$X = UDV^T$$

where $U = [u_1, \dots, u_p] : n \times p$ orthogonal

$V = [v_1, \dots, v_p] : p \times p$ eigenvectors

$D = \text{diag}(d_1, \dots, d_p)$ singular values

So

$$X = \sum_{j=1}^p d_j u_j v_j^T \quad [\text{bilinear form}]$$

Now consider $c = v_j$

$$Xc = d_j u_j \quad [V \text{ orthogonal}]$$

$$|Xc| = d_j |c| \quad [U, V \text{ orthogonal}]$$

So for any $\delta > 0$, if $d_j < \delta$ then there is a collinearity according to Definition 1.2.

1.3.2 Estimate Effects

The initial reason for Hoerl and Kennard developing the technique of ridge regression was their observation that the estimates of β were "too long" in the presence of

collinearity. This observation is seen by considering the norm of the ordinary least squares estimate $\hat{\beta}$.

$$\begin{aligned}
 E[\hat{\beta}^T \hat{\beta}] &= E[(\hat{\beta} - \beta + \beta)^T (\hat{\beta} - \beta + \beta)] \\
 &= \beta^T \beta + E[\text{tr}(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] \\
 &= \beta^T \beta + \text{tr}(\text{Var}(\hat{\beta})) \\
 &= \beta^T \beta + \text{tr}((X^T X)^{-1}) \sigma^2 \\
 &= \beta^T \beta + \sigma^2 \text{tr}(V D^{-1} V^T) \text{ [singular value decomposition]}
 \end{aligned}$$

So the magnitude of β in a collinear system is overestimated relative to an orthogonal data matrix X . Further, the magnitude is inversely proportional to the singular values of X . Similarly with the variance estimates, since the covariance matrix is a direct function of the reciprocals of the singular values, ie.

$$\text{Var}(\hat{\beta}) = \sigma^2 V D^{-1} V^T$$

In particular the variance of β can be decomposed into p components. Each one is uniquely associated with just one of the singular values as follows

$$\text{Var}(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^p \frac{v_{kj}^2}{d_j}$$

Hence if two coefficients have high proportions of their variances explained by the same related singular value, then there is an approximate linear dependency between them. This leads to the following definition of the variance decomposition proportion, π_{jk} , being the amount of the variance of β_k explained by the dependency associated with the j th singular value.

Definition 1.2 **Variance Decomposition** Proportions of the matrix $X:n \times p$

$$\pi_{jk} = \frac{\phi_{kj}}{\phi_k} \quad k, j = 1(1)p$$

$$\text{where } \phi_{kj} = \frac{v_{kj}}{d_j} \quad \phi_k = \sum_{j=1}^p \phi_{kj}$$

These proportions form the basis of the identification algorithm developed by [BELSEY, KUH, WELSCH, 80] for the detection of multiple collinearities. This algorithm is shown to extend to the general case in section 3.2.

1.3.3 Inference Effects

Almost all statistics associated with linear regression can potentially be "degraded" by the effects of collinearity. For instance, in testing the hypothesis

$$H_0: \beta_j = 0$$

$$H_a: \beta_j \neq 0$$

the test statistic is

$$t = \frac{\hat{\beta}_j}{\sigma \sqrt{1 - R_j^2}}^{1/2}$$

But recalling the relationship between the coefficient of determination and the singular values, it is seen that the t statistic can be effected by collinearity. In practice the t statistic is often smaller in the presence of collinearity, but this can be compensated for by the "overlarge" magnitude of β . As [GUNST,83] notes, no explicit statements can be made about the magnitudes of the t statistic. However he notes the non centrality parameter of the t statistic is

$$\lambda = (1 - R_j)^{1/2} \frac{\beta_j}{2\sigma}$$

So for fixed β_j and σ , the stronger the offending collinearity involving x_j , the smaller is the non centrality parameter of the test statistic and consequently the lower the power of the test.

As is noted by [BELSEY,KUH,WELSCH,80], the effects of collinearity involving a subset of predictor variables need not effect the estimates of the remaining coefficients, if the two subsets are orthogonal. This is to be expected by comparing it with the introduction of orthogonal variates in a stepwise procedure.

1.3.4 Predictor Effects

The effects of collinearity are often widely different in predictions using "in sample" and "out of sample" data. If the predictor variables lie in the subspace of the original

data, then relatively precise estimation is possible. However "out of sample" predictions may be severely effected by collinearity.

Consider for example the ordinary least squares predictions given by the expression

$$\begin{aligned} y &= x^T \beta \\ &= \beta_0 + x_1^T \beta_1 \end{aligned}$$

Now

$$\begin{aligned} \text{Var}(y) &= \sigma^2 (1/n + x_1^T (X^T X)^{-1} x_1) \\ &= \sigma^2 (1/n + x_1^T T \Lambda^{-1} T^T x_1) \end{aligned}$$

where $T = (t_1, \dots, t_p)$ eigenvectors of $X^T X$

$\Lambda = (\lambda_1, \dots, \lambda_p)$ eigenvalues of $X^T X$

Now if x_1^T is a linear combination of the eigenvectors associated with the "large" eigenvalues, ie

$$x_1^T = a^T T^T$$

$$\text{where } a_i = \begin{cases} 1 & \lambda_i \text{ "large"} \\ 0 & \text{else} \end{cases}$$

then

$$\text{Var}(y) = \sigma^2 (1/n + \sum (1/\lambda_i))$$

where the summation is over "large" eigenvalues. So in this case relatively precise estimation is possible.

2.0 Collinearity in Generalized Linear Models

2.1 Definition of a Generalized Linear Model

The generalized linear model as developed by [NELDER,WEDDERBURN,72] is an umbrella that encompasses amongst others, linear models with normal errors, logit and probit models for quantal responses, and log linear models for counts.

The underlying features of these models are :

- the observations are independent
- the response variance is equal to, or directly proportional to, some function of its expectation
- the response is modelled additively as a function of a linear combination of the predictor variables and error

[WEDDERBURN,74] develops a theory based on the second feature above, the relation of the variance to the mean. He terms this quasi-likelihood, because; although no distribution is explicitly assumed, it is sufficient for estimation purposes and yields analagous results to Fishers classical likelihood. [WEDDERBURN,74] shows that for a one parameter exponential likelihood, the log likelihood is the same as the quasi likelihood. Consequently the weakest sort of distribution that can be assumed for this class of models can be shown to be the generalized one parameter exponential

family. This leads to the following definition, which is based on the notation of [McCULLAGH, NELDER, 83].

Definition 2.1 Generalized Linear Model

$$y = g^{-1}(X\beta) + \epsilon$$

is completely defined by the following three components.

- (i) a random component, y , independently distributed with an $n \times 1$ mean value vector $\mu = E(y)$ and an $n \times 1$ error vector ϵ . The probability density of y follows a generalized exponential density

$$f(y; \theta, \phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y, \phi)\}$$

where θ is called the canonical parameter and

ϕ is called the dispersion parameter.

- (ii) a systematic component, X being an $n \times p$ matrix of predictor variables, and an $p \times 1$ coefficient vector β , giving an $n \times 1$ linear predictor

$$\hat{\eta} = X\beta$$

- (iii) a link function, $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$, connecting the random and systematic components as follows

$$\hat{\eta} = g(\mu) \quad \text{or} \quad \mu = g^{-1}(X\beta)$$

It can be shown that sufficient statistics exist for the parameters in the linear predictor $\hat{\eta} = X\beta$ when $X\beta$ is equal to

the canonical parameter θ . In this case the link function $g(\cdot)$ is called the canonical link of the distribution.

Although it is not necessary to use canonical links, for the purposes of this thesis they will be assumed, as considerable analytic simplification can be made (see Appendix 2A).

Tables 2.1 to 2.3 give distribution details for selected members of the generalized exponential family. Note that the non-central hypergeometric and the multinomial distributions were excluded because their means, variances and link functions are complicated expressions. Table 2.1 lists the density functions for the selected members, while Table 2.2 gives the corresponding density parameters a , b , c for the general exponential distribution. The first column of Table 2.3 gives the canonical link $\theta = \eta$, while the second gives the relationship between the mean vector μ and the canonical parameter θ (see Appendix 1A for details). The third and fourth columns of table 2.3 give the variance of y (ie. v) as a function of the mean and the canonical parameter respectively. It is to be noted that the relationship of v and μ in column three uniquely characterizes the member.

Table 2.1 Densities of the general
exponential family

normal $f(y; \mu, \theta) = \left[\frac{1}{2\pi\theta} \right]^{1/2} \exp\left(-\frac{(y-\mu)^2}{2\theta}\right)$

poisson $f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!} \quad y=0, 1, \dots$

gamma[#] $f(y; \mu, \theta) = \frac{1}{\Gamma(\theta)} e^{-\left[\frac{\theta y}{\mu}\right]} \left[\frac{\theta y}{\mu}\right]^{\theta-1} \frac{1}{y} \quad y > 0$

inverse[#]
Gaussian $f(y; \mu, \theta) = \left[\frac{1}{2\pi\theta y^3} \right]^{1/2} \exp\left(-\frac{\theta(y-\mu)^2}{2\mu^3 y}\right) \quad y > 0$

pth order[#] $f(y; \mu, \theta) = \exp\left(\mu^{-p} \left[\frac{\mu y}{1-p} - \frac{\mu^2}{2-p} \right] + h(y, \theta)\right)$

binomial
proportion $f(y; \mu, n) = \binom{n}{y} \mu^y (1-\mu)^{n-y} \quad y=0/n, 1/n, \dots$

negative
binomial $f(y; \mu, k) = \binom{k+y-1}{y-1} \left[\frac{\mu}{k+\mu} \right]^y \left[\frac{k}{k+\mu} \right]^k \quad y=0, 1, \dots$

1. gamma with $\theta=1$ is the exponential
2. binomial proportion with $n=1$ is logistic
3. # denotes the "p series"

Table 2.2 Parameters of the general
exponential family

	$a(\theta)$	$b(\theta)$	$c(\theta, y)$
normal	0	$\frac{1}{2}\theta^2$	$-\frac{1}{2}[y^2/\theta + \log(2\pi\theta)]$
poisson	1	e^θ	$-\log(y!)$
gamma [#]	0	$-\log(-\theta)$	$-\log\Gamma(\theta) + \theta \log(\theta y)$
inverse [#] Gaussian	0	$-(-2\theta)^{1/2}$	$-1/(2\theta y)$ $-(1/3)\log(2\pi\theta y^3)$
pth order [#]	0	$\frac{-1}{(p-2)} [(1-p)\theta] \left[\frac{p-1}{p-2} \right]$	$h(y, \theta)$
binomial proportion	1/n	$\log(1+e^\theta)$	$\log \binom{n}{ny}$
negative binomial	1	$-k \log(1-e^\theta)$	$\log \binom{k+y-1}{y}$

1. # denotes the "p series"

Table 2.3 Densities of the general
exponential family

	$g(\mu)$	$\mu(\theta)$	$v(\mu)$	$v(\theta)$
normal	μ	θ	1	1
poisson	$\log(\mu)$	e^θ	μ	e^θ
gamma [#]	$-\frac{1}{\mu}$	$-\frac{1}{\theta}$	μ^2	$\frac{1}{\theta^2}$
inverse [#] Gaussian	$-\frac{1}{2\mu^2}$	$\left[-\frac{1}{2\theta}\right]^{1/2}$	μ^3	$\left[-\frac{1}{2\theta}\right]^{3/2}$
pth order [#]	$\frac{-1}{(p-1)\mu^{(p-1)}}$	$\left[\frac{-1}{(p-1)\theta}\right]^{1/(p-1)}$	μ^p	$\left[\frac{-1}{(p-1)\theta}\right]^{p/(p-1)}$
binomial proportion	$\log\left[\frac{\mu}{1-\mu}\right]$	$\frac{e^\theta}{1 + e^\theta}$	$\mu(1-\mu)$	$\frac{e^\theta}{(1 + e^\theta)^2}$
negative binomial	$\log\left[\frac{\mu}{k+\mu}\right]$	$\frac{ke^\theta}{1 - ke^\theta}$	$\mu + \mu^2/k$	$\frac{ke^\theta(1+e^\theta(1-k))}{(1-ke^\theta)^2}$

1. # denotes the "p series"

Further details of these distributions and their applications are to be found in [McCULLAGH,NELDER,83] and [McCULLAGH,82].

In the ensuing development, much reference is made to the maximum likelihood estimation procedures. The important results and brief details of their derivation are given in Appendix 2A. As is shown in this appendix Fisher's scoring technique is equivalent to iteratively reweighted least squares, and an algorithm for this procedure is given in Appendix 2B.

2.2 Definition of Collinearity in Generalized Linear Models

Collinearity in a generalized linear model is more difficult to define than its counterpart in the standard linear case. Intuitively it is again said to be present when:

- two or more predictor variables are highly correlated
- coefficient estimates are different in magnitude or sign from those hypothesised
- coefficient estimates are highly variable
- data collection is deficient
- matrices used in the computation of the coefficient estimates are ill conditioned.

There are several approaches to defining it. One approach based on [BELSEY, KUH, WELSCH, 80] uses linearisation of the model. A second approach is suggested from the iterative calculations for generalized linear models (see Appendix 2A).

2.2.1 Linearisation of the Link Function

[BELSEY, KUH, WELSCH, 80] when commenting on the non linear case in their directions for future research section, suggest linearising the model

$$y = g^{-1}(XB) + \epsilon$$

to the form

$$\tilde{y} = \tilde{X}B + \tilde{\epsilon}$$

Intuitively, by analogy with the standard linear case, it is plausible to define collinearity in the general system as being an approximate linear dependency amongst the columns of $\tilde{X}$. This linearisation is expressed formally in the following theorem.

Theorem 2.1 Linearisation of $y = g^{-1}(XB) + \epsilon$

$y = g^{-1}(XB) + \epsilon$ can be expressed in the form

$$\tilde{y} = \tilde{X}B + \tilde{\epsilon}$$

where $\tilde{x}_i = v_i x_i / a(0)$

with $X = [x_1, \dots, x_n]^T$

$\tilde{X} = [\tilde{x}_1, \dots, \tilde{x}_n]^T$

$V = \text{diag}(v_i)$

This can be proved as follows. First the residual due to error from a non linear procedure is expanded about the estimate β , giving

$$\begin{aligned}\phi(\beta) &= \phi(\hat{\beta}) + J(\hat{\beta})(\beta - \hat{\beta}) + O_{\text{int}}(\beta - \hat{\beta}) \\ &\approx \phi(\hat{\beta}) - J(\hat{\beta})\hat{\beta} + J(\hat{\beta})\beta\end{aligned}$$

where $J(\beta)$ is the gradient matrix of ϕ with respect to β .

Now by letting

$$\tilde{y} = \phi(\beta) - J(\hat{\beta})\beta$$

$$\tilde{X} = -J(\hat{\beta})$$

$$\tilde{\epsilon} = -\epsilon$$

the first equation follows. The second result appeals to the theory of Appendix 2A as follows.

$$\begin{aligned}\tilde{X}_{1,j} &= -J(\hat{\beta})_{1,j} \\ &= -\left[\frac{\partial \phi_1}{\partial \beta_j}\right] \\ &= -\frac{\partial}{\partial \beta_j} [y_1 - g^{-1}(x_1\beta)] \\ &= \left[\frac{\partial}{\partial \beta_j} g^{-1}(x_1\beta)\right]\end{aligned}$$

By definition of the generalised linear model this is just

$$\begin{aligned}&= \frac{\partial \mu_1}{\partial \beta_j} \\ &= \left[\frac{\partial \mu_1}{\partial \eta_1}\right] \left[\frac{\partial \eta_1}{\partial \beta_j}\right] \\ &= \begin{bmatrix} 1 & \frac{\partial \mu_1}{\partial \eta_1} \end{bmatrix} v_1 \begin{bmatrix} \frac{\partial \eta_1}{\partial \beta_j} \end{bmatrix}\end{aligned}$$

Now using the simplification of the canonical link, as in Appendix 2A, and the definition of the linear predictor $\hat{\eta}$, this simplifies to

$$= \frac{v_1 x_1}{a(\emptyset)}$$

It is to be admitted that in the proof above, the unweighted residual does not give an efficient estimate as that due to the weighted residuals, $V^w \in$. However as a first approach, it is useful in suggesting a suitable starting point for defining collinearity in the general case.

2.2.2 Iteratively Reweighted Least Squares Approach

The computation of the estimates can be cast in a weighted least squares formulation as in Appendix 2a ie.

$$\begin{aligned} \hat{\beta}^{(t+1)} &= (X^T W^{(t)} X)^{-1} X^T W^{(t)} Z^{(t)} \\ &= (X^T X)^{-1} X^T Z^{(t)} \end{aligned}$$

This can be viewed as ordinary least squares with the transformation

$$\tilde{X} = W^{\frac{1}{2}} X \quad \tilde{Z} = W^{\frac{1}{2}} Z$$

applied to the data. So it would again be reasonable to define collinearity in the generalized linear model as an approximate linear dependency amongst the columns of $W^{\frac{1}{2}} X$. Now using the definition of the weight as in Appendix 2A, ie.

$$w_1 = \left[\frac{d\mu_1}{d\hat{\eta}_1} \right]^2 \frac{1}{v_1} = \frac{v_1}{a^2(\emptyset)}$$

gives

$$\tilde{x}_{1,j} = \frac{v_1^{1/2}}{a(0)} x_{1,j}$$

2.2.3 Choice of Approach

The link linearisation and iteratively reweighted least squares produce similar transformations, namely

$$\tilde{x}_{1,j} = \frac{v_1 x_{1,j}}{a(0)} \quad \text{and} \quad \tilde{x}_{1,j} = \frac{v_1^{1/2} x_{1,j}}{a(0)}$$

respectively. However the intuitive notions of collinearity discussed earlier are seen more easily from the iteratively reweighted least squares approach. This will be illustrated in the sections 2.4 and 2.5 on sources and effects of collinearity in the generalized linear model. Also [SCHAEFER,83] for similar reasons in the case of logistic regression selects the same approach. Hence, the following definition is adopted.

Definition 2.2 Collinearity in a Generalized Linear Model

$y = g^{-1}(X\beta) + \epsilon$ is said to exist when

For some suitably chosen $\delta > 0$ there exists a $p \times 1$ vector $c \neq 0$ such that

$$|\tilde{X}c| \leq \delta |c| \quad \text{where } \tilde{X} = W^{1/2}X$$

2.3 Relationship of the Standard Linear Model and the Generalized Model Collinearity Definitions

Now v_i is a function of the canonical parameter $\theta_i (=X\beta)$ as in Table 2.3. So substituting for this in Definition 2.2 gives a direct expressions for $\tilde{X}$. The first column of Table 2.4 lists the ij th element of $\tilde{X}$ for the selected members of the general exponential family. It is to be noted from the second column of Table 2.4 that collinearity in the standard linear model carries over to the generalized linear model for all members of the family except the so called p series.

Since, informally, when there is a β such that $|X\beta|$ is "small", then $\tilde{X} \rightarrow X$ for all except the p series (ie except gamma ($p=1$), inverse Gaussian ($p=2$) and p th order). Hence for the other members considered of the family (ie. Poisson, binomial proportion and negative binomial) collinearity in the standard linear model implies collinearity in the generalized linear model.

Table 2.4 Linearised Predictor Elements for the
general exponential family

	$\tilde{x}_{1j}$	$\lim_{\theta \rightarrow 0} \tilde{x}_{1j}$
normal	$\frac{1}{\theta} x_{1j}$	$\frac{1}{\theta} x_{1j}$
poisson	$e^{\theta/\mu} x_{1j}$	x_{1j}
gamma [#]	$\frac{1}{\theta^2} x_{1j}$	∞
inverse [#] Gaussian	$\frac{1}{\theta} \left[\frac{-1}{2\theta} \right]^{3/4} x_{1j}$	∞
pth order	$\frac{1}{\theta} \left[\frac{-1}{(p-1)\theta} \right]^{p/\mu(p-1)} x_{1j}$	∞
binomial proportion	$\frac{1}{n} \frac{e^{\theta/\mu}}{(1+e^{\theta})} x_{1j}$	$\frac{1}{2n} x_{1j}$
negative binomial	$\frac{[k(e^{\theta}(1 + (1-k)e^{\theta}))^{1/2}]}{(1-ke^{\theta})} x_{1j}$	$\frac{[k(2-k)]}{(1-k)} x_{1j}$

1. # denotes the "p series".

However it would be incorrect to define collinearity in the general linear model case as collinearity in X . This is because it is possible for the X system to be collinear, but the premultiplication by W may remove the collinearity in the general linear case for the p series. For example consider a gamma model with $\theta=1$ and

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

Now $\tilde{X} = W^*X$ where $w_1 = \frac{1}{x_{11}\beta}$

So

$$\tilde{X} = \begin{bmatrix} \frac{x_{11}}{\beta_1 x_{11} + \beta_2 x_{12}} & \frac{x_{12}}{\beta_1 x_{11} + \beta_2 x_{12}} \\ \frac{x_{21}}{\beta_1 x_{21} + \beta_2 x_{22}} & \frac{x_{22}}{\beta_1 x_{21} + \beta_2 x_{22}} \end{bmatrix}$$

$$|Xc| = (c_1 x_{11} + c_2 x_{12})^2 + (c_1 x_{21} + c_2 x_{22})^2$$

$$|\tilde{X}c| = \left[\frac{c_1 x_{11} + c_2 x_{12}}{\beta_1 x_{11} + \beta_2 x_{12}} \right]^2 + \left[\frac{c_1 x_{21} + c_2 x_{22}}{\beta_1 x_{21} + \beta_2 x_{22}} \right]^2$$

Now X may be collinear (eg. $x_{11} \approx x_{12} \approx x_{21} \approx x_{22}$ and

$c_1 = 1, c_2 = -1$) with $|Xc| < \delta$, but for β_1, β_2 sufficiently small $|\tilde{X}c| > \delta$.

2.4 Sources of Collinearity in a Generalized Linear Model

In many practical applications the scaling matrix W^* relating the X and $\tilde{X}$ systems will be approximately a constant scalar multiple of the identity matrix.

In particular this occurs when the linear combinations of the rows and β are approximately equal, as is shown in the following theorem.

Theorem 2.1 Condition for a Constant Scaling Matrix W^n

The scaling matrix $W^{\frac{1}{2}}$ will be an approximate scalar multiple of the identity matrix if $x_i\beta$ are approximately equal, where x_i are the rows of X .

This is proved as follows. If w_i equals a constant then so does v_i , since they are equal up to the dispersion parameter when the canonical link is used. But there is a one to one correspondence between v and μ and θ as is listed in Table 2.3. Hence w_i constant implies θ_i is constant. For example consider the binomial proportions case where

$$\mu = 1 + (1 - 4v)^{\frac{1}{2}} = C \text{ say}$$

But

$$\mu = \frac{e^\theta}{1 + e^\theta}$$

so

$$\theta = \log(c/(1-c)) = C'$$

Hence in this case if w_i is constant then so is $\theta_i (=x_i\beta)$.

So in these cases, the same sources of collinearity will also arise in the general linear case, namely :

- large pairwise correlations
- data collection

- model specification
- overdefined model
- outliers

However in the general linear case, it is better to think theoretically in terms of the scaled data matrix $\tilde{X}$. So the above notions are captured by Definition 2.2 with the appropriate substitution. For example the large pairwise correlation of $\tilde{x}_i$ and $\tilde{x}_j$ gives rise to collinearity as in Definition 2.2.

2.5 Effects of Collinearity in a Generalized Linear Model

2.5.1 Estimation Effects

In analogous manner to the standard linear case, both the estimates and their variances are affected by collinearity.

The estimates of β may be expressed iteratively as in result (see Appendix 2A)

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (X^T W^{(t)} X)^{-1} X^T (y - \mu^{(t)})$$

Omitting superscripts for clarity, $X^T W X$ can be decomposed as follows.

$$\begin{aligned} (X^T W X)^{-1} &= (\tilde{X}^T \tilde{X})^{-1} \\ &\quad \text{where } \tilde{X} = W^{1/2} X \\ &= (V D U^T U D V^T)^{-1} && \text{[singular value decomposition]} \\ &= (V D^{-2} V^T)^{-1} \\ &= V D^{-2} V^T \\ &= \sum_{j=1}^p \frac{v_j v_j^T}{d_j^2} \\ &= \sum_{j=1}^p \frac{1}{d_j^2} I_{p \times p} && \text{[V orthonormal]} \end{aligned}$$

So at any iteration (t) , including the final one, collinearity (ie. small singular values of $\tilde{X}$) is seen to directly affect the parameter estimates through the small singular values. Hence collinearity in the generalized linear model can develop throughout the iterative estimation procedure and this suggests that it may be worthwhile for any

estimation procedure to adjust for collinearity at each iteration.

The variance estimates at convergence can be approximated by the asymptotic result

$$\text{Var}(\hat{\beta}) = (X^T W X)^{-1} \sigma^2$$

So again collinearity directly affects the estimation, by inflating the variance estimates according to the small singular values of $\tilde{X}$ (since $X^T V X$ is positive definite and hence has positive singular values).

2.4.2 Inference and Predictor Effects

In a similar manner to the standard linear case the equivalent statistics are affected by collinearity. However the effect is not so direct. For example the analogue of the t statistic is the deviance difference where the deviance is

$$D(y, \hat{\mu}) = 2 \sum w_i [y(\tilde{\theta}_i) - \hat{\theta}_i] - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$$

$$\text{where } \tilde{\theta}_i = \theta(\hat{\mu})$$

$$\hat{\theta}_i = \theta(y_i)$$

is a complex form of $\hat{\beta}$ and W which are affected by collinearity.

Similarly with the generalized Pearson statistic

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}$$

which is used to calculate the dispersion parameter θ .

2.6 Appendix 2A Maximum Likelihood and the Generalized Linear Model

The following derivation of the standard results for the maximum likelihood iterative estimates is taken from [McCULLAGH, NELDER, 83].

Since the observations are independent, consider initially the log likelihood contribution of the i th observation in canonical form l_i , with canonical parameter θ_i and dispersion parameter ϕ . Using the notation of Definition 2.1, and applying Fisher's scoring technique (see [KENDALL, STUART, 67.2] for details) gives the following results for the iterative maximum likelihood estimates.

$$l_i = \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} \right] + c(y_i, \phi)$$

$$\frac{\partial l_i}{\partial \theta_i} = \left[\frac{y_i - \dot{b}(\theta_i)}{a(\phi)} \right] = \frac{y_i - \mu_i}{a(\phi)}$$

Hence the mean vector μ can be represented in terms of the canonical parameter θ as

$$E(y_i) = \mu_i = \dot{b}(\theta_i)$$

by using the standard result

$$E \left[\frac{\partial l_i}{\partial \theta_i} \right] = 0$$

Now

$$\frac{\partial \ln l_1}{\partial \theta_1} = \frac{-\dot{b}(\theta_1)}{a(\theta)}$$

so the variance function v , can be similarly represented in terms of the canonical parameter θ as

$$\text{Var}(y_1) = v_1 = \dot{b}(\theta_1)a(\theta)$$

by using the standard result

$$E\left[\frac{\partial \ln l}{\partial \theta}\right] + E\left[\frac{\partial l}{\partial \theta}\right] = 0.$$

The maximum likelihood equations are obtained in terms of μ and v in the following steps.

$$\left[\frac{\partial l_1}{\partial \mu_1}\right] = \left[\frac{\partial l_1}{\partial \theta_1}\right] \left[\frac{\partial \theta_1}{\partial \mu_1}\right] = \frac{y_1 - \mu_1}{v_1}$$

$$\left[\frac{\partial l_1}{\partial \beta_{rv}}\right] = \left[\frac{\partial l_1}{\partial \mu_1}\right] \left[\frac{\partial \mu_1}{\partial \hat{n}_1}\right] \left[\frac{\partial \hat{n}_1}{\partial \beta_{rv}}\right] = \frac{y_1 - \mu_1}{v_1} \left[\frac{\partial \mu_1}{\partial \hat{n}_1}\right]$$

Now the total contribution to any linear component such as the log likelihood or the score is just the sum of the individual independent contributions. Hence the score vector s , which is the first derivative of the log likelihood, can be expressed as

$$s_{rv} = \left[\frac{\partial l}{\partial \beta_{rv}}\right] = \sum_i \left[\frac{\partial l_i}{\partial \beta_{rv}}\right] = \sum_i \left[\frac{y_i - \mu_i}{v_i}\right] \left[\frac{\partial \mu_i}{\partial \hat{n}_i}\right]$$

For canonical links the following simplification is used repeatedly.

$$\left[\frac{\partial \mu_1}{\partial \hat{n}_1}\right] \frac{1}{v_1} = \left[\frac{\partial \mu_1}{\partial \theta_1}\right] \frac{1}{v_1} = \frac{\dot{b}(\theta_1)}{v_1} = \frac{1}{a(\theta)}$$

So the score component simplifies to

$$s_{1r} = \sum_i \frac{(y_i - \mu_i) x_{1r}}{a(\theta)}$$

Hence the maximum likelihood equations in component form are

$$\sum_i \left[\frac{y_i - \mu_i}{v_i} \right] \left[\frac{\partial \mu_i}{\partial \hat{\eta}_i} \right] x_{1r} = 0 \quad r=1(1)p$$

which for canonical links simplifies to

$$\sum_i \left[\frac{y_i - \mu_i}{a(\theta)} \right] x_{1r} = 0 \quad r=1(1)p$$

or in matrix form

$$\frac{X^T(y - \mu)}{a(\theta)} = 0$$

The information matrix, which was defined by Fisher as minus the expected value of the second derivative of the log likelihood, is derived in the following steps.

$$\begin{aligned} \frac{\partial^2 l_i}{\partial \beta_r \partial \beta_s} &= \frac{\partial}{\partial \beta_s} \left\{ (y_i - \mu_i) \left[\frac{\partial \mu_i}{\partial \hat{\eta}_i} \right] \frac{x_{1r}}{v_i} + \frac{(y_i - \mu_i)}{v_i} \frac{\partial}{\partial \beta_s} \left[\frac{\partial \mu_i}{\partial \hat{\eta}_i} \right] \frac{x_{1r}}{v_i} \right\} \\ &= \left[\frac{\partial \mu_i}{\partial \beta_s} \right] \left[\frac{\partial \mu_i}{\partial \hat{\eta}_i} \right] \frac{x_{1r}}{v_i} \end{aligned}$$

where the second term is zero since it contains the derivative of a constant due to the canonical link simplification. Hence the elements are

$$\begin{aligned} -E \left[\frac{\partial^2 l_i}{\partial \beta_r \partial \beta_s} \right] &= \left[\frac{\partial \mu_i}{\partial \hat{\eta}_i} \right] \left[\frac{\partial \hat{\eta}_i}{\partial \beta_s} \right] \left[\frac{\partial \mu_i}{\partial \hat{\eta}_i} \right] \frac{x_{1r}}{v_i} \\ &= \frac{x_{1r} x_{1s}}{v_i} \left[\frac{\partial \mu_i}{\partial \hat{\eta}_i} \right]^2 \end{aligned}$$

Define weight matrix $W = \text{diag}(w_1, \dots, w_n)$ as

$$w_i = \frac{\left[\frac{\partial \mu_i}{\partial \hat{\eta}_i} \right]^2}{v_i}$$

$$= \frac{v_i}{a''(\eta_i)}$$

So the negative expected value expression simplifies to

$$-E \left[\frac{\partial^2 l_i}{\partial \beta_r \partial \beta_s} \right] = w_i x_{ir} x_{is}$$

So summing to get the total contribution, the r th element of the information matrix is

$$I_{rs} = \sum_i w_i x_{ir} x_{is}$$

or in matrix form

$$I = X^T W X$$

So using Fisher's scoring scheme, the maximum likelihood estimates of β are given by

$$\beta^{(t+1)} = \beta^{(t)} + I^{-1} s$$

For further details of this see [KENDALL, STUART, 67.2]. Now because of the canonical link simplification the information matrix is the same as the Hessian, and so the above scoring scheme is equivalent to the Newton Raphson method (eg. [BARD, 74]) typically expressed as

$$\hat{\delta \beta} = H^{-1} s \quad \text{or}$$

$$H \hat{\delta \beta} = s$$

So

$$H \hat{\beta}^{(t+1)} = H \hat{\beta}^{(t)} + s \quad \text{or}$$

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (X^T W X)^{-1} s$$

$$= \hat{\beta}^{(t)} + (X^T W X)^{-1} X^T (y - \mu^{(t)})$$

But

$$\begin{aligned}
 (\hat{H}\hat{\beta})_1 &= \sum_j H_{1,j} \hat{\beta}_j \\
 &= \sum_j \sum_k w_k x_{k1} x_{kj} \hat{\beta}_j \\
 &= \sum_k w_k x_{k1} \hat{\eta}_k
 \end{aligned}$$

So

$$\hat{H}\hat{\beta}^{(t+1)} = \sum w_k^{(t)} x_{k1} \left[\hat{\eta}_k^{(t)} + \frac{y_k - \mu_k^{(t)}}{\left[\frac{d\mu_k}{d\hat{\eta}_k} \right]^{(t)}} \right]$$

But this is just the form of weighted least squares

$$(X^T W^{(t)} X) \hat{\beta}^{(t)} = X^T Z^{(t)}$$

with pseudo dependent (working) variable

$$z_k = \hat{\eta}_k + \frac{y_k - \mu_k}{\left[\frac{d\mu_k}{d\hat{\eta}_k} \right]}$$

This generally converges, so the final estimate can be represented as

$$\hat{\beta} = (X^T W X)^{-1} X^T W Z$$

2.7 Appendix 2B Iteratively Reweighted Least Squares Algorithm

The following algorithm for calculating β is taken from [McCULLAGH, NELDER, 83].

Algorithm 2.1 Iteratively reweighted least squares

STEP 0

Obtain initial estimates of the linear predictor

$$\hat{\eta}^{(0)} = X\hat{\beta}^{(0)} \text{ from the data}$$

$$\hat{\mu}^{(0)} = y$$

$$\hat{\eta}^{(0)} = g(\hat{\mu}^{(0)})$$

$$\left[\frac{d\mu}{d\eta} \right]^{(0)} = \left[\frac{d\mu}{d\eta} \right] \bigg|_{\hat{\mu}^{(0)}}$$

STEP 1

REPEAT $t = 0(1)$ UNTIL convergence

Set up working dependent variable

$$z^{(t)} = \hat{\eta}_{ik} + \frac{y_{ik} - \hat{\mu}_{ik}}{\left[\frac{d\mu_{ik}}{d\eta_{ik}} \right]}$$

Set up weight W

$$W^{(t)} = \text{diag}(w_1^{(t)}, \dots, w_{n_1}^{(t)})$$

$$w_1^{(t)} = \frac{\left[\frac{d\mu}{d\eta} \right]^{(t)2}}{v_1^{(t)}}$$

Regress $z^{(t)}$ on $x_1, \dots, x_p$ with weight W to get

$$\hat{\beta}^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} z^{(t)}$$

$$\hat{n}^{(t+1)} = X \hat{\beta}^{(t+1)}$$

$$\hat{\mu}^{(t+1)} = g(\hat{\mu}^{(t+1)})$$

$$\left[\frac{d\mu}{d\hat{n}} \right]^{(t+1)} = \left[\frac{d\mu}{d\hat{n}} \right] \bigg|_{\hat{\mu}^{(t+1)}}$$

END REPEAT

3. Diagnostics for Collinearity in Generalized Linear Models

3.1 Desirable Properties for Diagnostic Measures

A diagnostic measure should not be based on methods that try to sense collinearity from its possible effects such as incorrectly hypothesised signs or sensitivity in row deletion statistics. This is because most of these are not necessary and sufficient conditions for collinearity (eg. [MULLET,76] illustrates the signs case). Rather, the diagnostic should be directly linked to the definition of approximate linear dependency amongst the columns of the predictor variables, ie.

$$|Xc| \leq \delta |c|$$

This gives two desirable properties of a diagnostic measure.

First, it should be capable of identifying the linear combinations of offending predictor variables.

Second, it should be able to quantify the effects of collinearity. However the use of the quantitative aspect of the diagnostic measure must be subjective, since there are no distribution assumptions about the nature of the predictor variables X .

Now while collinearity is primarily a data problem, it has consequences for the estimation of the parameters, and in any subsequent hypothesis testing or prediction. In the standard linear case, the measure of degradation due to collinearity (say the singular values of X) directly enters into the expressions for the estimates, tests and predictions. In a similar manner, this natural embedding should be present in the general case.

3.2 Measures of Collinearity

Now while Definition 2.2 exactly expresses the idea of collinearity in the generalized linear model setting, it is an impractical measure, as there are no easy ways of finding the linear combinations.

However, there is a one to one correspondence between collinearity as in Definition 1.2 and the instability of the X system as characterised by small singular values. This is proved in the following theorem.

Theorem 3.1 Equivalence of Collinearity and a Small Singular Value of X for the Standard Linear Model $y = X\beta + \epsilon$

For some suitably chosen $\delta \geq 0$ there exists a $p \times 1$ vector $c \neq 0$ such that $|Xc| \leq \delta |c|$
if and only if

$d_1 \leq \delta$ where d_1 is the smallest singular value of X .

This is proved as follows. The fact that small singular values give rise to collinearity was established in section 1.3.1. So it remains to show that collinearity according to Definition 1.2 implies a small singular value. Without loss of generality, suppose that the c of Definition 1.2 has unit norm (this is equivalent to scaling). Consider

$$\begin{aligned} |Xc|^2 &= c^T X^T X c \\ &\geq \min_u u^T X^T X u \quad \text{where } u^T u = 1 \\ &= d_1^2 \end{aligned} \quad \begin{array}{l} \text{[extrema of ratio of} \\ \text{quadratic forms} \quad] \end{array}$$

Hence if $|Xc| \leq \delta$ then $d_1 \leq \delta$. So d_1 is small. .

Although not emphasized in chapter one, collinearity is a multivariate problem, in that there may be k linearly independent c_i such that $|Xc_i| < \delta$. Now the above theorem can be extended to the case of multiple collinearities as follows.

Corollary 3.1 **Equivalence of Multiple Collinearities and Small Singular Values of X for the Standard Linear Model**

$$y = XB + \epsilon$$

k collinearities are present in the standard linear model if and only if

the k th smallest singular values of X are "small" •

The "only if" argument was established in section 1.3.1 while the "if" argument follows directly from the above theorem by the analogue with the estimation of the variance principal components. •

Further, this carries through to the general linear setting since it is equivalent to a linear transformation of the variables. Hence the following definition of a measure

Definition 3.1 Collinearity Measure in a Generalized Linear Model $y = g^{-1}(XB) + \epsilon$

is given by the "small" singular values of $X = W^m X$ •

The above definition is reasonable as it comes near to having all the desirable properties discussed earlier.

First it is possible to apply the identification algorithm of [BELSEY, KUH, WELSCH, 80] to deduce the offending combinations of pseudo predictor variables x_j corresponding to the singular values. Now it would be highly desirable to relate the singular values directly to the original predictor variables x_j . However this is in general not possible because the natural variables in the general setting are the pseudo

ones x_j . Nevertheless in many practical situations the scaling matrix W is an approximate scalar multiple of the identity matrix and has little effect on the collinear structure.

Second, the magnitudes of the collinearities can be measured by those of the corresponding singularities. To do this a collinearity index is defined as follows,

Definition 3.2 Collinearity Index for Generalized Linear Model $y = g^{-1}(XB) + \epsilon$ is

$$\alpha_i = \frac{d_{\max}}{d_i} \quad i=1(1)p$$

where $(d_1, \dots, d_p)$ are the singular values of $\tilde{X}$.

While it is subjective there is a large body of experience, such as referred to in [MONTGOMERY, PECK, 82], that suggests a collinearity is present if the corresponding collinearity index is greater than 10.0.

Finally the effects of collinearity are clearly manifested in the estimation, inference and prediction calculations as was discussed in section 2.4.

3.3 Model Dependency in Collinear Systems

From the definition it is seen that the collinearity measure for a generalised linear model depends specifically

upon the distribution chosen for the response y (eg. binomial proportion, Poisson, etc).

It is desirable to be able to quantify this indeterminacy in the collinearity. The following theorem conjectured by the author and proved by [MOYLS,85] comes part way in answering this.

Theorem 3.1 Collinearity Index Relationships

$$\frac{w_{\min}}{w_{\max}} \alpha \leq \tilde{\alpha} \leq \frac{w_{\max}}{w_{\min}} \alpha$$

where α , $\tilde{\alpha}$ are the collinearity indices of X , $\tilde{X}$ and $w_{\min}$, $w_{\max}$ are the minimal and maximal diagonal elements of the weight matrix W .

This is proved as follows. Let the subscripts m and M stand for $\min$ and $\max$. Suppose Λ, T is the eigen system of $X^T X$, and $\tilde{\Lambda}, \tilde{T}$ the eigensystem of $X^T W X$. Then

$$\begin{aligned} (X^T X) t_M &= \lambda_M t_M \\ t_M^T (X^T X) t_M &= \lambda_M t_M^T t_M \\ &= \lambda_M \end{aligned} \quad [T \text{ orthogonal}]$$

Let $z = X t_M$ so

$$\lambda_M = z^T z$$

Consider $u^T X^T W X u$. It can be thought of as the ratio of the extrema of quadratic forms, subject to the scaling restriction $u^T u = 1$ and is so maximised by $u = t_M$. So

$$\max_{\tilde{U}} (\tilde{U}^T X^T W X \tilde{U}) = \tilde{t}_M^T (X^T W X) \tilde{t}_M = \tilde{\lambda}_M \geq t_M^T (X^T W X) t_M$$

Hence

$$\tilde{\lambda}_M \geq z^T W z$$

$$= \sum w_i z_i^T z_i$$

$$\geq w_m \sum z_i^T z_i$$

$$= w_m \lambda_m$$

Now by the ratio of the extrema of quadratic forms

$$\tilde{\lambda}_m = \tilde{t}_m^T X^T W X \tilde{t}_m$$

$$= \tilde{z}^T W \tilde{z} \quad \text{where} \quad \tilde{z} = X \tilde{t}_m$$

$$= \sum w_i \tilde{z}_i^T \tilde{z}_i$$

$$\leq w_m \sum \tilde{z}_i^T \tilde{z}_i$$

$$= w_m \tilde{t}_m^T X^T X \tilde{t}_m$$

$$\leq w_m \tilde{t}_m^T X^T X \tilde{t}_m \quad \text{[extrema of ratio]}$$

$$= w_m \lambda_m$$

Thus

$$w_m \lambda_m \leq \tilde{\lambda}_m \leq w_m \lambda_m$$

Similarly

$$w_m \lambda_m \leq \tilde{\lambda}_m \leq w_m \lambda_m$$

Dividing these appropriately yields the result, since the covariance and weight matrices are positive semi definite. •

So this theorem says that the collinearity of the general system is bounded by that of the standard linear system with factors being the ratios of the minimum and maximum scaling weights. This is a useful result for two reasons.

First, given a collinearity analysis with a standard linear model, the effects of collinearity with other general models can be calculated without doing another eigenanalysis.

Second, it shows that if the scaling matrix W is nearly a scalar multiple of the identity matrix, then the collinear structure of the transformed predictor is approximately that of the original predictor matrix. Since this is often the case with many practical applications, this fact is of considerable use in the following chapter on estimation.

By considering the expression $v(\theta)$ in the fourth column of Table 2.3 it is possible to compare the magnitudes of these bounds for the members of the family. In particular for the Poisson

$$\frac{\min e^{x_1 \beta}}{\max e^{x_1 \beta}} \alpha \leq \tilde{\alpha} \leq \frac{\max e^{x_1 \beta}}{\min e^{x_1 \beta}} \alpha$$

while for the gamma the bounds are

$$\frac{\max (x_1 \beta)^{\frac{1}{\alpha}}}{\min (x_1 \beta)^{\frac{1}{\alpha}}} \alpha \leq \tilde{\alpha} \leq \frac{\min (x_1 \beta)^{\frac{1}{\alpha}}}{\max (x_1 \beta)^{\frac{1}{\alpha}}} \alpha$$

It is seen immediately from the above lower bound, that collinearity in the standard linear model becomes more of a problem when a gamma model is used. This is another example of the problems caused by the so called p series.

4.0 Estimation for Generalized Linear Models in the Presence of Collinearity

Having detected collinearity by the methods of the previous chapter, the problem of estimation in its presence remains. The solution to this depends on the causes of the collinearity.

However a cautionary word is necessary. While precise estimation is possible for certain linear combinations of the coefficients as was shown in section 1.3.4, [ALLEN,77] p96 shows with an example

"We should recognise the existence of situations where no estimation by any method is warranted"

Before proceeding to examine the various possible remedial solutions in detail, a general overview is given with respect to the sources and effects discussed in chapter one.

4.1 Remedies for Collinear Sources

From a pragmatic viewpoint, often the effect of the scaling matrix W is negligible, so some progress can be made by treating the model as a standard linear one. This is the approach taken in following sections, except where noted otherwise.

4.1.1 Large Pairwise Correlations in $\tilde{X}$

While this could be alleviated by a Bayesian approach, it is unlikely sufficient information exists to formulate a prior distribution or likelihood. If the pattern is very simple it may be possible to remedy this by variable deletion.

4.1.2 Data Collection

If the dependency is suspected to be due to extrinsic factors, then collection of extra data according to a proper sampling plan is the obvious solution. However, this may often be infeasible due to cost, inconsistency with the previous data, or just lack of available data.

In contrast, if the dependency is intrinsic, then a Bayesian approach may be used to incorporate the information on the underlying dependency mechanism.

4.1.3 Model Specification

The common solution to the case of exact dependencies, is to eliminate the redundant predictor variables (eg. one of the levels of a categorical variable).

With approximate dependencies, since the source is clearly specified, there is good justification for eliminating predictor variables. However as [GUNST,83] notes, one must be

careful to make sure that this is the sole cause of the collinearity in the data.

4.1.4 Overdefined Model

One method to handle this is principal components regression, where the eigenvalues of $\tilde{X}^T \tilde{X}$ are used to eliminate some predictor variables. This has the advantage that exact distribution theory exists for the decision statistics, unlike other variable selection procedures, which have their statistics influenced by the collinearity. However [GUNST,83] p2237 notes there are situations where

"There appears to be no credible substitution for collecting additional data or limiting the goals of the study when multicollinearity is attributed to overdefined models."

4.1.5 Outliers

[PREGIBON,81] shows that outliers in a generalized linear model may be detected by examining the diagonal elements of the generalized projection matrix

$$M = I - W^{-1}X(X^TWX)^{-1}X^TW$$

So using similar criteria to those in the standard linear case, the offending data can be dropped, so alleviating the collinearity.

4.2 Remedies for Collinear Effects

The major observable effect of collinearity is the ill conditioning of the $\tilde{X}$ matrix, and in most practical cases

there may be no obvious reasons for the collinearity other than complex interrelationships in the predictor variables. In this case a tradeoff between increased bias and reduced variance (ie. ridge estimation) may be worthwhile.

4.3 Estimation Methods

4.3.1 Ridge Estimation

Historically ridge analysis stemmed from estimation methods developed to handle the instability of response surface coefficients lying along the ridge of a canonical surface [HOERL,62]. The simplest estimator is the biased ridge estimator

$$\hat{\beta}_R = (X^T X + kI)^{-1} X^T y$$

as proposed in the pioneering paper [HOERL,KENNARD,70]. That paper shows there exists a k_0 such that the mean square error of the ridge estimator is less than that of the ordinary least squares estimator, ie. an introduction of bias results in a reduced mean square error. Numerous empirical methods have been developed for choosing k , and several variants of the simple estimator have been subsequently proposed.

There has been much criticism of ridge estimation. [SMITH,CAMPBELL,80] p75 is a spirited attack against the use of ridge regression as

"mechanical data manipulation that is insensitive to the particular phenomena being modeled and to information about coefficients"

They argue that ridge estimators are based on "ad hoc pseudoinformation", and instead advocate a semi Bayesian approach. However the critics of this paper point out that, while the prior information held is too weak to be formulated in terms of a prior distribution; it is nevertheless essential and ridge regression approximately incorporates this imprecise information. Furthermore, as [GUNST,80] p100 notes

"ridge regression has been successfully applied too frequently, when little prior information is available, for its use to be restricted only to data for which formal Bayesian priors are known"

4.3.1.1 Standard Linear Case

In their derivation of a ridge estimator [HOERL,KENNARD,70] reason as follows

"the worse the conditioning of $X^T X$, the more $\hat{\beta}$ can be expected to be too long. On the other hand, the worse the conditioning, the further one can move from $\hat{\beta}$ without an appreciable increase in the residual sum of squares. In view of $E[\hat{\beta}^T \hat{\beta}] = \beta^T \beta + \sigma^2 \text{tr}(X^T X)^{-1}$ it seems reasonable that if one moves away from the minimum sum of squares point, the movement should be in a direction which will shorten the length of the regression vector."

So the ridge estimator is the shortest β whose sum of squares due to error is constrained to be within d^2 say of that of the ordinary least squares estimator β_{OLS} ie.

$$\begin{aligned} d^2 &= \text{SSE}(\beta_R) - \text{SSE}(\hat{\beta}_{OLS}) \\ &= (y - \hat{y})^T (y - \hat{y}) - (y - \hat{y}_{OLS})^T (y - \hat{y}_{OLS}) \end{aligned}$$

$$\begin{aligned}
&= (y - X\beta_R)^T(y - X\beta_R) - (y - X\hat{\beta}_{OLS})^T(y - X\hat{\beta}_{OLS}) \\
&= (\beta_R - \hat{\beta}_{OLS})^T X^T X (\beta_R - \hat{\beta}_{OLS}) + \\
&\quad 2(\beta_R - \hat{\beta}_{OLS})^T X^T (y - X\hat{\beta}_{OLS}) \\
&= (\beta_R - \hat{\beta}_{OLS})^T X^T X (\beta_R - \hat{\beta}_{OLS}) \quad [\text{normal equations}]
\end{aligned}$$

Thus the ridge trace is the solution of

minimise $\beta_R^T \beta_R$

subject to $(\beta_R - \hat{\beta}_{OLS})^T X^T X (\beta_R - \hat{\beta}_{OLS}) < d^2$

Now the Lagrangian expression with multiplier $1/k$ is

$$L = \beta_R^T \beta_R + 1/k (\beta_R - \hat{\beta}_{OLS})^T X^T X (\beta_R - \hat{\beta}_{OLS})$$

which gives has solution of the form

$$\hat{\beta}_R = (X^T X + kI)^{-1} \hat{\beta}_{OLS}$$

the simple ridge estimator.

4.3.1.2 General Linear Case

This extends to the general setting as follows with the following definition of a ridge estimator β_R defined in terms of the usual iteratively reweighted least squares solution $\hat{\beta}_{IRLS}$.

Definition 4.1 Ridge Estimation for the generalized linear model $y = g^{-1}(X\beta) + \epsilon$ is given by the solution

minimise $\beta_R^T \beta_R$

subject to $(\beta_R - \hat{\beta}_{IRLS})^T X^T W X (\beta_R - \hat{\beta}_{IRLS}) < d^2$

This has solution

$$\hat{\beta}_R = (X^T W X + kI)^{-1} X^T W X \hat{\beta}_{IRLS}$$

The proof of this follows by extending [SCHAEFER,79] to the general setting as follows. $\hat{\beta}_{IRLS}$ is the iteratively reweighted least squares estimator, with transformation $\tilde{X} = W^{1/2} X$, in the sense that it minimises the weighted sums of squares due to error

$$WSSE = (y - g^{-1}(X\beta))^T W^{-1} (y - g^{-1}(X\beta))$$

Again the sum of squares is constrained to within say d^2 of the usual estimator β_{IRLS} . Hence

$$\begin{aligned} d^2 &= WSSE(\beta_R) - WSSE(\hat{\beta}_{IRLS}) \\ &= (y - g^{-1}(X\beta_R))^T W^{-1} (y - g^{-1}(X\beta_R)) + \\ &\quad (y - g^{-1}(X\hat{\beta}_{IRLS}))^T W^{-1} (y - g^{-1}(X\hat{\beta}_{IRLS})) \\ &= (y - \mu(\beta_R))^T W^{-1} (y - \mu(\beta_R)) - \\ &\quad (y - \mu(\hat{\beta}_{IRLS}))^T W^{-1} (y - \mu(\hat{\beta}_{IRLS})) \\ &= (\mu(\beta_R) - \mu(\hat{\beta}_{IRLS}))^T W^{-1} (\mu(\beta_R) - \mu(\hat{\beta}_{IRLS})) + \\ &\quad 2(\mu(\beta_R) - \mu(\hat{\beta}_{IRLS}))^T W^{-1} (y - \mu(\hat{\beta}_{IRLS})) \end{aligned}$$

Now $\mu(\beta_R)$ and $\mu(\hat{\beta}_{IRLS})$ can be approximated by first order vector Taylor series as

$$\mu(\beta_R) = \mu(\beta) + W X (\beta_R - \beta)$$

$$\mu(\hat{\beta}_{IRLS}) = \mu(\beta) + W X (\hat{\beta}_{IRLS} - \beta)$$

so the above simplifies to

$$\begin{aligned} d^2 &= (\beta_R - \hat{\beta}_{IRLS})^T X^T W X (\beta_R - \hat{\beta}_{IRLS}) + \\ &\quad (\beta_R - \hat{\beta}_{IRLS})^T X^T (y - W X \hat{\beta}_{IRLS}) \end{aligned}$$

and the generalized linear model likelihood equation simplifies it further to

$$= (\beta_R - \hat{\beta}_{IRLS})^T X^T W X (\beta_R - \hat{\beta}_{IRLS})$$

So following [HOERL,KENNARD,70] in an analogous manner gives the minimisation result and using a Lagrange multiplier the result follows.

Now [HOERL,KENNARD,70] give a justification for their method in an existence-construction theorem. The existence result shows there is a non zero value of k , such that the mean square error of this estimator is less than that of the ordinary least squares estimator. This result extends to the general setting as is shown in the following theorem.

Theorem 4.2 Existence of k_0 : $MSE(\hat{\beta}_R(k_0)) \leq MSE(\hat{\beta}_{IRLS})$

for generalized linear model $y = g^{-1}(X\beta) + \epsilon$

If

- (i) x_{ij} are uniformly bounded
- (ii) $(X^T V X)$ is of order $O(1/n^2)$
- (iii) $V(\mu) \quad \dot{V} = \frac{dV}{d\mu} \quad \ddot{V}$ bounded
- (iv) 3rd and 4th central moments of ϵ exist

Then for n sufficiently large there exists

$$k_0 \geq 0 : MSE(\hat{\beta}_R(k_0)) \leq MSE(\hat{\beta}_{IRLS})$$

This theorem was originally proved by [SCHAEFER,79] for the case of logistic regression. The extension of the proof to the general case above, is straightforward if condition (iii) is included and the analogue between the 1st and 2nd

moments in logistic regression estimation and those in the general setting is recognised. This proof will be omitted as the multivariate algebra is long and tedious.

As with the standard linear case, it must be noted that while the theorem is appealing theoretically it does not give any assurance that k can be chosen optimally from the data. Further it is of limited value as no bounds have been found for k as yet. Also since the degree of collinearity depends solely upon the singular values of X , it is disturbing to have the side condition for sufficiently large n present. So one is led back to the alternative empirical argument of [GUNST,83] referenced earlier.

[HOERL,KENNARD,70] give a construction section in their theorem, where they deduce bounds for k . However [SCHAEFER,79] produced a counterexample that shows, in the logistic regression case, a construction result is not achievable. So it is not possible to get any guidance on the choice of k in the general setting. However by analogy with the standard linear case, several candidates are suggested.

[MONTGOMERY,PECK,82] p340 in reviewing various rules state that

"no single procedure emerges from these studies as best overall... Our own preference in practice is for ordinary ridge regression with k selected by inspection of the ridge trace ... It is also occasionally useful to find the "optimum" value of k suggested by Hoerl, Kennard and Baldwin [1975] and the iteratively estimated "optimum" of Hoerl and

Kennard [1976] and compare the resulting models with the one obtained via the ridge trace."

Hence the following extended definition

Definition 4.1B Selection of k in Ridge Estimation

for the general linear model $y = g^{-1}(X\beta) + \epsilon$

Rule I : k is that value which causes the estimate $\beta_R(k)$ to stabilise, and is found by inspection of the ridge trace ie. $\beta_R(k)$ plotted against k .

Rule II : [HOERL, KENNARD, BALDWIN, 75]

$$k = \frac{p\hat{\sigma}^2}{\hat{\beta}_{IRLS}^T \hat{\beta}_{IRLS}}$$

where $\hat{\sigma}^2$ is the dispersion parameter estimate of the generalized linear model.

Rule III: [HOERL, KENNARD, 76]

Algorithm

$$k^{(0)} = k_0$$

[Rule II]

REPEAT $i=1(1)$ UNTIL convergence

$$k^{(i+1)} = \frac{p\hat{\sigma}^2}{\hat{\beta}(k^{(i)})^T \hat{\beta}(k^{(i)})}$$

$$\hat{\beta}(k^{(i+1)}) = (X^T W X + k^{(i+1)} I)^{-1} X^T W y$$

END REPEAT

There is a legion of other methods to choose from, including the richer method of generalized ridge estimation, but the above three methods serve to illustrate the general idea.

4.3.2 Bayesian Estimation

4.3.2.1 Standard Linear Case

[LINDLEY, SMITH, 72] show, in the standard linear case, if y has the normal distribution $N_p(X\beta, \sigma^2 I)$ and β has the multivariate normal prior distribution $N(0, \sigma_a^2 I)$ then the posterior mean of β is

$$\hat{\beta} = [X^T X + kI]^{-1} X^T y$$

$$\text{where } k = \sigma^2 / \sigma_a^2$$

So by adding more information in terms of a prior distribution for β a ridge type estimator is obtained. The results of this paper are motivated by a strictly Bayesian philosophy, pivoting around the concept of exchangeable priors. Unfortunately this concept is a little unpalatable for applied work, since it means the β must come from a standard multivariate normal distribution. This implies a standardised model must always be used, which detracts from the method. However [NELDER, 72], in reviewing [LINDLEY, SMITH, 72], points out that the Bayesian formulation is equivalent to the concept of prior likelihood [EDWARDS, 69]

$$\text{posterior likelihood} = \text{prior likelihood} * \text{likelihood}$$

This is a simpler approach and fits in better with the likelihood framework of generalized linear models.

4.3.2.2 General Linear Case

In the absence of evidence from other experiments, the choice of a normal prior for prior likelihood is a natural one since as is known that the sampling estimates β are asymptotically normal. So using the philosophy of [EDWARDS,69] the posterior log likelihood of β is

$$l_p = \left(\sum_{i=1}^n \left[\frac{y_i x_i \beta - b(x_i \beta)}{a(\theta)} \right] + c(y_i, \theta) \right) + \left(\frac{-\frac{1}{2} \beta^T \beta}{\sigma_{\theta}^2} - \frac{p}{2} \log \sigma_{\theta}^2 \right)$$

where the first term is the log likelihood of the generalized exponential distribution as in definition 1.3, and the second term is the prior log likelihood using a normal prior with parameters $(0, \sigma_{\theta}^2)$. The maximum likelihood estimates are obtained from the partial derivatives and are presented in Appendix 4A. Hence the following definition of a posterior mode estimator.

Definition 4.2 Posterior Mode Estimator for the generalized linear model $y = g^{-1}(X\beta) + \epsilon$ is the value β_{PL} that maximises the density expression

$$l_p = \left(\sum_{i=1}^n \left[\frac{y_i x_i \beta - b(x_i \beta)}{a(\theta)} \right] + c(y_i, \theta) \right) + \left(\frac{-\frac{1}{2} \beta^T \beta}{\sigma_{\theta}^2} - \frac{p}{2} \log \sigma_{\theta}^2 \right)$$

4.3.3 Principal Component Estimation

4.3.3.1 Standard Linear Case

Principal components, in the standard linear case, produces biased estimates by using a reduced set of predictor variables, where the variables removed are those corresponding to small singular values in the design matrix. The results are easiest derived by considering the canonical form of the standard linear model, ie.

$$y^* = X^* \beta^* + \epsilon$$

$$\text{where } X^* = XT \quad X^T X = T^T T \quad T^T X^T X T = \Lambda$$

$$\Lambda = [\lambda_1, \dots, \lambda_p] \text{ eigenvalues of } X^T X$$

$$T = [t_1, \dots, t_p] \text{ eigenvectors of } X^T X$$

$$\beta^* = T^T \beta_{OLS}$$

Then the principal components estimator is defined as

$$\hat{\beta}_{PC} = TB\hat{\beta}^*$$

$$\text{where } B = \text{diag}(b_1, \dots, b_p)$$

$$b_i = \begin{cases} 0 & \lambda_i \text{ "small"} \\ 1 & \text{else} \end{cases}$$

So substituting for β^* from above

$$\begin{aligned} \hat{\beta}_{PC} &= TBT^T \hat{\beta}_{OLS} \\ &= TBT^T (X^T X)^{-1} X^T X \hat{\beta}_{OLS} && [* / X^T X] \\ &= TBT^T T \Lambda^{-1} T^T X^T X \hat{\beta}_{OLS} && [\text{diagonalisation } (X^T X)^{-1}] \\ &= T \Lambda^{-1} T^T X^T X \hat{\beta}_{OLS} && [T \text{ orthogonal}] \\ &= \tilde{\Lambda}^{-1} T^T X^T X \hat{\beta}_{OLS} && [\tilde{\Lambda}^{-1} = B \Lambda^{-1}] \end{aligned}$$

$$= (X^T X)^{-1} X^T X \hat{\beta}_{OLS}$$

where $(X^T X)^{-1} = \sum_{i=1}^{p-m} \frac{1}{\lambda_i} t_i t_i^T$ is an approximate inverse

Now

$$\begin{aligned} \text{Var}(\beta_{PC}) &= \text{Var}(TB\beta^*) \\ &= TB \text{Var}(\beta^*) B^T T^T && [\text{Var}(AX) = A \text{Var}(X) A^T] \\ &= TB (X^{*T} X^*)^{-1} B^T T^T \sigma^2 && [\text{by construction}] \\ &= TB \Lambda^{-1} B^T T^T \sigma^2 && [\text{diagonalisation}] \end{aligned}$$

which is to be compared with the variance of the ordinary least squares estimator

$$\text{Var}(\hat{\beta}_{OLS}) = T \Lambda^{-1} T^T \sigma^2$$

Hence principal components alleviates the consequences of collinearity in the variances of the effects, by removing the large terms due to those variables corresponding to small singular values.

4.3.3.2 General Linear Case

In a similar manner it is possible to extend this definition to the general setting as follows.

Definition 4.3 Principal Components Estimator

for the general linear model $y = g^{-1}(X\beta) + \epsilon$ is given by

$$\hat{\beta}_{PC} = (X^T W X)^{-1} X^T W X \hat{\beta}_{IRLS}$$

where

$$(X^T W X)^{-1} = \tilde{T} \tilde{\Lambda}^{-1} \tilde{T}^T$$

$$\tilde{\Lambda}^{-1} = B\Lambda^{-1}$$

$$\Lambda = [\lambda_1, \dots, \lambda_p] \text{ eigenvalues of } X^T W X$$

$$T = [t_1, \dots, t_p] \text{ eigenvectors of } X^T W X$$

$$B = \text{diag}(b_1, \dots, b_p)$$

$$b_i = \begin{cases} 0 & \lambda_i \text{ "small"} \\ 1 & \text{else} \end{cases}$$

Hence by direct analogy the effects of collinearity on the variances of the estimates in the generalized linear model are similarly reduced.

4.4 Appendix 4A Maximum Likelihood Estimates of the Posterior Distribution

The maximum likelihood equations are obtained in the usual manner by differentiating with respect to the parameters. However the problem is then slightly non standard because there is no associated distribution to use for the expected values in obtaining the information matrix. However [PATEFIELD,77] shows that when nuisance parameters like the variance of β are involved, the the formation matrix of [EDWARDS,69a] is more appropriate. He further shows that it is acceptable to remove the nuisance parameter by substituting it out, using the maximum likelihood equations. Hence the following calculations

$$l_p = \left(\sum_{i=1}^n \left[-\frac{y_i x_i \beta - b(x_i \beta)}{a(\theta)} \right] + c(y_i, \theta) \right) + \left(-\frac{\beta^T \beta}{\sigma_a^2} - \frac{p}{2} \log \sigma_a^2 \right)$$

So the following partial differentials are obtained

$$\frac{\partial l_p}{\partial \beta_r} = \frac{\sum (y_i - \mu_i) x_i}{n} - \frac{\beta_r}{\sigma_a^2}$$

$$\frac{\partial l_p}{\partial \sigma_a^2} = \frac{\beta^T \beta}{2(\sigma_a^2)^2} - \frac{p}{2\sigma_a^2}$$

$$\frac{\partial l_p}{\partial \theta} = - \left[\frac{\sum y_i x_i \beta - b(x_i \beta)}{\theta^2} + \frac{\partial c(y_i, \theta)}{\partial \theta} \right]$$

$$\frac{\partial^2 l_p}{\partial \beta_r \partial \beta_s} = \sum \frac{x_{1r} x_{1s}}{v_1} \left[\frac{\partial \mu_1}{\partial \beta_1} \right] - \frac{\delta_{rs}}{\sigma_a^2} \quad [\delta_{rs} \text{ Kronecker delta}]$$

$$\frac{\partial l_P}{\partial \beta_r \partial \sigma_n^2} = \frac{\sum (y_i - \mu_i) x_i}{\sigma_n^2} - \frac{\beta_r}{(\sigma_n^2)^2}$$

$$\frac{\partial l_P}{\partial \beta_r \partial \theta} = - \frac{\sum (y_i - \mu_i) x_i}{\theta^2}$$

$$\frac{\partial l_P}{\partial \sigma_n^2 \partial \theta} = 0$$

$$\frac{\partial l_P}{\partial \sigma_n^2} = \frac{-\beta^T \beta}{(\sigma_n^2)^2} + \frac{p}{(\sigma_n^2)^2}$$

Eliminating the nuisance parameter σ_n^2 using the maximum equations gives

$$\sigma_n^2 = \frac{\beta^T \beta}{p}$$

The score vector is

$$s = \begin{bmatrix} \frac{\partial l_P}{\partial \beta} \\ \frac{\partial l_P}{\partial \theta} \end{bmatrix}$$

and the formation matrix is

$$F = - \begin{bmatrix} \frac{\partial^2 l_P}{\partial \beta \partial \beta^T} & \frac{\partial^2 l_P}{\partial \beta \partial \theta} \\ \frac{\partial^2 l_P}{\partial \beta \partial \theta} & \frac{\partial^2 l_P}{\partial \theta^2} \end{bmatrix}$$

with the usual scoring equation being

$$\begin{bmatrix} \beta \\ \theta \end{bmatrix}^{(t+1)} = \begin{bmatrix} \beta \\ \theta \end{bmatrix}^{(t)} + F^{-1} s$$

5.0 Illustrative Example

5.1 Introduction

There are two approaches to validating the methods developed in the previous chapters. One approach would be to obtain a large data set which has collinear symptoms. The diagnostic measures of chapter two could then be used to verify this, and the methods of chapter three applied to obtain estimators. However, this approach has the disadvantage, that it is not possible to directly calculate the mean square error, since the "true" coefficients are unknown. Because of this, a simulation has been chosen.

5.2 Scope of the Simulation

Other simulations such as [DEMPSTER,SCHATZOFF,WERMOUTH,77] , [McDONALD,GALARNEAU,75] and [SCHAEFER,82] have studied the effects of several parameters - sample size (n), number of predictor variables (p), degree of collinearity (d), alignment of β (a), etc. But this is beyond the resources of this thesis, and so a restricted simulation, that serves to illustrate some of the methods of chapter four, has been adopted. In this simulation parameters n , p , d are restricted to one level and a to two levels.

5.3 Generation of Collinear Data

5.3.1 Standard Linear Case

The procedure that has been followed in many simulations (eg. [McDONALD,GALARNEAU,75], [HEMMERLE,BRANTLE,78], [WICHERN,CHURCHILL,78], [GIBBONS,81]) has been to generate a collinear design matrix from a theoretical linear dependency. Two "true" coefficient vectors have then been selected, as the eigenvectors corresponding to the smallest and largest eigenvalues of the $X^T X$ system. It is known that the mean square error is maximised and minimised respectively, with these choices. [GIBBONS,81] cites [THISTED,76] p74 as justification for these choices

" "extreme-case simulation" experiments appear to be the most economical and informative, especially for preliminary studies of new results"

Finally, the response vector is generated as the sum of the systematic part due to $X\beta$ and a random normal error.

Now much criticism has been levelled against this design eg. [DRAPER,VAN NOSTRAND,79] and much work needs to be done to construct a fair basis for comparison. However, this is beyond the resources of this thesis.

5.3.2 General Linear Case

The problem of generating a collinear data set from a generalized linear model is not as simple as in the standard

linear case, since for the p series the $X^T X$ system must be used.

A solution, in a similar manner to the standard linear case might appear to be to generate X as collinear, selecting β as the eigenvector corresponding to the smallest eigenvalue of $X^T X$. Then the relation $\tilde{X} = W^{1/2}(X, \beta)X$ could be used to recover the original X matrix. However taking β as the eigenvectors of $X^T X$, which is itself a function of β leads to an overconstrained system with no non trivial solution. Hence it is not possible to choose the alignment of β as recommended by [THISTED,76].

The alternative is to start with X collinear and select β so as to maintain collinearity in $\tilde{X}$ by realising that premultiplication by $W^{1/2}$ is effectively row scaling of X .

For convenience the case of the gamma density with unit dispersion parameter (ie. $\theta=1$) is chosen. The method of [GIBBONS,81] is used to generate X with a controllable degree of collinearity. This method is as follows.

Algorithm 5.1 Modified [GIBBONS,81] Collinear Generator of X

STEP 1

Generate z_{ij} as independent $N(0,1)$ pseudo random variables

$$i = 1(1)n \quad i=1(1)p+1$$

STEP 2

Select α where $\alpha^{(m)}$ is the correlation between any two variables

Compute

$$x_{ij} = (1 - \alpha^{(m)})z_{ij} + \alpha z_{i, p+1} \quad i = 1(1)n \quad j=1(1)p$$

STEP 3

Compute the $\tilde{X}$ matrix as

$$\tilde{X} = W^{1/2}X \quad \text{where } w_i^{1/2} = 1/x_i^T \beta$$

In the generalized linear model $y = g^{-1}(X\beta) + \epsilon$ the response y is the sum of the systematic component $g^{-1}(X\beta)$ and the error term ϵ , that has the one parameter exponential distribution. So it is not meaningful to add a specific error term to the systematic component to generate y . Rather y can be thought of as having come from a one parameter exponential distribution with mean vector $\mu = g^{-1}(X\beta)$. So for the gamma distribution considered, the mean vector is

$$\mu_i = 1/x_i^T \beta$$

5.4 Simulation

5.4.1 Simulation Setup

The model to be considered in the simulation is

$$y = g^{-1}(X\beta) + \epsilon \quad \text{where}$$

y : $n \times 1$ gamma variate dispersion parameter $\theta=1$

X : $n \times p$

n : sample size = 30

p : number predictor variables (including intercept) = 4

β : $p \times 1$ 2 levels $\beta^{(L)}$ $\beta^{(H)}$

$$\beta^{(L)} = \begin{bmatrix} 0.01 \\ 0.01 \\ 0.01 \\ 0.01 \end{bmatrix} \quad \beta^{(H)} = \begin{bmatrix} -0.001 \\ 0.01 \\ -0.01 \\ 0.01 \end{bmatrix}$$

Two levels were chosen for the degree of collinearity d as measured by α^2 , the correlation coefficient between any two predictor variables

$$d^{(1)} = 0.80 \quad d^{(2)} = 0.95$$

$\beta^{(L)}$, $\beta^{(H)}$ were chosen as above for the following reason. Given X generated by Algorithm 5.1 and $\beta^{(L)}$ chosen as above the eigenvectors corresponding to the smallest and largest eigenvalues ($t^{(L)}$, $t^{(H)}$) of $\tilde{X}$ with d at the two levels are

$$d^{(1)} = 0.80$$

$$d^{(2)} = 0.95$$

$$t^{(L)} \quad t^{(2)}$$

$$t^{(L)} \quad t^{(2)}$$

$$\begin{bmatrix} -0.63 \\ -0.48 \\ -0.43 \\ -0.44 \end{bmatrix} \quad \begin{bmatrix} 0.28 \\ -0.49 \\ 0.65 \\ -0.50 \end{bmatrix}$$

$$\begin{bmatrix} -0.75 \\ -0.39 \\ -0.36 \\ -0.37 \end{bmatrix} \quad \begin{bmatrix} 0.05 \\ -0.34 \\ 0.78 \\ -0.52 \end{bmatrix}$$

Now as discussed above it is not possible to align β with the eigenvectors of $X(\beta)$ analytically, but by the suitable choice of $\beta^{(L)}$ it is seen that $\beta^{(L)}$ and $\beta^{(2)}$ are approximately aligned in the directions of the respective eigenvectors of $X(\beta^{(L)})$ for $d = 0.95$. As little improvement is expected with the $\beta^{(2)}$ alignment only the setting $(\beta^{(2)}, d^{(2)})$ is used.

A pilot simulation of ten runs was made to estimate a reasonable run sample size. The criterion was that of being able to detect a 10 percent change in the estimates β at the 5 percent level. This gave a run size (m) of 20.

Because of resource constraints, just three estimators were considered

β_{IRLS} : iteratively reweighted least squares estimator

β_{RII} : [HOERL, KENNARD, BALDWIN, 75] ridge estimator

β_{PC} : principal components estimator

To assess the effectiveness of the estimators, the usual criterion of the so called average square error

$$ASE(\hat{\beta}) = (\hat{\beta} - \beta)^T (\hat{\beta} - \beta)$$

is used.

Now it is possible to obtain approximations for the variance of the ridge estimators to order $O(1/n^2)$ (see [SCHAEFER,79] chapter II) in terms of the third and fourth moments of the distribution of y . However this is omitted here, but should be considered in any future work.

5.4.2 Simulation Implementation and Difficulties

The simulation was divided into three stages. The first stage used the SAS procedure PROC MATRIX to generate the collinear data sets using Algorithm 5.1. The second stage used the generalized linear models package GLIM to get the iteratively reweighted least squares estimates. The third stage again used PROC MATRIX and calculated the biased estimators and statistics.

While the SAS PROC MATRIX routine is pleasant to use because of its matrix style syntax and is effective due to machine coding of the matrix routines; its input/output ability leaves much to be desired. The result of this deficiency is that very large matrices must be stored (with

numerous consequent physical i/o problems) if efficient use is to be made of the interpreter code.

During the running of the second stage, a computational fault was found in GLIM. This effected approximately five percent of the data sets. It would appear that when a gamma model is used with severe collinearity (as was here), GLIM has no adjustment for negative fitted values and sensibly aborts with the warning message

*****FAULT 59 *****

To overcome this extra data sets were substituted.

5.4.3 Simulation Results and Conclusions

The results of the simulations are displayed in Tables 5.1 and 5.2 and Figure 5.1.

Table 5.1 lists in each cell the average estimated parameter values β_1 , β_2 , β_3 , β_4 from the m runs with the estimator type by column and the degree of collinearity d'' nested within alignment β'' by row.

Table 5.2 in a similar layout gives the values of the actual squared error and its associated standard error in parentheses. This table also displays the percentage reduction ratio in ASE by using the biased estimators $\hat{\beta}_{BII}$ and $\hat{\beta}_{BC}$ as opposed to the usual estimator $\hat{\beta}_{IRLB}$.

(ie. $\frac{\hat{\beta}}{\hat{\beta}_{IRLS}} * 100$). In accordance with the sample size criteria the only comparisons considered were the multiple comparison of the three estimators within each setting of $\beta^{(i)}$ and $d^{(i)}$. Tukey's pairwise comparison test at the five percent significance level is displayed pictorially in Table 5.2 where the bars represent homogeneous groups.

Figure 5.1 displays graphically the simulation estimates of β for the three settings $(\beta^{(L)}, d^{(1)})$, $(\beta^{(L)}, d^{(2)})$ and $(\beta^{(H)}, d^{(2)})$. Note with $\beta^{(H)}$ the scaling has been modified to that of $\beta^{(L)}$ to make the graph easier to comprehend.

Table 5.1 Estimators for the Gamma Model

		$\hat{\beta}_{IRLS}$	$\hat{\beta}_{RII}$	$\hat{\beta}_{EC}$
$\beta^{(L)}$	$d^{(1)}$	0.0128	0.0107	0.00940
		0.00941	0.00958	0.0101
		0.0101	0.00914	0.0106
		0.0128	0.0105	0.0103
	$d^{(2)}$	0.0127	0.0110	0.00995
		0.00804	0.0107	0.0101
		0.00749	0.00735	0.0104
		0.0163	0.0123	0.0101
$\beta^{(S)}$	$d^{(2)}$	0.00763	0.00413	0.00175
		0.0100	0.00672	0.00229
		0.0101	0.00314	-0.00279
		0.0105	0.00548	0.00262

1. $\beta^{(L)\top} = (0.01, 0.01, 0.01, 0.01)$

$\beta^{(S)\top} = (-0.001, 0.01, -0.01, 0.01)$

2. $d^{(1)} = 0.80$ $d^{(2)} = 0.95$

3. Each cell contains $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)^\top$

Table 5.2 Actual Squared Error of Estimators
for the Gamma Model

		$\hat{\beta}_{IRLS}$	$\hat{\beta}_{RII}$	$\hat{\beta}_{PC}$
$\beta^{(L)}$	$d^{(1)}$	0.00171 [0.00209]	0.000450 [0.000984] 26%	0.000167 [0.00315]] 9%
	$d^{(2)}$	0.0139 [0.000177]	0.00268 [0.00722]] 19%	0.0000942 [0.000142]] 1%
$\beta^{(S)}$	$d^{(2)}$	0.000281 [0.000454]	0.000197 [0.000193] 55%	0.000286 [0.0000465]] 100%

1. $\beta^{(L)T} = (0.01, 0.01, 0.01, 0.01)$

$\beta^{(S)T} = (-0.001, 0.01, -0.01, 0.01)$

2. $d^{(1)} = 0.80$ $d^{(2)} = 0.95$

3. Each cell contains ASE average for m runs

[standard error]

reduction ratio % using a biased
estimator

4. Bars represent homogeneous groups with Tukey's pairwise
comparison at the 5% level

Figure 5.1 Mean $\pm$ Standard Deviation Intervals of Estimators for the Gamma Model

$\beta^{(L)T} = (0.01, 0.01, 0.01, 0.01)$
 $\beta^{(S)T} = (-0.001, 0.01, -0.01, 0.01)$
 $d^{(1)} = 0.80 \quad d^{(2)} = 0.95$

The major conclusions that can be drawn from the results are as follows.

When the alignment is favourable (ie. $\beta = \beta^{(L)}$), then there is considerable reduction in variance from using a biased estimator. Also, the amount of reduction is proportional to the degree of collinearity present. In this case principal components seems better than the ridge estimator (Duncan's multiple range test finds a difference at the five percent level).

In the case when the alignment is unfavourable (ie. $\beta = \beta^{(M)}$) there appears to be some gain from using ridge estimation, provided the bias induced could be considered tolerable.

The effect of using biased estimators is well shown in Figure 5.1(c). In particular the identity

$$\text{mean square error} = \text{variance} + \text{bias}^2$$

can be seen. The IRLS and PC2 estimators have the same ASE, because the variance reduction in PC2 is offset by its bias.

It is encouraging to note that the conclusions that can be drawn from Table 5.2 are in reasonable agreement with the major summary conclusions of [GIBBONS,81] p137-8.

- " 1. All estimators are better than the LS estimators when the underlying coefficient vector is favorable; that is $\beta = \beta_{\text{LS}}$.
2. No estimator is always better than the LS when the underlying coefficient vector is unfavorable; that is $\beta = \beta_{\text{LS}}$.
3. Estimates HKB, GHW and RIDGM performed well overall. "

where HKB refers to [HOERL, KENNARD, BALDWIN, 75]

So in conclusion, there is some evidence to suggest there are definite gains to be made by using some form of biased estimator when collinearity is present in a generalized linear model with a gamma distribution and an artificial collinear structure as was generated. However far more extensive testing, using both simulation and real data, must be carried out before any broader statements can be made.

6.0 Summary and Conclusions

This thesis has investigated suitable definitions for collinearity in the generalised linear model, diagnostic and identification procedures and estimation methods.

The definition of collinearity in a standard linear model as an approximate linear dependency amongst the columns of the predictor matrix X [GUNST,83], is shown to extend naturally to the general setting with the scaling transformation $\tilde{X} = W^{1/2}X$ where W is the weight matrix associated with the general linear model. Using this, it was shown that collinearity in the X system carries over to the general system for all models except the so called p series. That is collinearity in the general setting is model dependent.

The [GUNST,83] extended definition was shown to be mathematically equivalent to small singular values in the transformed predictor matrix $\tilde{X}$. This property was used to extend the [BELSEY,KUH,WELSCH,80] diagnostic and identification scheme. To quantify the model dependency, bounds were derived for the collinearity index of the generalized linear model $\tilde{\alpha}$, in terms of that of the standard linear model collinearity index α .

Although ridge regression intuitively seemed the easiest estimation method to consider, difficulties were encountered in trying to implement it by extending the [HOERL, KENNARD, 70] bound result for the parameter k . An appealing alternative, briefly discussed, was the posterior likelihood approach, which incorporates an "equivalent amount of information". This method is an area for future research.

Given the problems found in trying to set up a Monte Carlo simulation for the estimation of parameters in the generalized linear model, and the criticism directed against previous simulation studies, there is a need for more work in this area too.

While the simulation was extremely restricted in its aims and conclusions, it did succeed in demonstrating that problems exist with estimation in the presence of collinearity and can be alleviated by the methods described. However much more experience is needed both practically and theoretically before some form of biased estimation methods could be advocated for routine use.

Bibliography

[ALLEN,77] Allen,DM "Comment on Simulation of Alternatives to Least Squares" Journal of the American Statistical Association 72 357 95-96.

[BARD,74] Bard,Yonathon Nonlinear Parameter Estimation Academic Press New York.

[BELSEY,KUH,WELSCH,80] Belsey,DA Kuh,E Welsch,RE Regression Diagnostics : Identifying Influential Data and Sources of Collinearity John Wiley & Sons New York.

[BRADLEY,SVRIVASTAVA,79] Bradley,RA Svrivastava,SS "Correlation in Polynomial Regression" American Statistician 33 11-14.

[DEMPSTER,SCHATZOFF,WERMUTH,77] Dempster,AP Schatzoff,M Wermuth,N "A Simulation Study of Alternatives to Ordinary Least Squares" Journal of the American Statistical Association 72 77-106.

[DRAPER,VAN NOSTRAND,79] Draper,NR Van Nostrand,RC "Ridge regression and James Stein Estimators : Review and Comments" Technometrics 21 451-466.

[EDWARDS,69] Edwards,AWF "Statistical Methods in Scientific Inference" Nature 222 1233-1237.

[EDWARDS,69a] Edwards,AWF Likelihood Cambridge University Press London

[GIBBONS,81] Gibbons,DG "A Simulation Study of Some Ridge Estimators" Journal of the American Statistical Association 76 373 131-139

[GUNST,80] Gunst,RF "Comment on A Critique of Some Ridge Regression Methods" Journal of the American Statistical Association 75 369 98-100.

[GUNST,83] Gunst,RF "Regression Analysis with Multicollinear Predictor Variables : Definition, Detection and Effects" Communications in Statistics - Theory and Methods 12(19) 2217-2260.

[HEMMERLE,BRANTLE,78] Hemmerle,WJ Brantle,TF "Explicit and Constrained Generalized Ridge Regression" Technometrics 20 109-120.

[HOERL,62] Hoerl,AE "Application of Ridge Analysis to Regression Problems" Chemical Engineering Progress 58 54-59.

[HOERL,KENNARD,70] Hoerl,AE Kennard,RW "Ridge regression : Biased Estimation for Non-orthogonal Problems" Technometrics 12 56-67.

[HOERL,KENNARD,76] Hoerl,AE Kennard,RW "Ridge Regression : Iterative Estimation of the Biasing Parameter" Communications in Statistics A5 77-88.

[HOERL,KENNARD,BALWIN,75] Hoerl,AE Kennard,RW Baldwin,KF "Ridge Regression : Some Simulations" Communications in Statistics 4 105-123.

[KENDALL,57] Kendall,MG A Course in Multivariate Analysis Griffin London.

[KENDALL,STUART,67.2] Kendall,MG Stuart,A The Advanced Theory of Statistics Volume 2 Inference and Relationship Griffin London.

[LINDLEY,SMITH,72] Lindley,DV Smith,AFM "Bayes estimates for the Linear Model" Journal of the Royal Statistical Society B 42 31-34.

[McCULLAGH,82] McCullagh,P "Categorical Data Analysis" Lecture Notes University of British Columbia.

[McCULLAGH,NELDER,83] McCullagh,P Nelder,JA Generalized Linear Models Chapman and Hall London.

[McDONALD,GALARNEAU,75] McDonald,GC Galarneau,DI "A Monte Carlo Evaluation of Some Ridge Type Estimators" Journal of the American Statistical Association 70 407-416.

[MONTGOMERY,PECK,82] Montgomery,DC Peck,EA Introduction to Linear Regression Analysis J Wiley & Sons New York

[MOYLS,85] Moysls,B Personal communication

[MULLET,76] Mullet,GM "Why Regression Coefficients have the Wrong Sign" Journal of Quality Technology 8 121-126.

[NELDER,77] Nelder,JA "Discussion on Bayes Estimates for the Linear Model" Journal of the Royal Statistical Society Series B 34 18-20.

[NELDER,WEDDERBURN,72] Nelder,JA Wedderburn,RWM "General Linear Models" Journal of the Royal Statistical Society Series A 135 370-384.

[PATEFIELD,77] Patefield,WM "On the Maximized Likelihood Function" Sankhya Series B 39 92-96.

[PREGIBON,81] Pregibon,D "Logistic Regression Diagnostics" The Annals of Statistics 9 4 705-724.

[SCHAEFER,79] Schaefer,RL Multicollinearity in Logistic Regression PhD Thesis University of Michigan #792522D

[SCHAEFER,82] Schaefer,RF "Alternative Estimators in Logistic Regression when the Data are Collinear" Proceedings of the American Statistical Association Statistics and Computing Section 159-164.

[SILVEY,69] Silvey,SD "Multicollinearity and Imprecise Estimation" Journal of the Royal Statistical Society Series B 31 539-552.

[SMITH,CAMPBELL,80] Smith,G Campbell,F "A Critique of Ridge Regression" Journal of the American Statistical Association 75 369 74-81.

[THISTED,76] Thisted,RA "Ridge Regression, Minimax Estimation and Empirical Bayes Methods" Technical Report No 28 Stanford University Division of Biostatistics

[WEDDERBURN,74] Wedderburn,RWM "Quasi Likelihood Functions, Generalized Linear Models and the Gauss-Newton Method" Biometrika 61 3 439-447.