Imperial College of Science and Technology

Department of Mathematics

QUASI-LIKELIHOOD ESTIMATION: EFFICIENCY AND OTHER ASPECTS

by

David Firth

Thesis submitted to the University of London for the degree of Doctor of Philosophy (Ph.D.)

December 1986

ABSTRACT

A quasi-likelihood method has been proposed by Wedderburn (*Biometrika* **61** (1974) , 439-47) for the estimation of parameters in regression models when there is some assumed relationship between the mean and variance of each observation but not necessarily a fully specified likelihood. Some aspects of this method are studied, three main topics being efficiency, standard errors and the connections with some other recent developments.

If the underlying distribution comes from a natural exponential family the quasi-likelihood estimates maximize the likelihood and so have full asymptotic efficiency; under more general distributions this is not the case. The asymptotic efficiency of quasi-likelihood estimation is calculated under some particular distributions, and then more generally via an approximation for 'small departures' from the corresponding natural exponential family. The possibility of refinement of the quasi-likelihood approach, to incorporate additional information about the underlying distribution, is considered.

Standard errors for quasi-likelihood estimates are usually based on the covariance matrix of a large-sample normal approximation to their distribution. Under the same variance assumptions on which quasi-likelihood is based, the form of this covariance matrix is simple and well known. A 'robust' alternative is investigated, based on the asymptotic normal distribution of quasi-likelihood estimates under departures from the assumed second moment behaviour. Aspects discussed include bias correction and the relation of the method to a 'partially Bayes' approach.

A number of authors have recently proposed 'approximate likelihood' methods to allow comparison of different variance specifications. Connections between some of these methods are explored and made explicit.

Acknowledgements

I wish to thank Professor Sir David Cox and Professor P.J. Brown for many valuable and enjoyable discussions during the course of this work. I am also grateful to Anders Ekholm of the University of Helsinki for his kind hospitality during a fruitful visit there; and to Julie McCormack for her help in typing the thesis. The work was supported by a Research Studentship from the Science and Engineering Research Council.

CONTENTS

Appendices

Tables and Figures

CHAPTER 1

Introduction: quasi-likelihood estimation

1.1 A class of models

Consider independent random variables $Y_1,...,Y_n$ with

$$E(Y_i) = \mu_i(\beta) \qquad (i=1,...,n) \qquad (1.1)$$

and

$$\text{var}(Y_i) = \phi V(\mu_i) \qquad (i=1,...,n). \qquad (1.2)$$

McCullagh & Nelder (1983, chapter 8) have called this a 'quasi-likelihood model'. The first-moment specification (1.1) defines the parameters of primary interest, in general a vector $\beta=(\beta_0,...,\beta_p)$; the functions $\mu_i(.)$ are all known. The positive-valued *variance function* $V(.)$ is taken as known, and $\phi>0$ is a *dispersion parameter* whose value may be unknown but is not usually of primary interest. The model is 'semi-parametric' in that the form of the distribution of Y_i as a function of μ_i is only partially specified.

In applications the functions $\mu_i(.)$ often express dependence on explanatory variables $x_{i0},...,x_{ip}$ whose values are known. A form of dependence that has been found particularly useful is the *generalized linear model*, introduced by Nelder & Wedderburn (1972), in which

$$g(\mu_i) = \sum_{r=0}^{p} x_{ir}\beta_r \qquad (i = 1,...,n) \qquad (1.3)$$

for some specified *link function* $g(.)$. This type of model, in which the parameters β enter via a linear component, often has the advantages of ease of interpretation and computational simplicity, while retaining considerable flexibility through the choice of link function; for further discussion see, for example, McCullagh (1984).

This thesis discusses aspects of methods for estimating the regression parameters β in models of the type given by (1.1) and (1.2), with the generalized linear form (1.3) being assumed when it helps to make arguments clearer or more concrete. A more general formulation, allowing correlation between observations, has been given by McCullagh (1983), but this will not be considered here; as McCullagh & Nelder (1983, p169) assert, 'Apart from the multinomial case where dependence is induced by fixing the total, the most interesting class [of quasi-likelihood models] involves uncorrelated observations.'

Perhaps the simplest and most familiar example of a quasi-likelihood model is linear regression with constant error variance, in which

$$E(Y_i) = \mu_i = \sum_{r=0}^{p} x_{ir}\beta_r \qquad (i = 1,...,n) \tag{1.4}$$

and

$$\mathrm{var}(Y_i) = \sigma^2 \qquad (i = 1,...,n)\ ; \tag{1.5}$$

in terms of the earlier notation this has $g(\mu)=\mu$, $V(\mu)=1$ and $\phi=\sigma^2$.

1.2 Estimation of the regression parameters

1.2.1 *Quasi-likelihood estimation*

Given observations $y_1,...,y_n$ from the constant variance linear regression in (1.4) and (1.5), least squares estimates for the parameters $\beta=(\beta_0,...,\beta_p)$ are solutions of the equations

$$\sum_{i=1}^{n} (y_i-\mu_i)x_{ir} = 0 \qquad (r = 0,...,p)\ . \tag{1.6}$$

Wedderburn (1974) has shown how, in more general models of the type given by (1.1) and (1.2), the least squares equations (1.6) may be

generalized to the *quasi-likelihood equations*

$$\sum_{i=1}^{n} \frac{(y_i-\mu_i)}{V(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \beta_r} = 0 \qquad (r = 0,...,p) \qquad (1.7)$$

which in the case of a generalized linear model become

$$\sum_{i=1}^{n} \frac{(y_i-\mu_i)}{V(\mu_i)} \cdot \frac{x_{ir}}{g'(\mu_i)} = 0 \qquad (r = 0,...,p) \ . \qquad (1.8)$$

The name 'quasi-likelihood' arises because $(y_i-\mu_i)/V(\mu_i)$ behaves in many ways like a likelihood-based score function: see Wedderburn (1974,§3) for details. In fact the equations are, in some important cases, the same as maximum likelihood equations based on a particular family of distributions. For example, quasi-likelihood equations with $V(\mu)=1$ are simply the least squares equations, i.e. maximum likelihood based on the assumption $Y_i \sim N(\mu_i,1)$. Similarly quasi-likelihood equations with $V(\mu)=\mu$ are the same as maximum likelihood equations based on a Poisson(μ_i) distribution for Y_i ; and in general quasi-likelihood estimation based on solving (1.7) is the same as maximum likelihood based on the *natural exponential family* with variance function $V(\mu)$, when such a family exists.

Particularly important in what follows will be the large-sample behaviour of quasi-likelihood estimates, as $n \to \infty$ with p fixed. Under the mean and variance assumptions (1.1) and (1.2), McCullagh (1983) has given conditions for the existence of a solution $\tilde{\beta}$ of (1.7) which is consistent and asymptotically normal,

$$n^{1/2}(\tilde{\beta}-\beta) \xrightarrow{\mathcal{D}} N_{p+1}[\ 0\ ,\ n\phi\{D^T \mathrm{diag}(1/V(\mu_i))D\}^{-1}\ ]\ , \qquad (1.9)$$

where $D=(\partial \mu_i/\partial \beta_r)$ is the $n\times(p+1)$ matrix of derivatives of $\mu_i(\beta)$. The main condition needed is the existence of a (positive definite) limiting value for the covariance matrix in (1.9). We will write

$$\mathrm{cov}(\tilde{\beta}) = \phi\{D^T \mathrm{diag}(1/V(\mu_i))D\}^{-1} \qquad (1.10)$$

to describe the *asymptotic* covariance matrix of $\tilde{\beta}$, and similarly for other estimators. In the case of a generalized linear model this may be written as

$$\text{cov}(\tilde{\beta}) = \phi\{X^T \text{diag}(w_i) X\}^{-1} \quad , \tag{1.11}$$

where $$w_i = \frac{1}{V(\mu_i)\{g'(\mu_i)\}^2} \tag{1.12}$$

and $X=(x_{ir})$ is the $n\times(p+1)$ matrix of explanatory variables.

The quasi-likelihood equations remain unbiased estimating equations for β in the sense of, for example, Godambe & Thompson (1978), even if the variance specification (1.2) is incorrect. Under some fairly mild extra conditions (e.g. Inagaki, 1973) $\tilde{\beta}$ remains consistent for β in (1.1) and asymptotically normal; writing the equations as $z_n(\beta)=0$ the asymptotic covariance matrix is $\Lambda_n^{-1} S_n (\Lambda_n^{-1})^T$ where $S_n = \text{cov}\{z_n(\beta)\}$ and $\Lambda_n = \partial E\{z_n(\beta)\}/\partial\beta$ evaluated at the true parameter value. If the true variance of Y_i is ϕ_i, say, these quantities are easily calculated as

$$\Lambda_n = -D^T \text{diag}\{1/V(\mu_i)\} D$$

and $$S_n = D^T \text{diag}[\phi_i/\{V(\mu_i)\}^2] D$$

so the asymptotic covariance matrix is

$$\text{cov}(\tilde{\beta}) = [D^T \text{diag}\{1/V(\mu_i)\} D]^{-1} D^T \text{diag}[\phi_i/\{V(\mu_i)\}^2] D [D^T \text{diag}\{1/V(\mu_i)\} D]^{-1}; \tag{1.13}$$

of course if $\phi_i=\phi V(\mu_i)$ this agrees with (1.10) exactly.

Standard errors for quasi-likelihood estimates may be based on either (1.10) or (1.13); this will be discussed further in Chapter 3.

1.2.2 *Least squares, Gaussian and maximum likelihood estimation*

Some other possible methods for estimating the parameters β in (1.1) are now discussed briefly.

A widely used method in regression problems is *unweighted least squares*, based on the estimating equations

$$\sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \mu_i}{\partial \beta_r} = 0 \qquad (r = 0,...,p) \ . \tag{1.14}$$

This may also be thought of as maximum likelihood based on the assumption $Y_i \sim N(\mu_i, 1)$, or as quasi-likelihood estimation based on variance function $V(\mu)$=constant. While estimates based on (1.14) are quite generally consistent and asymptotically normal, it is easily shown that, unless $V(\mu)$ in (1.2) is constant, they are inferior, in the sense of asymptotic variance, to quasi-likelihood estimates based on the correct variance function; unweighted least squares estimation does not make use of (1.2) at all. In fact it may be shown that under (1.1) and (1.2) the quasi-likelihood equations are asymptotically *optimal* among unbiased estimating equations that are, like least squares, linear in the observations. General versions of this result are given by Morton (1981), Crowder (1982), McCullagh (1983) and Gourieroux, Monfort & Trognon (1984a). The optimality is most easily demonstrated using a model in which β is a single, scalar parameter; a general linear unbiased estimating equation may then be written as

$$\sum_{i=1}^{n} a_i \{y_i - \mu_i(\beta)\} = 0$$

for some $a_1,...,a_n$ which do not depend on the observations. The asymptotic variance of a consistent solution may be calculated as before, and is of the form

$$\left\{ \sum_{i=1}^{n} \phi a_i^2 V(\mu_i) \right\} \left\{ \sum_{i=1}^{n} -a_i \frac{\partial \mu_i}{\partial \beta} \right\}^{-2} \ ;$$

a simple application of Cauchy's inequality now shows that this is minimized when a_i is proportional to $\{V(\mu_i)\}^{-1}(\partial\mu_i/\partial\beta)$, with minimum value $\phi[\Sigma\{V(\mu_i)\}^{-1}(\partial\mu_i/\partial\beta)^2]^{-1}$ as in (1.10).

Unweighted least squares is the same as maximum likelihood based on $Y_i \sim N(\mu_i, 1)$; an obvious alternative method, which makes use of the mean-variance relationship (1.2), is maximum likelihood based on the assumption $Y_i \sim N(\mu_i, \phi V(\mu_i))$. This has been called *Gaussian estimation*, see e.g. Whittle (1961), and the estimating equations are

$$\sum_{i=1}^{n}\left[\frac{(y_i-\mu_i)}{V(\mu_i)} + \frac{V'(\mu_i)}{2\{V(\mu_i)\}^2}\{(y_i-\mu_i)^2-\phi V(\mu_i)\}\right]\frac{\partial\mu_i}{\partial\beta_r} = 0 \qquad (r=0,\ldots,p) \ . \qquad (1.15)$$

These equations, which provide estimates with full asymptotic efficiency when the underlying distribution is $N(\mu_i, \phi V(\mu_i))$, remain unbiased estimating equations under the weaker assumptions (1.1) and (1.2); they are not in general unbiased when (1.2) fails, and in this sense Gaussian estimation is less robust than quasi-likelihood estimation based on (1.7). Another consideration is that estimates based on (1.15), even when consistent and asymptotically normal, have an asymptotic covariance matrix that depends on the third and fourth moments of the underlying distribution; without further assumptions these are unknown, so the precision of Gaussian estimates may be difficult to assess.

Gaussian estimation is one of a large class of *maximum likelihood* methods, each based on some particular family of distributions satisfying the assumptions (1.1) and (1.2); the unbiasedness of the estimating equations (1.15) under only these assumptions is, however, untypical of this class. In general, maximum likelihood equations are not linear or quadratic in the observations, and so cannot be unbiased under assumptions about only the first two moments. On the other hand, maximum likelihood based on the *true* underlying distribution would not only be consistent but also

asymptotically efficient; while this is of no use in a practical sense because the precise form of the true distribution is not assumed known, it provides a means of assessing, theoretically, the efficiency of other methods, in particular quasi-likelihood estimation. This will be pursued further in Chapter 2.

CHAPTER 2

Efficiency of quasi-likelihood estimation

2.1 Introduction

It was noted in Chapter 1 that the quasi-likelihood equations (1.7) are the same as maximum likelihood equations based on the natural exponential family with variance function $V(\mu)$, when such a family exists. In the present chapter we study the loss of efficiency involved in quasi-likelihood estimation, relative to maximum likelihood, when the two methods differ, i.e. when the true distribution satisfies (1.1) and (1.2) but is not from a natural exponential family; the asymptotic covariance matrix of the quasi-likelihood estimate is compared with that of the maximum likelihood estimate based on the true underlying distribution. We focus on three types of problem in which quasi-likelihood methods seem to have been used most: first, in §2.2, we consider models with constant variance, $V(\mu)=1$; section 2.3 then examines models with constant coefficient of variation, $V(\mu)=\mu^2$; and in §2.4 we investigate problems in which there is 'overdispersion' relative to some natural exponential family. In each case we discuss also the possibility of extending the quasi-likelihood approach to incorporate intermediate knowledge about the underlying distribution, such as the form of the third moment.

The arguments of this chapter are expressed in terms of the generalized linear model (1.3), as certain parts of the discussion are thereby made more concrete. Abstraction to the general regression (1.1), while straightforward, provides no additional insight.

2.2 Models with constant variance

2.2.1 *Asymptotic relative efficiency*

Here we restrict attention to quasi-likelihood models with constant variance function, $V(\mu)=1$. Explicitly then, writing the regression as a generalized linear model, we assume

$$E(Y_i)=\mu_i \ , \quad g(\mu_i)=\sum_{r=0}^{p} x_{ir}\beta_r \ \text{ and } \ \text{var}(Y_i)=\phi \qquad (i=1,...,n) \ . \tag{2.1}$$

The quasi-likelihood equations for estimating $\beta=\beta_0,...,\beta_p$ are now simply the least squares equations,

$$\sum_{i=1}^{n} (y_i-\mu_i)\frac{x_{ir}}{g'(\mu_i)} = 0 \qquad (r=0,...,p) \ . \tag{2.2}$$

In the special case where each Y_i is from a normal distribution with mean μ_i and variance ϕ, these equations are maximum likelihood equations and so, subject to standard conditions, least squares estimation has full efficiency for fixed p as $n\to\infty$. In general, however, the underlying distribution may be non-normal and there is then some loss of efficiency.

The efficiency of least squares estimation in linear models with a constant error variance, i.e. models of the form (2.1) with $g(\mu)=\mu$, has been investigated by Cox & Hinkley (1968). The development here, although a slight generalization, will be substantially the same as theirs: the main points will be sketched and an error will be corrected. We follow Cox & Hinkley in making the additional assumption that the distribution of $\epsilon_i=Y_i-\mu_i$ is the same for all i, so that μ_i is simply a location parameter for the distribution of Y_i.

Consider first the situation where ϕ is known. Write the log likelihood from a sample $y_1,...,y_n$ as

$$l(\beta,\phi) = \Sigma \log f(y_i;\ \mu_i\ ,\phi) = \Sigma\ l_i(\mu_i,\phi) \ ,$$

say. Then, assuming sufficient regularity, the Fisher information matrix has

elements

$$E(-\partial^2 l/\partial\beta_r\partial\beta_s) = \sum_{i=1}^{n} E(-\partial^2 l_i/\partial\mu_i^2)\ \frac{x_{ir}x_{is}}{\{g'(\mu_i)\}^2}\ .$$

This is simplified by the assumption that μ_i is a location parameter only, which implies that $E(-\partial^2 l_i/\partial\mu_i^2)=A_\epsilon$, say, the same for all $i=1,\ldots,n$. Thus

$$E(-\partial^2 l/\partial\beta_r\partial\beta_s) = A_\epsilon \sum_{i=1}^{n} \frac{x_{ir}\,x_{is}}{\{g'(\mu_i)\}^2} = A_\epsilon\{X^T\mathrm{diag}(w_i)X\} \qquad (2.3)$$

where w_i is as defined in (1.12). Under standard conditions, assumed satisfied here, the maximum likelihood estimate $\hat{\beta}$ is asymptotically $(p+1)$-variate normal with covariance matrix given by the inverse of (2.3). Comparison with (1.11) shows that $\mathrm{cov}(\hat{\beta})$ and $\mathrm{cov}(\tilde{\beta})$ are proportional and there is a single measure of *asymptotic relative efficiency*,

$$\mathrm{ARE}(\tilde{\beta} : \hat{\beta}) = (\phi A_\epsilon)^{-1} = \phi^{-1}\{E(-\partial^2 l_i/\partial\mu_i^2)\}^{-1}, \qquad (2.4)$$

where l_i is the true log likelihood for a single observation.

To extend the analysis to problems with ϕ unknown it is convenient to assume that the general mean is included in the model and is not a parameter of primary interest; then, without loss of generality, suppose that

$$x_{i0} = 1 \quad (i=1,\ldots,n) \quad , \qquad \sum_{i=1}^{n} x_{ir} = 0 \quad (r=1,\ldots,p) \qquad (2.5)$$

and take $\beta_1,\ldots,\beta_p$ as parameters of interest, orthogonal to the general mean. Again examining the Fisher information matrix we find that

$$E(-\partial^2 l/\partial\beta_r\partial\phi) = \sum_{i=1}^{n} E(-\partial^2 l_i/\partial\mu_i\partial\phi)\ \frac{x_{ir}}{g'(\mu_i)} \qquad (r=1,\ldots,p) \qquad (2.6)$$

and

$$E(-\partial^2 l/\partial\beta_r\partial\beta_0) = \sum_{i=1}^{n} E(-\partial^2 l_i/\partial\mu_i^2)\ \frac{x_{ir}x_{i0}}{\{g'(\mu_i)\}^2} \qquad (r=1,\ldots,p)\ . \qquad (2.7)$$

The expectations $E(\partial^2 l_i/\partial\mu_i\partial\phi)$ and $E(\partial^2 l_i/\partial\mu_i^2)$ here do not depend on i. Thus, provided $g(\mu)=\mu$, the identity link, the expressions in (2.6) and (2.7) all vanish on account of the orthogonality (2.5). The full information matrix then partitions into the form

$$\begin{bmatrix} I_1 & 0 \\ 0 & I_2 \end{bmatrix} \tag{2.8}$$

where I_1 refers to β_0 and ϕ, and I_2 to $\beta_1,\ldots,\beta_p$; the elements of I_2 are still given by (2.3). So if we consider only the parameters of interest, $\beta=(\beta_1,\ldots,\beta_p)$, the asymptotic covariance matrices of $\hat{\beta}$ and $\tilde{\beta}$ remain proportional, and asymptotic relative efficiency is still given by the scalar quantity $(\phi A_\epsilon)^{-1}$.

In most applications of constant variance models the effects are taken as additive, i.e. the link function is the identity. In models with other link functions there is a problem with the above treatment for unknown ϕ, for then there is some non-orthogonality between the two sets of parameters $(\beta_1,\ldots,\beta_p)$ and (β_0,ϕ). This in turn means that there is no single constant of proportionality between the asymptotic covariance matrices of $\tilde{\beta}$ and $\hat{\beta}$. However use of $(\phi A_\epsilon)^{-1}$ as a measure of asymptotic relative efficiency can still be justified, particularly if β_0, the 'intercept' parameter, is large compared with the other effects. For when this is the case the link function $g(\mu)$ may be approximated by a linear function over the relevant range of μ-values; then $g'(\mu)$ is approximately constant over this range and the expectations (2.6) and (2.7) are close to zero. This argument could be made formal via an expansion of $g'(\mu)$ about $\mu_0=g^{-1}(\beta_0)$; the point is that there may still be *approximate* proportionality between the matrices $\mathrm{cov}(\tilde{\beta})$ and $\mathrm{cov}(\hat{\beta})$, with $(\phi A_\epsilon)^{-1}$ remaining an index of asymptotic relative efficiency.

We proceed now to calculate A_ϵ for some possible families of underlying distributions. For the Normal(μ_i,ϕ) family, of course, $E(-\partial^2 l_i/\partial\mu_i^2)=\phi^{-1}$, i.e. asymptotic efficiency is 1, since in this case least squares and maximum likelihood coincide. Cox & Hinkley consider two other special families, the Pearson Type VII and the log gamma, and then generalize the calculations by considering an Edgeworth series in which, in terms of a notional parameter N, the higher order standardized cumulants

are assumed to behave as $\gamma_1=O(N^{-\frac{1}{2}})$, $\gamma_2=O(N^{-1})$, etc . This leads after appreciable calculation, see Appendix 1, to the approximation

$$E(-\partial^2 l_i/\partial\mu_i^2) = \phi^{-1}\{ 1 + \tfrac{1}{2}\gamma_1^2 + (\tfrac{7}{4}\gamma_1^4 - \tfrac{5}{4}\gamma_1^2\gamma_2 + \tfrac{1}{6}\gamma_2^2) + o(N^{-2}) \} \qquad (2.9)$$

which corrects an error in the $O(N^{-2})$ term of Cox & Hinkley's expression (20). The quality of this approximation may be checked using the Type VII and log gamma distributions as examples; in both cases the 'non-normality' parameter N can be identified with the shape parameter of the distribution. The efficiency values given by Cox & Hinkley for the Type VII distribution remain unaffected by the above correction, since $\gamma_1=0$ in that case. For the log gamma, Table 1 gives the exact efficiency and that based on the corrected approximation (2.9); the exact efficiency is $(\nu\psi'(\nu))^{-1}$, where ν is the index of the underlying gamma distribution and $\psi'(.)$ is the trigamma function. The asymptotic formulae

$$\psi'(\nu) = \nu^{-1}\{1 + \tfrac{1}{2}\nu^{-1} + \tfrac{1}{6}\nu^{-2} + o(\nu^{-2})\}$$

$$\psi''(\nu) = -\nu^{-2}\{1 + \nu^{-1} + o(\nu^{-1})\}$$

and $$\psi^{(3)}(\nu) = \nu^{-3}\{2 + 3\nu^{-1} + o(\nu^{-1})\}$$

(Abramowitz & Stegun, 1965, p260) may be used as a check on (2.9): calculation from the exact efficiency gives

$$\phi A_\epsilon = \nu\psi'(\nu) = 1 + \tfrac{1}{2}\nu^{-1} + \tfrac{1}{6}\nu^{-2} + o(\nu^{-2}) \; ; \qquad (2.10)$$

the log gamma distribution has $\gamma_1=\psi''(\nu)\{\psi'(\nu)\}^{-3/2}$ and $\gamma_2=\psi^{(3)}(\nu)\{\psi'(\nu)\}^{-2}$, and the asymptotic approximation obtained by substituting these into (2.9) is easily seen to agree with (2.10). Table 1 shows that the approximation, taken either to order ν^{-1} or to order ν^{-2}, is close even for ν-values as small as 1.0 .

TABLE 1. *Asymptotic efficiency of least squares estimates when the errors have a log gamma distribution*

ν	γ_1	γ_2	Exact efficiency	Efficiency from (2.9) to $O(\nu^{-2})$	Efficiency from (2.9) to $O(\nu^{-1})$
0.5	-1.535	4.000	0.405	0.360	0.463
1.0	-1.140	2.400	0.608	0.601	0.606
2.5	-0.688	0.931	0.816	0.818	0.809
5.0	-0.469	0.437	0.904	0.904	0.902
10.0	-0.324	0.210	0.951	0.951	0.950

Perhaps the main observation to be made about the form of (2.9) in general is that the leading term involves only γ_1^2, suggesting that *skewness* is the most important factor affecting the efficiency of least squares, at least for small departures from normality. If γ_1 is zero the leading term is of order N^{-2} and involves only γ_2^2 , so *kurtosis* is then an appropriate index of the effect of non-normality.

2.2.2 *Refinement of least squares estimation*

The least squares estimating equations are based only on second-moment assumptions ; we consider now the possibility of improved efficiency when there is information about higher moments of the underlying distribution.

The least squares equations for estimating $\beta_0,\ldots,\beta_p$ may be written as

$$\sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \mu_i} \log f_N(y_i;\, \mu_i, \phi) \right\} \frac{x_{ir}}{g'(\mu_i)} = 0 \qquad (r=0,\ldots,p) \qquad (2.11)$$

where $f_N(y;\, \mu,\phi)=(2\pi\phi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y-\mu)^2/\phi\}$ is the normal density; use of the true density $f(y;\, \mu,\phi)$ in place of $f_N(y;\, \mu,\phi)$ would give the maximum likelihood equations. It has already been noted that, to first order, the loss of efficiency in using least squares depends on the skewness of the underlying distribution; if we suppose the standardized skewness γ_1 to be

known then, subject to regular behaviour of higher order cumulants, the Edgeworth series approximation

$$f(y;\ \mu,\phi) \cong f_N(y;\ \mu,\phi)[\ 1 + \tfrac{1}{6}\gamma_1 H_3\{(y-\mu)/\sqrt{\phi}\}\ ] \qquad (2.12)$$

should be closer than $f_N(y;\mu,\phi)$, at least if γ_1 is small. This suggests a refinement of least squares which replaces $f_N(y;\mu,\phi)$ in (2.11) with an approximation like (2.12) ; to avoid problems with negative values we shall in fact consider using

$$f^*(y;\ \mu,\phi) = f_N(y;\ \mu,\phi)\exp[\tfrac{1}{6}\gamma_1 H_3\{(y-\mu)/\sqrt{\phi}\}] \qquad (2.13)$$

in place of f_N in (2.11). Explicitly, then, the estimating equations become

$$\sum_{i=1}^{n}\left[\frac{y_i-\mu_i}{\sqrt{\phi}} - \tfrac{1}{2}\gamma_1\left\{\frac{(y_i-\mu_i)^2}{\phi} - 1\right\}\right]\frac{x_{ir}}{g'(\mu_i)} = 0 \qquad (r=0,\ldots,p)\ . \qquad (2.14)$$

Suppose first that ϕ is known. The equations (2.14) are unbiased estimating equations for β and under standard conditions (e.g. Inagaki, 1973) they have a solution, β^* say, which is consistent and asymptotically normal with mean β_T, the true value; as in §1.2, if we write the equations as $z_n(\beta)=0$ the asymptotic covariance matrix is $\mathrm{cov}(\beta^*)=\Lambda_n^{-1}S_n(\Lambda_n^{-1})^T$, where $S_n=\mathrm{cov}\{z_n(\beta_T)\}$ and $\Lambda_n=[\partial E\{z_n(\beta)\}/\partial\beta]$ evaluated at β_T . In the present case this gives

$$\mathrm{cov}(\beta^*) = \phi(1 - \tfrac{1}{2}\gamma_1^{\,2} + \tfrac{1}{4}\gamma_1^2\gamma_2)\{X^T\mathrm{diag}(w_i)X\}^{-1}, \qquad (2.15)$$

proportional to $\mathrm{cov}(\tilde{\beta})$. Comparison with the reciprocal of (2.9) gives the asymptotic efficiency of β^*, relative to the maximum likelihood estimator $\hat{\beta}$, as

$$\mathrm{ARE}(\beta^*\ ;\ \hat{\beta}) = \frac{1 - \tfrac{1}{2}\gamma_1^2 + (\tfrac{5}{4}\gamma_1^2\gamma_2 - \tfrac{3}{2}\gamma_1^4 - \tfrac{1}{6}\gamma_2^2) + o(N^{-2})}{1 - \tfrac{1}{2}\gamma_1^2 + \tfrac{1}{4}\gamma_1^2\gamma_2} \qquad (2.16)$$

$$= 1 - \tfrac{1}{6}(\gamma_2 - 3\gamma_1^2)^2 + o(N^{-2})\ ,$$

where we assume $\gamma_1=O(N^{-\frac{1}{2}})$, $\gamma_2=O(N^{-1})$, etc., as before.

The asymptotic efficiency of β^* should be compared with the numerator of (2.16), the asymptotic efficiency of the least squares estimate $\tilde{\beta}$. We see that exploitation of third moment information reduces the loss of efficiency from $O(N^{-1})$ to $O(N^{-2})$, and an index for the efficiency of β^* relative to maximum likelihood is $\gamma_2-3\gamma_1^2$; when this is zero the loss of efficiency is reduced still further to $O(N^{-3})$. As a numerical example consider again the log gamma distribution; Table 2 gives the asymptotic efficiency of β^* at the same values of the shape parameter used in Table 1. Comparison with the efficiency of least squares shows the third-moment refinement to be very effective for small to moderate departures from normality, although least squares is more efficient when ν is less than about 1 , which represents extreme non-normality.

TABLE 2. *Asymptotic efficiency of 'refined least squares' estimates under log gamma errors*

ν	0.5	1.0	2.5	5.0	10.0
$ARE(\beta^*\ ;\ \hat{\beta})$	0.186	0.538	0.934	0.989	0.998

The equations (2.14) may be shown to have a certain optimality property among quadratic unbiased estimating equations. To illustrate, consider the 'single sample' model $E(Y_i)=\mu$, $\mathrm{var}(Y_i)=\phi$ with ϕ known. Possible unbiased estimating equations for μ based on a single observation are $y_i-\mu=0$ and $(y_i-\mu)^2-\phi=0$; assuming symmetry, represent the combination of these over all observations as

$$\sum_{i=1}^{n} [\ \lambda(y_i-\mu) + (1-\lambda)\{(y_i-\mu)^2-\phi\}\ ] = 0 \qquad (-\infty < \lambda < \infty) \qquad (2.17)$$

where λ is a scalar weight. The asymptotic variance of a consistent solution of (2.17) may be calculated as before and shown to be minimized at $\lambda^0=\{2+\gamma_2\}/\{2+\gamma_2-(\gamma_1/\sqrt{\phi})\}$ with corresponding minimum variance $n^{-1}\phi\{1-\gamma_1^2/(\gamma_2+2)\} = n^{-1}\phi\{1-\frac{1}{2}\gamma_1^2+\frac{1}{4}\gamma_1^2\gamma_2+o(N^{-2})\}$; this is the same, to $O(N^{-2})$,

as the variance given by (2.15), so the refined estimating equation (2.14) is 'optimal to second order' among quadratic unbiased estimating equations. Of course if γ_2 is also known we can use $\lambda=\lambda^0$ in (2.17) to achieve exact optimality within this class.

An interesting feature here is the behaviour of λ^0, the optimizing value of λ, when γ_1 is large. The familiar restriction on third and fourth cumulants,

$$\gamma_1^2 \leqslant \gamma_2 + 2$$

implies, in particular, that $\lambda^0 \to 1$ as $\gamma_1 \to \pm\infty$. Thus it appears that, as skewness increases, the optimal quadratic estimating equation reverts towards least squares. However the asymptotic relative efficiency of these two methods is $1-\gamma_1^2/(\gamma_2+2)$, which need not have a limiting value as $\gamma_1^2 \to \infty$; in particular it does not, in general, converge to 1.

Solution of (2.14) generally requires iteration. It is illuminating, though, to consider a case where there is an explicit solution: in the single sample problem with ϕ known the solution of (2.17) with $\lambda=2\sqrt{\phi}/(2\sqrt{\phi}-\gamma_1)$ is of the form

$$\mu^* = \tilde{\mu} - (\tfrac{1}{2}\gamma_1/\sqrt{\phi}) \{ n^{-1}\Sigma(y_i-\tilde{\mu})^2 - \phi\} + O(\gamma_1^2) , \qquad (2.18)$$

where in this case $\tilde{\mu}$ is simply the sample mean. The form of (2.18) has intuitive appeal ; for example if $\gamma_1>0$ and the sample variance exceeds the known population variance then we estimate the population mean to be less than the sample mean, reflecting the preference of outliers for the upper tail.

In problems where the variance is unknown, but still assumed constant, consider replacing ϕ in (2.14) by $n^{-1}\Sigma(y_i-\mu_i)^2=\tilde{\phi}$, say. Then in the single sample model just discussed, the 'refined' estimating equation reverts to least squares, whatever the value of γ_1 . However in more complicated models the assumption of constant variance may make

knowledge of the standardized skewness more useful. To illustrate, consider regression through the origin, $\mu_i=x_i\beta$; the estimating equation for β in this case may be written as

$$(\tilde{\phi})^{-\frac{1}{2}}\Sigma\,(y_i-\mu_i)x_i \;-\; \tfrac{1}{2}\gamma_1(\tilde{\phi})^{-1}\Sigma\,\{(y_i-\mu_i)^2-\tilde{\phi}\}x_i = 0 \quad . \tag{2.19}$$

Again writing $\epsilon_i=Y_i-\mu_i$, the two sums in (2.19) give the sample covariances of x_i with ϵ_i and with ϵ_i^2 ; equating either of these covariances individually to zero would yield an unbiased estimating equation for β , and (2.19) combines the two estimating functions in a way that makes use of the known skewness. Efficiency considerations parallel those in §2.2.1 for maximum likelihood with ϕ unknown; in particular it may be shown that, with the orthogonality relations (2.5) and the identity link, the asymptotic covariance matrix of $\beta^*=(\beta_1^*,\ldots,\beta_p^*)$ is still given by (2.15).

In general the construction of refined estimators based on exact knowledge of the standardized third moment would seem to be mainly of theoretical interest, since in practice it is unlikely that γ_1 can reasonably be specified at all precisely except possibly as zero. When there is imprecise information about γ_1 the particular refinement studied here has the advantage that the estimating equations remain unbiased even when an incorrect value of γ_1 is used; and, as shown in Appendix 2, the asymptotic covariance matrix (2.15) is unchanged if an estimate $\hat{\gamma}_1=\gamma_1+O_p(n^{-\frac{1}{2}})$ replaces γ_1 in (2.14).

If γ_1 is known to be zero the Edgeworth series approximation

$$f(y;\mu,\phi) \;\cong\; f_N(y\;;\mu,\phi)[\;1 \;+\; (\gamma_2/24)H_4\{(y-\mu)/\sqrt{\phi}\}\;]$$

suggests, in the same way as before, 'refined' estimating equations

$$\sum_{i=1}^{n}\{\,(1+\tfrac{1}{2}\gamma_2)(y_i-\mu_i) \;-\; (\gamma_2/6\phi)(y_i-\mu_i)^3\}\,\frac{x_{ir}}{g'(\mu_i)} = 0 \qquad (r=0,\ldots,p)\;. \tag{2.20}$$

The asymptotic covariance matrix of a consistent solution β^* is

$$\mathrm{cov}(\beta^*) \;=\; \{\;1 \;-\; \gamma_2^2/6 \;+\; O(N^{-3})\;\}\;\mathrm{cov}(\tilde{\beta})\;,$$

and comparison with the reciprocal of (2.9) shows that the loss of asymptotic efficiency is reduced from $O(N^{-2})$ to $O(N^{-3})$ by incorporating information about kurtosis in this way. In practice γ_2 is not usually known but, as shown in Appendix 2, an estimate $\hat{\gamma}_2=\gamma_2+O_p(n^{-\frac{1}{2}})$ gives the same asymptotic efficiency.

2.3 Models with constant coefficient of variation

2.3.1 *Asymptotic relative efficiency*

In this section we consider quasi-likelihood models that have variance function $V(\mu_i)=\mu_i^2$, i.e. $\text{var}(Y_i)=\phi\{E(Y_i)\}^2$, so that $\sqrt{\phi}$ is a *constant coefficient of variation.* A general discussion of this type of model is given, with examples, by McCullagh & Nelder (1983, chapter 7). One type of problem in which such models have found practical application is where data are in the form of continuous measurements whose variance increases with the mean. In particular, measurements with multiplicative error, $Y_i=\mu_i\epsilon_i$, where $E(\epsilon_i)=1$ and $\text{var}(\epsilon_i)=\phi$, have a constant coefficient of variation. A second application is in modelling waiting times with 'overdispersion' relative to the standard assumption of an exponential distribution.

Here the quasi-likelihood equations are obtained by putting $V(\mu_i)=\mu_i^2$ in (1.7); for positive observations they are the same as the maximum likelihood equations that would arise if we assumed each Y_i to have a *gamma* distribution with mean μ_i and index $\nu=\phi^{-1}$, i.e. to have density

$$f_G(y_i\,;\,\mu_i\,,\phi) = \frac{\exp(-y_i\nu/\mu_i)\;\; y_i^{\nu-1}}{(\mu_i/\nu)^{\nu}\,\Gamma(\nu)} \qquad (y_i>0\,;\;\mu_i,\nu>0). \qquad (2.21)$$

When the true underlying family of distributions is *not* the gamma family the quasi-likelihood estimate remains, under certain conditions, consistent and asymptotically normal with covariance matrix given by (1.11), which we now compare with the asymptotic covariance matrix of the maximum likelihood estimate. To simplify the analysis, as in §2.2.1, we make an assumption limiting the extent to which the parameter μ_i affects the distribution of Y_i: here it is assumed that the distribution of $\epsilon_i=Y_i/\mu_i$ is the same for all i, so that μ_i is simply a scale parameter for the

distribution of Y_i . Then arguments leading to a single measure of efficiency closely parallel those given in §2.2.1 for the constant variance model.

Consider first the case when the value of the coefficient of variation is known. The log likelihood may be written as

$$l(\beta,\phi) = \sum_{i=1}^{n} l_i(\mu_i,\phi) = \sum_{i=1}^{n} \{ h(\epsilon_i;\phi) - \log \mu_i \} ,$$

where $h(\epsilon_i ; \phi)$ is the log-density of $\epsilon_i = Y_i/\mu_i$; the $\{\epsilon_i\}$ are assumed to be identically distributed and so $h(\epsilon_i ; \phi)$ does not depend on μ_i. Again assuming that the usual regularity conditions are satisfied, the (r,s)th element of the Fisher information matrix is

$$\begin{aligned} E(-\partial^2 l/\partial\beta_r\partial\beta_s) &= -\sum_{i=1}^{n} \frac{x_{ir}x_{is}}{\{g'(\mu_i)\}^2} E(\partial^2 l_i/\partial\mu_i^2) \\ &= -\sum_{i=1}^{n} \frac{x_{ir}x_{is}}{\{g'(\mu_i)\}^2} E\left[\frac{\partial}{\partial\mu_i}\left\{\frac{-h'(\epsilon_i;\phi)y_i}{\mu_i^2} - \frac{1}{\mu_i}\right\}\right] \\ &= -\sum_{i=1}^{n} \frac{x_{ir}x_{is}}{\{g'(\mu_i)\}^2} E\left\{\frac{y_i^2}{\mu_i^4}h''(\epsilon_i;\phi) + \frac{2y_i}{\mu_i^3}h'(\epsilon_i;\phi) + \frac{1}{\mu_i^2}\right\} \end{aligned}$$

which can be written

$$E(-\partial^2 l/\partial\beta_r\partial\beta_s) = \left[-\sum_{i=1}^{n} \frac{x_{ir}x_{is}}{\mu_i^2\{g'(\mu_i)\}^2} \right] E\{\epsilon_i^2 h''(\epsilon_i;\phi) + 2\epsilon_i h'(\epsilon_i;\phi) + 1\} \quad (2.22)$$

since the distribution of ϵ_i is the same for all i. Now the asymptotic covariance matrix of $\hat{\beta}$, the maximum likelihood estimate, is given by the inverse of the information matrix with elements (2.22). Comparison with (1.11) shows immediately that $\mathrm{cov}(\hat{\beta})$ and $\mathrm{cov}(\tilde{\beta})$ are proportional, with asymptotic relative efficiency given by the scalar quantity

$$\begin{aligned} \mathrm{ARE}(\tilde{\beta} ; \hat{\beta}) &= [-\phi E\{\epsilon_i^2 h''(\epsilon_i;\phi) + 2\epsilon_i h'(\epsilon_i;\phi) + 1\}]^{-1} \quad (2.23) \\ &= (\phi A_\epsilon)^{-1} , \end{aligned}$$

say, where now $A_\epsilon = \mu_i^2 E(-\partial^2 l_i/\partial\mu_i^2)$, which does not depend on i. It should be noted that the relative efficiency does not depend on either the model

matrix X or the link function $g(.)$.

Consider now the case when ϕ is unknown. As in §2.2.1, it is convenient to assume the presence of an intercept term in the linear model, and the orthogonality relations (2.5). We can write

$$\frac{\partial^2 l}{\partial \beta_r \partial \phi} = \sum_{i=1}^{n} \frac{x_{ir}}{\mu_i g'(\mu_i)} \left\{ \epsilon_i \frac{\partial^2 h(\epsilon_i\,;\,\phi)}{\partial \epsilon_i \partial \phi} \right\}$$

and the expectation of the bracketed term here does not depend on i, so

$$E(-\partial^2 l/\partial \beta_r \partial \phi) = 0 \qquad (r=1,...,p)\,, \tag{2.24}$$

provided that $g'(\mu) \propto \mu^{-1}$, i.e. provided we have the *logarithmic link function* $g(\mu)=\log(\mu)$. The same conditions can be shown to imply also that

$$E(-\partial^2 l/\partial \beta_r \partial \beta_0) = 0 \qquad (r=1,...,p)\,, \tag{2.25}$$

so that the information matrix for the full set of unknown parameters again partitions as in (2.8) into two parts, one part referring to the parameters $(\beta_1,...,\beta_p)$ and the other part to (β_0,ϕ). Thus if we consider the parameters of interest to be $\beta=(\beta_1,...,\beta_p)$ we see that $\mathrm{cov}(\hat{\beta})$ is still given by the inverse of the matrix with (r,s)th element as in (2.22), and the asymptotic relative efficiency for estimating these parameters is still $\mathrm{ARE}(\tilde{\beta}\,;\,\hat{\beta})=(\phi A_\epsilon)^{-1}$ as in (2.23).

When ϕ is unknown and the link function is some function other than $g(\mu)=\log(\mu)$, orthogonality between the two sets of parameters $(\beta_1,...,\beta_p)$ and (β_0,ϕ) no longer holds; this creates a problem for the above analysis, since it implies that $\mathrm{cov}(\hat{\beta})$ is no longer proportional to $\mathrm{cov}(\tilde{\beta})$ and so there can be no single measure of asymptotic relative efficiency. However, remarks corresponding to those made about the link function in §2.2.1 also apply here. First, the log link is frequently the one used in practice; it gives a model with multiplicative effects, which seems appropriate when a constant coefficient of variation is assumed, particularly if that assumption is based on the notion of multiplicative errors mentioned above. Secondly, as explained in section 2, the orthogonality relations (2.24) and

(2.25) may still hold *approximately*, especially if the intercept term β_0 is large compared with the other terms in the linear model; this would imply approximate proportionality between the matrices cov($\hat{\beta}$) and cov($\tilde{\beta}$), and $(\phi A_\epsilon)^{-1}$ would remain a suitable measure of asymptotic relative efficiency.

It now remains to calculate A_ϵ for some families of distributions satisfying $E(Y_i)=\mu_i$, $\text{var}(Y_i)=\phi\mu_i^2$. Consider first just four illustrative examples; from now on, when convenient, the subscript i will be dropped so that y, μ and l will refer, respectively, to an observation, its expectation and the corresponding log likelihood. The four examples are :

(i) a normal family, $N(\mu, \phi\mu^2)$, which has cumulants $\kappa_r=0$ for all $r>2$;

(ii) a lognormal family with density, on $y>0$,

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

and higher order cumulants $\kappa_3=\mu^3\phi^2(3+\phi)$, $\kappa_4=\mu^4\phi^3(16+15\phi+6\phi^2+\phi^3)$, etc.;

(iii) an inverse Gaussian family with density, on $y>0$,

$$f(y;\ \mu,\ \phi) = (2\pi\phi y^3/\mu)^{-\frac{1}{2}} \exp\left\{ - \frac{\mu(y^2 - 2\mu y + \mu^2)}{2\phi\mu^2 y} \right\}$$

which has higher order cumulants $\kappa_r=\mu^r\phi^{r-1}(2r-3)!/\{2^{r-2}(r-2)!\}$; and

(iv) a mixture of exponential distributions in which Y~exponential with mean M, where M has an inverse gamma distribution with mean μ and squared coefficient of variation $\frac{1}{2}(\phi-1)$, giving the Pareto density

$$f(y;\ \mu,\phi) = \frac{\delta\{(\delta-1)\mu\}^\delta}{\{y + (\delta-1)\mu\}^{\delta+1}} \qquad (y>0) , \qquad (2.26)$$

where $\delta=2\phi/(\phi-1)$ and ϕ is restricted here to be greater than 1.

The families (i) and (ii) are perhaps the most obvious alternatives to the gamma distribution as models for data with constant coefficient of variation; family (iii), a particular parameterization of the inverse Gaussian distribution, is rather less familiar; and family (iv) is an example of a

family giving overdispersion relative to the exponential distribution. All four families satisfy the assumption that μ is a scale parameter, i.e. the distribution of Y/μ does not depend on μ ; calculation of A_ϵ gives

(i) for the normal distribution, $A_\epsilon = (1+2\phi)/\phi$ $(\phi>0)$,

(ii) for the lognormal, $A_\epsilon = \{\log(1+\phi)\}^{-1}$ $(\phi>0)$,

(iii) for the inverse Gaussian, $A_\epsilon = (1+\frac{1}{2}\phi)/\phi$ $(\phi>0)$, and

(iv) for the mixed exponential, $A_\epsilon = \phi/(2\phi-1)$ $(\phi>1)$.

To generalize these results beyond the examples, first note that family (iv) is different from (i)-(iii) since it permits only $\phi>1$; generalization of the calculations for family (iv) is deferred until §2.4, where a wider class of 'overdispersed' distributions is discussed. The families (i)-(iii), however, have much in common, including limiting normality and a limiting value of 1 for the efficiency $(\phi A_\epsilon)^{-1}$ as $\phi\to 0$. The same properties are shared by the gamma family (2.21) which has, of course, $(\phi A_\epsilon)^{-1}=1$ for all $\phi>0$. A more specific observation may be made, about the behaviour of higher order cumulants as $\phi\to 0$, which may be summarized as

$$\kappa_r = \{ \tau_r + (r-1)! \}\phi^{r-1}\mu^r \qquad (r=3,4,\ldots) , \tag{2.27}$$

where the constants τ_r depend on the particular distribution and are $O(1)$ as $\phi\to 0$; for the gamma distribution, $\tau_r=0$ for all r . Notice that the re-scaled variate $Z=Y/(\mu\phi)$ has cumulants $\kappa_r=\{\tau_r+(r-1)!\}\phi^{-1}$, so the behaviour (2.27) is connected with a sort of 'approximate infinite divisibility'.

The efficiency calculations for families (i)-(iii) are now generalized using the formal series expansion

$$f(y;\,\mu,\phi) \;=\; f_G(y;\,\mu,\phi)\,\{\,1 + c_3 L_3^{(\alpha)}(z) + c_4 L_4^{(\alpha)}(z) + \ldots\,\} \tag{2.28}$$

of the true underlying density about the gamma density (2.21). Here $L_r^{(\alpha)}(.)$ is the generalized Laguerre polynomial of degree r , $\alpha=(\phi^{-1}-1)$ and $z=y/(\mu\phi)$. The polynomials $\{L_r^{(\alpha)}(z)\}$ are orthogonal with respect to the gamma density

$f_G(y;\mu,\phi)$, and so it is a straightforward task to show that the coefficients are, in terms of the constants $\{\tau_r\}$ of (2.27),

$$c_3 = \frac{\phi^2}{(1+\phi)(1+2\phi)}(-\tau_3) ,$$

$$c_4 = \frac{\phi^3}{(1+\phi)(1+2\phi)(1+3\phi)}(\tau_4 - 12\tau_3) ,$$

$$c_5 = \frac{\phi^4}{(1+\phi)(1+2\phi)(1+3\phi)(1+4\phi)}(-\tau_5 + 20\tau_4 - 120\tau_3) ,$$

$$c_6 = \frac{\phi^4}{(1+\phi)(1+2\phi)(1+3\phi)(1+4\phi)(1+5\phi)}\{10\tau_3^2 + \phi(\tau_6 - 30\tau_5 + 300\tau_4 - 1200\tau_3)\} ,$$

and so on, with $c_7=O(\phi^5)$, $c_8=O(\phi^6)$, $c_9=O(\phi^6)$, $c_{10}=O(\phi^7)$, etc. as $\phi\to 0$, assuming $\tau_r=O(1)$ for all r. The evaluation of ϕA_ϵ from (2.28) proceeds by taking logarithms, differentiating twice with respect to μ and then taking expectations. After a rather large amount of calculation, details of which are given in Appendix 3, it is found that

$$\phi A_\epsilon = \phi\mu^2 E(-\partial^2 l/\partial\mu^2)$$

$$= 1 + \frac{\phi}{2}\tau_3^2 + \frac{\phi^2}{12}(156\tau_3^2 - 36\tau_3\tau_4 + 2\tau_4^2 + 120\tau_3^3 - 15\tau_3^2\tau_4 + 21\tau_3^4) + o(\phi^2). \quad (2.29)$$

The $O(\phi)$ term here contains no contribution from terms in the expansion (2.28) higher than $c_3 L_3^{(\alpha)}(z)$; the $O(\phi^2)$ term involves taking (2.28) up to the term $c_6 L_6^{(\alpha)}(z)$.

The expression (2.29) is readily checked against our families (i)-(iii). For example, the $N(\mu, \phi\mu^2)$ family has $\tau_3=-2$ and $\tau_4=-6$ so (2.29) gives $\phi A_\epsilon = 1+2\phi+o(\phi^2)$ which is in accord with the exact value $\phi A_\epsilon = 1+2\phi$ already calculated above. Table 3 gives values of $\mathrm{ARE}(\tilde{\beta};\hat{\beta}) = (\phi A_\epsilon)^{-1}$ for the families (i)-(iii), at several values of ϕ. For the lognormal distribution, approximations based on the reciprocal of (2.29) are also given; for the normal and the inverse Gaussian the approximation is *exact*, even when

taken only to $O(\phi)$. The table shows that the approximations, both to $O(\phi^2)$ and to $O(\phi)$, are very close for the lognormal also.

TABLE 3. *Asymptotic efficiency of quasi-likelihood estimates under particular distributions with* $V(\mu)=\mu^2$

	(i) NORMAL	(ii) LOGNORMAL			(iii) INVERSE GAUSSIAN
		exact	*based on* (2.29) *to* $O(\phi^2)$	*based on* (2.29) *to* $O(\phi)$	
$\phi = 0.1$	0.833	0.953	0.953	0.952	0.952
$\phi = 0.2$	0.714	0.912	0.912	0.909	0.909
$\phi = 0.5$	0.500	0.811	0.814	0.800	0.800
$\phi = 1.0$	0.333	0.693	0.706	0.667	0.667
$\phi = 2.0$	0.200	0.549	0.600	0.500	0.500

It is clear that the loss of efficiency incurred in using quasi-likelihood estimation can be quite substantial, depending heavily on the true distribution as well as on ϕ, the squared coefficient of variation, whose value will depend on the context. Survival data often have a coefficient of variation of 1 or more. Most laboratory measurements, however, might be expected to have a smaller coefficient of variation, with a value of ϕ less than, say, 0.1, at which quasi-likelihood estimation would have high efficiency in all of examples (i)-(iii).

The form of (2.29) shows that, to first order in ϕ , the loss of efficiency depends only on $\tau_3^2=\{\kappa_3-\kappa_3(\Gamma)\}^2/(\phi^4\mu^6)$, where $\kappa_3(\Gamma)$ is the third cumulant of the gamma distribution. Thus a suitable first order index of the effect of departure from the gamma distribution is the difference in *skewness* between the true distribution and the gamma distribution. When the skewness of the true distribution is the same as that of the gamma, i.e. $\tau_3=0$, the leading term in (2.29) is $O(\phi^2)$ and depends only on $\tau_4^2=\{\kappa_4-\kappa_4(\Gamma)\}^2/(\phi^6\mu^8)$, so the difference in *kurtosis* is then the most important factor affecting efficiency, at least when ϕ is small. It is interesting to compare these results with the discussion at the end of §2.2.1, where corresponding observations were made concerning the structure of small departures from normality.

2.3.2 *Refinement*

Suppose now that some information is available about the value of τ_3. This is equivalent to knowledge about the relationship between the coefficient of variation and the standardized third moment, and could come from the sample $y_1,\ldots,y_n$ or perhaps from experience with similar sets of data; for discussion, see Cox & Oakes (1984, pp26-28).

In models with constant coefficient of variation, quasi-likelihood estimation uses the gamma density $f_G(y;\ \mu,\phi)$ in place of the true density $f(y;\ \mu,\phi)$. Proceeding by analogy with the refinement of least squares in §2.2.2 , if τ_3 is known consider using instead the improved approximation

$$f(y;\ \mu,\phi) \cong f_G(y;\ \mu,\phi)\{\ 1 + c_3 L_3^{(\alpha)}(z)\ \}$$

based on (2.28). To avoid problems with negative values of $\{1+c_3L_3^{(\alpha)}(z)\}$ we appeal, as before, to the Taylor series expansion

$$\log\{\ 1 + c_3 L_3^{(\alpha)}(z)\ \} = c_3 L_3^{(\alpha)}(z) - \tfrac{1}{2}c_3^2\{L_3^{(\alpha)}(z)\}^2 + \ldots$$

It is found, however, that terms in this expansion do not behave regularly as $\phi\to 0$; in particular an approximation based on only the first term, as in (2.13), is of no use. To remove the $O(\phi)$ term from the loss of efficiency would in fact require three terms, and would lead to an estimating equation involving powers of y_i higher than the third; in addition to being cumbersome, such an equation would not be unbiased without the introduction of further assumptions.

A more promising approach is suggested by the alternative motivation for the refined least squares estimator of §2.2.2 , as an optimal quadratic unbiased estimating equation. If, with ϕ known, we consider unbiased estimating equations of the form

$$\sum_{i=1}^{n} [\ \lambda(y_i-\mu_i)/\mu_i^2 + (1-\lambda)\{(y_i-\mu_i)^2 - \phi\mu_i^2\}/\mu_i^3\ ] \frac{x_{ir}}{g'(\mu_i)} = 0 \quad (r=0,\ldots,p\ ;\ -\infty<\lambda<\infty) \qquad (2.30)$$

then the asymptotic covariance matrix of a consistent solution, $\beta^*(\lambda)$ say, may be calculated as before and is of the form $B(\lambda)\mathrm{cov}(\tilde{\beta})$. The scalar $B(\lambda)$ is minimized at

$$\lambda^0 = (\kappa_4 - 2\phi\kappa_3 + 2\phi^2) / (\kappa_4 - 2\phi\kappa_3 - \kappa_3 + 4\phi^2) \qquad (2.31)$$

where here κ_r is the rth cumulant of $\epsilon_i = Y_i/\mu_i$; the minimum value is

$$B(\lambda^0) = 1 - \tfrac{1}{2}\phi\tau_3^2 + \tfrac{1}{4}\phi^2(\tau_3^2\tau_4 - 4\tau_3^3 + 2\tau_3^2) + O(\phi^3) \qquad (2.32)$$

if $\kappa_3 = O(\phi^2)$, $\kappa_4 = O(\phi^3)$, etc. An approximation to λ^0 based only on τ_3 is $\lambda_A^0 = 2/(2-\tau_3) = \lambda^0 + O(\phi)$, and in fact $B(\lambda_A^0) = B(\lambda^0) + O(\phi^3)$. Comparison of (2.32) with the reciprocal of (2.29) gives

$$\mathrm{ARE}(\beta^*(\lambda_A^0) ; \hat{\beta}) = 1 - \tfrac{1}{6}\phi^2\{\tau_4 - 3\tau_3(\tau_3 + 3)\}^2 + O(\phi^3)$$

so knowledge of τ_3 reduces the loss of efficiency from $O(\phi)$ to $O(\phi^2)$ and an index for the efficiency of $\beta^*(\lambda_A^0)$ is $\tau_4 - 3\tau_3(\tau_3+3)$: when this is zero the loss of efficiency is reduced still further to $O(\phi^3)$. In practice it is unlikely that τ_3 will be known exactly, but it may be shown, see Appendix 2, that an estimate $\tau_3 + O_p(n^{-\frac{1}{2}})$ allows the same asymptotic efficiency to be achieved. Whatever value of λ is used, (2.30) are unbiased estimating equations provided that the assumption of constant coefficient of variation is true; the value $\lambda=1$ gives the quasi-likelihood equations, and $\lambda=\frac{1}{2}$ gives maximum likelihood equations for the normal family, $N(\mu_i, \phi\mu_i^2)$.

Table 4 gives numerical values for the efficiency, under lognormal and inverse Gaussian distributions, of refinements based on (2.30). While, as expected, $\lambda=\lambda_A^0$ gives an improvement over quasi-likelihood estimation when ϕ is small, the approximation to λ^0 deteriorates rapidly as ϕ increases, and in fact for ϕ greater than about 0.18 the 'refinement' based on τ_3 is *less* efficient than quasi-likelihood estimation under these two distributions.

TABLE 4. *Asymptotic efficiency, under lognormal and inverse Gaussian distributions, of estimates based on* (2.30)

Lognormal

ϕ	λ^0	λ_A^0	Efficiency with $\lambda=1$	Efficiency with $\lambda=\lambda^0$	Efficiency with $\lambda=\lambda_A^0$
0.025	1.83	2.05	0.988	0.9996	0.9994
0.05	1.71	2.11	0.976	0.999	0.997
0.075	1.61	2.16	0.964	0.997	0.990
0.1	1.54	2.22	0.953	0.994	0.977
0.125	1.48	2.29	0.942	0.992	0.955
0.15	1.43	2.35	0.932	0.988	0.921
0.175	1.39	2.42	0.922	0.985	0.872
0.2	1.36	2.50	0.912	0.981	0.806

Inverse Gaussian

ϕ	λ^0	λ_A^0	Efficiency with $\lambda=1$	Efficiency with $\lambda=\lambda^0$	Efficiency with $\lambda=\lambda_A^0$
0.025	1.81	2.00	0.988	0.9991	0.9990
0.05	1.69	2.00	0.976	0.997	0.997
0.075	1.60	2.00	0.964	0.993	0.989
0.1	1.53	2.00	0.952	0.989	0.980
0.125	1.47	2.00	0.941	0.984	0.966
0.15	1.43	2.00	0.930	0.978	0.947
0.175	1.39	2.00	0.920	0.972	0.924
0.2	1.36	2.00	0.909	0.966	0.895

It is interesting to compare the efficiency of estimates based on $\lambda=\lambda^0$, the optimal choice of λ, with that of estimates based on $\lambda=1$, i.e. quasi-likelihood estimates. The behaviour for small ϕ has already been discussed above, via the approximation (2.32). The exact asymptotic relative efficiency is

$$\mathrm{ARE}(\tilde{\beta}; \beta^*(\lambda^0)) = B(\lambda^0) = 1 - \phi\tau_3^2/\{2 + \phi(2 + \tau_4 - 4\tau_3)\} ,$$

and in each of the examples (i)-(iii) this has a limiting value as $\phi\to\infty$. Under the normal distribution, $\beta^*(\lambda^0)$ is the maximum likelihood estimate and $B(\lambda^0)\to 0$ as $\phi\to\infty$; the limit values under the lognormal and the inverse Gaussian are 1 and $\frac{6}{7}$, respectively. Thus, while the improvement given by the optimum quadratic estimating equation has a certain uniformity

when ϕ is small, it varies, depending on the underlying distribution, between extremes when ϕ is large.

When ϕ is unknown its place in (2.30) may be taken by, for example, $n^{-1}\Sigma(y_i-\mu_i)^2/\mu_i^2$, and remarks corresponding to those made in §2.2.2 apply here also.

2.4 Overdispersion

2.4.1 *Asymptotic relative efficiency*

Consider now situations in which the quasi-likelihood model described by (1.1) and (1.2) arises from overdispersion, of a particular type, relative to the natural exponential family that has variance function $V(\mu)$. Because not all conceivable variance functions correspond to such a family, and for another reason that will become apparent, we restrict attention to the class of quadratic variance functions, and write $V(\mu)=a+b\mu+c\mu^2$. The natural exponential families defined by such variance functions have been discussed in detail by Morris (1982) under the name NEF-QVF, standing for Natural Exponential Family with Quadratic Variance Function; included are the normal, Poisson, gamma, binomial and negative binomial distributions.

More specifically, suppose each Y_i has conditional density, with respect to some measure, of the natural exponential family form

$$f_0(y_i|M_i) = \exp\{y_i\theta_i - u(\theta_i) + t(y_i)\} ,$$

with $$u'(\theta_i) = E(Y_i|M_i) = M_i, \text{ say,}$$

and $$u''(\theta_i) = \text{var}(Y_i|M_i) = V(M_i) ,$$

and suppose that the conditional means $\{M_i\}$ are themselves random variables, unobserved, drawn independently from a family of distributions with

$$E(M_i) = \mu_i \quad , \quad \text{var}(M_i) = V(\mu_i)(\phi-1)/(1+c) . \tag{2.33}$$

Then, unconditionally,

$$E(Y_i) = E(E(Y_i|M_i)) = \mu_i$$

and
$$\begin{aligned}\text{var}(Y_i) &= E(\text{var}(Y_i|M_i)) + \text{var}(E(Y_i|M_i)) \\ &= E(a + bM_i + cM_i^2) + (\phi-1)(a + b\mu_i + c\mu_i^2)/(1+c) \\ &= a+b\mu_i+c[\text{var}(M_i)+\{E(M_i)\}^2] + (\phi-1)(a+b\mu_i+c\mu_i^2)/(1+c) \\ &= \phi V(\mu_i) ,\end{aligned}$$

where ϕ is now restricted to be greater than 1. The restriction to variance functions that are quadratic is necessary here to ensure the desired form $\phi V(\mu_i)$ for the unconditional variance of Y_i .

The mixture model described above is a useful mechanism for overdispersion, allowing some 'random heterogeneity' in the mean parameter.

Again dropping the subscript i for convenience, our general mixture family is illustrated by some well known examples :

(i) $Y|M \sim$ Poisson(M), $M \sim$ gamma$(\mu,\nu\mu)$, leads to the negative binomial

$$\text{pr}(Y=y) = \Gamma(y+\nu\mu)\nu^{\nu\mu}/\{y!\Gamma(\nu\mu)(1+\nu)^{y+\nu\mu}\}$$

with $V(\mu)=\mu$ and $\phi=(\nu+1)/\nu$;

(ii) $Y|M \sim$ binomial$(r, M/r)$, $M/r \sim$ beta(p,q) , leads to the 'beta-binomial' with $E(Y)=\mu=rp/(p+q)$, $\text{var}(Y)=\mu\{1-(\mu/r)\}(p+q+r)/(p+q+1)$, so here $V(\mu)=\mu\{1-(\mu/r)\}$ and $\phi=(p+q+r)/(p+q+1)$;

(iii) the mixed exponential which was example (iv) in §2.3.1, with

$Y|M \sim$ exponential(mean M),

$M \sim$ inverse gamma [mean μ, coefficient of variation $\{(\phi-1)/2\}^{\frac{1}{2}}$]

gives the density (2.26); here $V(\mu)=\mu^2$.

Consider first, as an illustration, the case $V(\mu)=\mu^2$, in which the underlying natural exponential family is the *exponential* distribution, $f_0(y|M)=M^{-1}e^{-y/M}$. Write the unconditional log likelihood as

$$l(\beta,\phi) = \Sigma \log f(y_i;\mu_i,\phi) = \Sigma\, l_i(\mu_i,\phi) ,$$

say, where $f(y_i;\mu_i,\phi) = E(M_i^{-1}e^{-y_i/M_i})$, the expectation here being over the mixing distribution. The quasi-likelihood estimate in this case is the maximum likelihood estimate based on the exponential density $\mu_i^{-1}e^{-y_i/\mu_i}$ and the arguments of §2.3.1 identify $(\phi A_\epsilon)^{-1}$, where $A_\epsilon=\mu_i^2E(-\partial^2 l_i/\partial\mu_i^2)$, as a measure of asymptotic efficiency relative to maximum likelihood estimation based on the true likelihood. Again dropping the subscript i, write

$$f(y;\mu,\phi) = E\left\{\mu^{-1}e^{-y/\mu} + (M-\mu)\frac{d}{d\mu}(\mu^{-1}e^{-y/\mu}) + \tfrac{1}{2}(M-\mu)^2\frac{d^2}{d\mu^2}(\mu^{-1}e^{-y/\mu}) + \ldots\right\} \quad (2.34)$$

Now $d^r(\mu^{-1}e^{-y/\mu})/d\mu^r = \mu^{-(r+1)}e^{-y/\mu}(-1)^r r! L_r(y/\mu)$ where $L_r(.)$ is the Laguerre polynomial of degree r; thus (2.34) becomes

$$f(y;\mu,\phi) = \mu^{-1}e^{-y/\mu}\{1 + c_2L_2(y/\mu) + c_3L_3(y/\mu) + c_4L_4(y/\mu) + \ldots\} \quad (2.35)$$

with, from (2.33), $c_2 = \mathrm{var}(M)/\mu^2 = \tfrac{1}{2}(\phi-1)$, and in general $c_r = (-1)^r\{\tfrac{1}{2}(\phi-1)\}^{r/2}\rho_r$ where ρ_r is the standardized rth central moment of the mixing distribution.

Use of the expansion (2.35) to obtain an approximation to $(\phi A_\epsilon)^{-1}$ requires some assumptions about the form of the mixing distribution as its coefficient of variation tends to zero. Writing ψ for $(\phi-1)$, consider two types of behaviour :

(*a*) 'limiting normality', with cumulants $\kappa_3 = O(\psi^2)$, $\kappa_4 = O(\psi^3)$, etc.; and

(*b*) 'constant shape', with ρ_r (r=3,4,...) not depending on ϕ.

Calculation of the expected second derivative of the logarithm of (2.35) then gives

$$\phi A_\epsilon = \phi\mu^2 E(-\partial^2 l/\partial\mu^2) = 1 + \psi^2 - 6\psi c_3 - 8\psi^3 + o(\psi^3) \quad ; \quad (2.36)$$

see Appendix 4 for details. The term involving c_3 is $O(\psi^3)$ under assumption (*a*), and $O(\psi^{5/2})$ under (*b*).

The form of (2.36) may be checked against a particular mixing distribution such as the inverse gamma, example (iii) above, which has cumulant behaviour of type (*a*). Exact calculation from the density (2.26) gives

$$\phi A_\epsilon = \phi^2/(2\phi-1) \; . \quad (2.37)$$

The inverse gamma with coefficient of variation $\{\tfrac{1}{2}(\phi-1)\}^{\frac{1}{2}}$ has third cumulant $\kappa_3 = 2\mu^3(\phi-1)^2/(3-\phi)$, and substitution into (2.36) yields

$$\phi A_\epsilon = 1 + \psi^2 - 2\psi^3 + o(\psi^3) \; , \quad (2.38)$$

which is the same as the Taylor series expansion of (2.37). Table 5 gives values of $(\phi A_\epsilon)^{-1}$ based on (2.37) and on the approximation (2.38), and shows that quasi-likelihood estimation retains high efficiency under moderate overdispersion. The approximation based on (2.38) taken to $O(\psi^2)$ is within 10% of the exact value when ϕ is less than about 1.5 ; when taken to $O(\psi^3)$ the approximation is improved at values of ϕ close to 1 but becomes unreliable when ϕ is greater than about 1.4 .

TABLE 5. *Asymptotic efficiency of quasi-likelihood estimates under the exponential with inverse gamma mean,* (2.26)

	Exact, (2.37)	Approximation (2.38) to $O(\psi^2)$	Approximation (2.38) to $O(\psi^3)$	Efficiency of refinement based on (2.43)
$\phi = 1.1$	0.992	0.990	0.992	0.998
$\phi = 1.2$	0.972	0.962	0.977	0.942
$\phi = 1.3$	0.947	0.917	0.965	0.641
$\phi = 1.4$	0.918	0.862	0.969	0.197
$\phi = 1.5$	0.889	0.800	1.000	0.025
$\phi = 2.0$	0.750	0.500	∞	0

The main points to note about the form of the approximation (2.36) in general are that there is no $O(\psi)$ term and that, to $O(\psi^2)$, ϕA_ϵ depends only on ϕ and not on the shape of the mixing distribution. Thus Table 5 suggests that, for any mixing distribution with regular cumulant behaviour, quasi-likelihood estimation has efficiency greater than 90% if ϕ is not greater than about 1.3.

It would of course have been possible here, as in §2.3.1, to obtain the approximation (2.36) via an expansion of the form (2.28) about the *gamma* distribution with coefficient of variation $\sqrt{\phi}$, rather than about the exponential distribution as in (2.35). Expansion about the exponential distribution was used here because it may be generalized, as shown below, to mixtures of other NEF-QVF distributions. It would be surprising, though, if the first few terms in (2.36) did not have an interpretation, as in §2.3.1, in terms of differences between low order cumulants of the

gamma distribution and of the underlying mixed exponential. With this in mind, it is illuminating to consider the cumulant generating function, $\log E\{(1-itM)^{-1}\}$, of a mixture of exponential distributions, the expectation here being over the mixing distribution; a straightforward Taylor series expansion then gives the cumulants of the mixture, e.g.

$$\kappa_3(Y) = \mu^3\{2 + 6\psi + 6\rho_3(\tfrac{1}{2}\psi)^{3/2}\} .$$

If, as under assumption (*a*) or (*b*) above, the standardized central moments $\{\rho_r\}$ of the mixing distribution are $O(1)$ as $\psi\to 0$, this may be written as

$$\kappa_3(Y) = \mu^3\{2 + 6\psi + o(\psi)\} ,$$

which should be compared with the third cumulant of the gamma distribution,

$$\kappa_3(\Gamma) = 2\mu^3\phi^2 = \mu^3\{2 + 4\psi + o(\psi)\} .$$

As $\psi\to 0$, then, $\kappa_3(Y)-\kappa_3(\Gamma)$ has leading term $2\mu^3\psi$, which does not depend on the shape of the mixing distribution. Similarly the respective fourth cumulants are

$$\kappa_4(Y) = \mu^4\{6 + 36\psi + o(\psi)\}$$

and

$$\kappa_4(\Gamma) = 6\mu^4\phi^3 = \mu^4\{6 + 18\psi + o(\psi)\} ,$$

and again the leading term in the difference does not depend on the shape of the mixing distribution; the same pattern holds for differences between fifth, sixth and higher order cumulants. It is not at all surprising, then, that the leading term in (2.36) depends only on ϕ and not on the shape of the mixing distribution.

The above development is now generalized to mixtures of NEF-QVF distributions other than the exponential. In the general case write the density for a single observation as

$$f(y;\mu,\phi) = f_0(y\,|\,\mu) + \frac{(\phi-1)V(\mu)}{2(1+c)} f_0^{(2)}(y\,|\,\mu) + \left\{\frac{(\phi-1)V(\mu)}{(1+c)}\right\}^{3/2}\frac{\rho_3}{6} f_0^{(3)}(y\,|\,\mu) + \ldots \quad (2.39)$$

where $f_0^{(r)}$ denotes $d^r f_0/d\mu^r$ and ρ_r is the standardized rth moment of the mixing distribution as before. Now define

$$P_r(y,\ \mu)=V^r(\mu)f_0^{(r)}(y\,|\,\mu)/f_0(y\,|\,\mu) \qquad (r=0,1,...)\ ;$$

Morris (1982) has shown that $P_r(y,\mu)$ is a polynomial of degree r in both y and μ, and that $\{P_r : r=0,1,2,...\}$ is a set of *orthogonal* polynomials with respect to $f_0(y\,|\,\mu)$. Explicitly,

$$P_0(y,\mu)=1\ ,\ P_1(y,\mu)=y-\mu\ ,\ P_2(y,\mu)=(y-\mu)^2-V'(\mu)(y-\mu)-V(\mu)\ ,\ \text{etc.}\ ,$$

and the general class includes as special cases the Hermite and Laguerre polynomials used previously. The expansion (2.39) may be written in terms of these polynomials as

$$f(y;\ \mu,\phi) = f_0(y\,|\,\mu)[\ 1 + c_2P_2(y,\ \mu)/V(\mu) + c_3P_3(y,\ \mu)/\{V(\mu)\}^{3/2} + \ ...] \qquad (2.40)$$

where $c_r = \{\psi/(1+c)\}^{r/2}\rho_r/r!$ $(r=2,3,...)$. We assume, as before, regular behaviour of higher order cumulants of the mixing distribution, for example as in (*a*) or (*b*) above, as $\psi\to 0$; the standardized central moments $\{\rho_r\}$ are also assumed not to depend on μ . A straightforward calculation, some details of which are given in Appendix 5, then gives

$$\phi A_\epsilon = 1 + \tfrac{1}{2}\psi^2\{V'(\mu)\}^2/\{(1+c)V(\mu)\} + o(\psi^2)\ , \qquad (2.41)$$

where now $A_\epsilon=V(\mu)E(-\partial^2 l/\partial\mu^2)$; this generalizes (2.36) as far as the $O(\psi^2)$ term. Again there is no $O(\psi)$ term and the $O(\psi^2)$ term depends only on the variance function and not on the shape of the mixing distribution.

Notice that if $V(\mu)$=constant the leading term in (2.41) is $o(\psi^2)$ and will depend on the shape of the mixing distribution. This is no surprise; the 'parent' exponential family in this case is the normal, $Y_i\,|\,M_i \sim N(M_i,\sigma^2)$, and if we consider the particular mixing distribution $M_i \sim N(\mu_i,(\phi-1)\sigma^2)$ then, unconditionally, $Y_i \sim N(\mu_i,\phi\sigma^2)$. Thus the 'overdispersed' distribution is still in the natural exponential family and so quasi-likelihood estimation has full efficiency; it follows that (2.41) cannot have a term dependent only on the dispersion factor, ϕ.

If $V(\mu)$=constant or $V(\mu)\propto\mu^2$ the quantity $(\phi A_\epsilon)^{-1}$ has been shown in

§§2.2 and 2.3 to have a clear interpretation as the asymptotic efficiency, based on proportionality of asymptotic covariance matrices. More generally $\text{cov}(\hat{\beta})$ and $\text{cov}(\tilde{\beta})$ are proportional if $V(\mu_i)E(-\partial^2 l_i/\partial\mu_i^2)$ is a constant, not depending on i; this holds for the single sample model with μ_i=constant, but not in general. There may, however, be *approximate* proportionality in particular instances. For example if μ_i is approximately constant, as in a model with an intercept parameter that is large compared with other effects, changes in $V(\mu_i)E(-\partial^2 l_i/\partial\mu_i^2)$ may be small. Alternatively $V(\mu_i)E(-\partial^2 l_i/\partial\mu_i^2)$ may depend only weakly on μ_i ; for example the leading term in (2.41) depends on μ via $\{V'(\mu)\}^2/V(\mu)=(b+2c\mu)^2/(a+b\mu+c\mu^2)$, which has a constant limit as $\mu\to\infty$ if $c>0$, so $V(\mu_i)E(-\partial^2 l_i/\partial\mu_i^2)$ is approximately constant if values of μ_i are sufficiently large that $V(\mu_i)$ is dominated by the μ_i^2 term.

2.4.2 *Refinement*

In models of the type just discussed, quasi-likelihood estimation is maximum likelihood based on the 'parent' exponential family approximation $f_0(y\,|\,\mu)$. Given the additional information that the true distribution is a mixture $E\{f_0(y\,|\,M)\}$, with M having mean-variance relationship as in (2.33), a better approximation

$$f(y;\mu,\phi) \cong f_0(y\,|\,\mu)\{1+ c_2 P_2(y,\ \mu)/V(\mu)\}$$

is suggested by (2.40). To avoid problems with negative values, consider in fact using

$$f^*(y;\mu,\phi) = f_0(y\,|\,\mu)\exp\{c_2 P_2(y,\mu)/V(\mu)\}$$

to form a 'refined' quasi-likelihood. The resulting estimating equations are

$$\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\mu_i}\log f^*(y_i;\mu_i,\phi)\right\}\frac{x_{ir}}{g'(\mu_i)} = 0 \qquad (r=0,\ldots,p)$$

and it is easily shown, writing now just V for $V(\mu)$, that

$$\frac{\partial}{\partial\mu}\log f^*(y;\ \mu,\phi) = \frac{y-\mu}{V} + \left\{\frac{\psi}{2(1+c)V^2}\right\}\{V'V + (y-\mu)(V'^2-2V-VV'') - (y-\mu)^2V'\} \quad (2.42)$$

As an example, consider again $f_0(y\,|\,\mu)=\mu^{-1}e^{-y/\mu}$, the exponential 'parent'. This has $V(\mu)=\mu^2$ so (2.42) becomes

$$\begin{aligned}\frac{\partial}{\partial\mu}\log\ f^*(y;\ \mu,\phi) &= (y-\mu)/\mu^2 - \tfrac{1}{2}\psi\{(y-\mu)^2-\mu^2\}/\mu^3 \\ &= (y-\mu)/\mu^2 - \tfrac{1}{2}\psi\{(y-\mu)^2-\phi\mu^2\}/\mu^3 + O(\psi^2)\ . \qquad (2.43)\end{aligned}$$

If the $O(\psi^2)$ term is ignored, (2.43) gives a quadratic unbiased estimating equation of exactly the type discussed in §2.3.2 ; in terms of the form (2.30) it has $\lambda=2/(3-\phi)$, which is $\lambda^0+O(\psi^2)$ where λ^0 is the optimizing value (2.31). The estimating equation is 'close to optimal' in this sense, and may be shown to remove the $O(\psi^2)$ term from the loss of efficiency. In cases where the third or fourth cumulant of the mixture increases rapidly these advantages for 'small overdispersion' may be lost quite quickly as ϕ increases; for a numerical illustration of this see the last column of Table 5.

2.5 Discussion and remarks

2.5.1 *Strength of departure from the exponential family*

The numerical results in §§2.2.1, 2.3.1 and 2.4.1 suggest that quasi-likelihood estimation retains fairly high efficiency under 'moderate' departures from the corresponding natural exponential family. The notion of strength of departure has been made more explicit in the somewhat different asymptotic analysis of Cox (1983) which investigates mixture models like those discussed in §2.4 and concludes that quasi-likelihood estimation is likely to have high efficiency when overdispersion is 'on the borderline of detectibility'.

It might be expected that the efficiency of quasi-likelihood estimation would be related to Efron's (1975) measure of the 'statistical curvature' of a family of distributions, since both quantities reflect departure from the exponential family. The examples of §2.3 may be used to demonstrate that curvature, while certainly making a contribution, does not *determine* the asymptotic efficiency of quasi-likelihood estimation; see Table 6. The normal and inverse Gaussian here are curved exponential families. The lognormal family is a full exponential family and so has zero curvature; however quasi-likelihood estimation does not have full efficiency because the 'natural observation' is not y but $\log y$.

TABLE 6. *Asymptotic efficiency of quasi-likelihood estimates, and Efron's curvature, for some families with constant coefficient of variation*

Family	$\mathrm{ARE}(\tilde{\beta};\hat{\beta})$	*Curvature,* γ^2
Gamma	1	0
Normal	$\frac{1}{2}(\phi+\frac{1}{2})^{-1}$	$\frac{1}{4}\phi^2(\phi+\frac{1}{2})^{-3}$
Inverse Gaussian	$2(\phi+2)^{-1}$	$4\phi^2(\phi+2)^{-3}$
Lognormal	$\phi^{-1}\log(1+\phi)$	0

2.5.2. *A comparison arising from the connection between additive and multiplicative models*

There is a well known correspondence between multiplicative models for positive observations and additive models for their logarithms. For if

$$Y_i = \mu_i \epsilon_i \tag{2.44}$$

with $\log \mu_i = \beta_0 + \sum_{r=1}^{p} x_{ir}\beta_r$ and ϵ_i iid with $E(\epsilon_i)=1$, then

$$\log Y_i = \nu_i + \eta_i \tag{2.45}$$

with $\nu_i = \{\beta_0 + E(\log \epsilon_i)\} + \sum_{r=1}^{p} x_{ir}\beta_r$ and η_i iid with $E(\eta_i)=0$.

Thus, provided $\Sigma x_{ir}=0$ $(r=1,...,p)$, the non-intercept parameters $\beta_1,...,\beta_p$ are equivalently estimated from either (2.44) or (2.45). Now in §2.2.1 we investigated the efficiency of estimation in (2.45) based on the normal likelihood for $\{\eta_i\}$. In particular, Table 1 gives the efficiency when the $\{\eta_i\}$ are actually log gamma random variables. By the above correspondence, this is the same as the efficiency of lognormal-based estimates in (2.44) when the $\{\epsilon_i\}$ are actually gamma distributed. An immediate comparison may be made with the efficiency, calculated in §2.3.1, of gamma-based estimates when the $\{\epsilon_i\}$ are actually lognormal random variables: Table 7 gives some numerical values.

TABLE 7. *Reciprocal comparison of efficiencies in the multiplicative model* (2.44)

ϕ = variance of $\{\epsilon_i\}$	$\log(1+\phi)/\phi$ = asymptotic efficiency of gamma MLE under lognormal errors	$\phi/\{\psi'(1/\phi)\}$ = asymptotic efficiency of lognormal MLE under gamma errors
0.1	0.953	0.951
0.2	0.912	0.904
0.5	0.811	0.775
1.0	0.693	0.608
2.0	0.549	0.405

The differences in Table 7 are not large, particularly when ϕ is small, but there is an indication that quasi-likelihood estimation, i.e. maximum likelihood based on the gamma distribution, is 'safer' than maximum likelihood based on the lognormal in the sense that it is more efficient under reciprocal misspecification. An intuitive explanation for this could be that the lognormal estimate involves logarithms of observations whose propensity to be small increases, as the variance increases, more rapidly under the gamma distribution than under the lognormal.

By the correspondence described above, the same table may also be used to compare estimation based on the normal and log gamma likelihoods in the additive model (2.45).

2.5.3 *Behaviour of quasi-likelihood estimates under nonstandard conditions*

It has been implicitly assumed throughout this chapter that conditions necessary for the results of McCullagh (1983), concerning consistency and asymptotic normality of quasi-likelihood estimates, are satisfied. In particular, it has been assumed that the mean and variance specifications (1.1) and (1.2), on which the quasi-likelihood equations (1.7) are based, are correct. While the regression specified at (1.1) is taken to define the parameters of interest, and in that sense is not in question, the mean-variance relationship (1.2) is often a secondary aspect of the model, based perhaps on empirical experience and assumed for pragmatic reasons such as increased precision of estimation. As already noted in §1.2.1, quasi-likelihood estimates remain, subject to some conditions, consistent and asymptotically normal even when (1.2) fails. The asymptotic covariance matrix is given by (1.13). In general the efficiency of quasi-likelihood estimates based on an incorrect variance function depends heavily on the regression model as well as on the shape of the underlying distribution; in

the 'single sample' model with $\mu_i=\mu$, the quasi-likelihood estimate based on *any* variance function is just $\bar{y}$, the sample mean, but in more complex models estimates based on different variance functions may differ considerably. The structure of (1.13) in some particular cases has been investigated by Gourieroux, Monfort & Trognon (1984a,b).

Crowder (1986a) has constructed examples of models which violate the usual conditions to the extent that quasi-likelihood estimates have asymptotic efficiency zero or are even inconsistent.

2.5.4 *Problems with 'refined' quasi-likelihood estimators in practice*

The quadratic estimating equations of §§2.2.2, 2.3.2 and 2.4.2 were motivated primarily by the need to assess the extent to which, under only the mean and variance assumptions (1.1) and (1.2), it is possible to improve upon the (asymptotic) performance of linear estimating equations. Though they serve this theoretical purpose well, their practical utility is less clear. Four aspects that will be discussed briefly here are the existence of solutions, efficiency in finite samples, standard errors and robustness.

The *asymptotic* existence of a consistent solution to unbiased estimating equations like those discussed in this chapter is not in question: general results such as those of Crowder (1986b) apply directly. In finite samples, however, the existence of a solution is not always guaranteed even when the estimating equations are linear in the observations. Wedderburn (1976) discusses the existence of solutions to quasi-likelihood equations based on four particular variance functions. Preliminary considerations suggest that a more general treatment would be very complicated; here we merely use a simple example to illustrate that a quadratic estimating equation may fail where a linear one does not. Consider, then, the general quadratic estimating equation (2.17) in the 'single sample' model with constant, known variance ϕ. The value $\lambda=1$ gives the least squares

equation, linear in the observations, which has a unique real solution in every possible sample. However any other value of λ gives a quadratic equation that has a positive probability of no real root; the discriminant is

$$n^2[\,\lambda^2-4(1-\lambda)^2\{\overline{y^2}-\bar{y}^2-\phi\}\,]\ ,$$

which, although positive with increasing probability as n increases if $\lambda\neq 0$, may be negative in any given sample.

Throughout this chapter, 'efficiency' has meant 'asymptotic efficiency' relative to the maximum likelihood estimate. Efficiency in finite samples has not been explicitly considered. A source of concern in this respect must be the use of sample-based estimates for third or fourth moments in the 'refined' estimating equations of §§2.2.2 and 2.3.2. While the arguments of Appendix 2 show that substitution of a $\sqrt{n}$-consistent estimate for the true value achieves the same *asymptotic* efficiency, poor performance is to be expected in all but very large samples. A systematic study of this has not been attempted. Here we give a numerical illustration based on a simple example; in the single sample model with constant, known variance $\phi=1$, a small modification of (2.18) gives, replacing γ_1 by the sample k-statistic k_3, the estimate

$$\mu^{**}= k_1-\tfrac{1}{2}k_3(k_2-1)\quad .$$

Consider behaviour under the normal distribution, $N(\mu,1)$; μ^{**} is unbiased and may be shown to have variance

$$\mathrm{var}(\mu^{**}) \;=\; n^{-1}\left\{1+\frac{3n^2}{(n-1)^2(n-2)}+\frac{72n^2}{(n-1)^3(n-2)}\right\}\ .$$

The estimate with γ_1 known to be zero would be simply $\bar{y}$, the sample mean, with variance n^{-1}. Table 8 gives the relative efficiency at some values of n.

TABLE 8. *Finite-sample efficiency of* μ^{**} *under the normal distribution*

n	Efficiency=$\{n\text{var}(\mu^{**})\}^{-1}$	'Equivalent' fixed value of γ_1
3	0.011	± 10.82
5	0.084	± 4.18
10	0.37	± 1.75
20	0.70	± 0.89
100	0.96	± 0.28

The asymptotic efficiency is 1, since k_3 is $\sqrt{n}$-consistent for γ_1. In small and even moderate-sized samples, estimation of γ_1 inflates the variance considerably. The third column of Table 8 gives the value of γ_1 which, if taken as the known value and used in place of k_3, would give an estimator with the same variance as μ^{**}. Thus, for example, use of any fixed value of γ_1 in the range [-0.89,0.89] is preferable, under normality, to the sample-based estimate k_3 in samples of size less than 20.

Remarks made in §1.2.2 about standard errors and robustness of Gaussian estimates apply here more generally. Standard errors for solutions of quadratic estimating equations will involve third and fourth moments, and may be unreliable if these have to be estimated. It has already been noted in §1.2.2 that, under the mean and variance assumptions (1.1) and (1.2), the quasi-likelihood equations (1.7) are asymptotically optimal among estimating equations that are linear in the observations $y_1,\ldots,y_n$. Thus any refinement must be non-linear; to remain unbiased under only (1.1) and (1.2) it must in fact, like the refinements suggested in §§2.2.2, 2.3.2 and 2.4.2, be quadratic. Such quadratic estimating equations require, for consistency, that the assumed variance function is correct; the quasi-likelihood estimates are consistent more generally, and in this sense refinement involves a certain sacrifice of robustness.

2.5.5 *A link with robust estimation*

The cubic estimating equations (2.20) of §2.2.2, designed to allow for kurtosis in the underlying distribution when skewness is known to be zero, bear some similarity to certain '*M*-estimators' familiar in the literature on robustness. To simplify the discussion, consider the 'single sample' model with $E(Y_i)=\mu$ and $\text{var}(Y_i)=\phi$ for all i ; then (2.20) may be written as

$$\sum_{i=1}^{n} z_i w(z_i) = 0 \quad , \tag{2.46}$$

where $$z_i = (y_i-\mu)/\sqrt{\phi}$$

and $$w(z) = (1+\tfrac{1}{2}\gamma_2) - \tfrac{1}{6}\gamma_2 z^2 \quad .$$

The equation is of a general type (Green, 1984) which may be solved for μ by the method of iteratively reweighted least squares, with $\{w(z_i)\}$ as iterative 'weights'.

Consider the estimating equation (2.46) with $\gamma_2>0$; this 'refinement' has been shown in §2.2.2 to provide an improvement in efficiency, over ordinary least squares, when the underlying distribution has zero skewness and small positive kurtosis γ_2 . This local, asymptotic property does not, however, imply robustness in the commonly used sense of resistance to gross errors; not only is $w(z)$ unbounded, but it actually becomes *negative* for $z^2 > 3+6\gamma_2^{-1}$, which is clearly absurd. A possible remedy is to replace $w(z)$ by

$$w^*(z) = \max\{w(z),0\} \ ;$$

this is not only bounded, but is zero outside a finite range so that grossly discrepant observations are rejected completely. Weight functions like $w^*(z)$ are said to define *redescending M-estimators*; many others have been suggested in the context of robust estimation, perhaps the most familiar

being the *biweight* family, suggested by J.W.Tukey,

$$w_b(z) = \begin{cases} (r^2-z^2)^2 & \text{if } z^2 \leq r^2 \\ 0 & \text{otherwise} \end{cases} \qquad (2.47)$$

indexed by the scalar r. A weight function is defined only up to a constant of proportionality, and we may write, mimicking (2.47),

$$w^*(z) = \begin{cases} r^2-z^2 & \text{if } z^2 \leq r^2 \\ 0 & \text{otherwise} \end{cases} \qquad (2.48)$$

where $r^2 = 3 + 6\gamma_2^{-1}$; thus every $w^*(z)$ may be derived as the square root of a biweight function. The form (2.48) has also been discussed by Stigler (1980), who shows that an early robust estimator due to Smith (1888) is in fact a redescending M-estimator with precisely such a weight function.

The functions w^* and w_b are qualitatively very similar, both to one another and to many others suggested in the literature on robustness. Andrews *et al* (1972) compare the performances of estimators based on several such weight functions, including w_b but not w^*; it would be interesting to study w^* in the same way, via asymptotic calculation and extensive simulation, but this is outside the scope of the present work.

One possibly appealing feature of weight functions of the type (2.48) is the interpretation of r in terms of γ_2 . Each such weight function might thus be thought of as being aimed at some 'target kurtosis', and choice of a particular r could be based either on prior knowledge of γ_2 or on a data-based estimate such as the sample kurtosis. The latter, 'kurtosis-adaptive' approach would be similar in spirit to some estimators proposed for the Princeton Robustness Study by R.V.Hogg, in which a choice among four estimators with different but well-understood characteristics is made on the basis of sample kurtosis. However it is found by Andrews *et al* (1972) that even the best of Hogg's

kurtosis-adaptive estimators is largely out-performed by non-adaptive methods; it seems likely that *M*-estimates based on an adaptively chosen weight function w^* would actually be somewhat *less* stable in finite samples than those of Hogg, so this adaptive approach appears to offer little promise as a robust method.

CHAPTER 3

Standard errors for quasi-likelihood estimates

3.1 Introduction : two types of estimated standard error

3.1.1 *'Model-based' standard errors*

Estimation of the covariance matrix of $\tilde{\beta}$, the vector of solutions of (1.7), is usually based on a large-sample $(p+1)$-variate normal approximation to its distribution. Under the mean and variance assumptions (1.1) and (1.2) the asymptotic covariance matrix is given by (1.10); the standard procedure is to replace μ_i by $\tilde{\mu}_i$, where $\tilde{\mu}_i=\mu_i(\tilde{\beta})$, and to estimate ϕ, if it is unknown, by

$$\tilde{\phi} = (n-q)^{-1}\Sigma\{(y_i-\tilde{\mu}_i)^2/V(\tilde{\mu}_i)\} = C/(n-q) \quad \text{, say,} \tag{3.1}$$

where $q=p+1$. Other estimates of ϕ are possible: see McCullagh & Nelder (1983, pp172-73). In this chapter, as in the previous one, everything will be expressed in terms of the generalized linear form (1.3) for $\mu_i(\beta)$; discussion in terms of a more general nonlinear regression, while straightforward, seems rather less revealing. For quasi-likelihood estimates in a generalized linear model, then, the 'model-based' covariance matrix estimate is derived from (1.11) as

$$\mathrm{cov}_M(\tilde{\beta}) = \tilde{\phi}\{X^T\mathrm{diag}(\tilde{w}_i)X\}^{-1} = \tilde{\phi}L^{-1} \ , \tag{3.2}$$

say, where $\tilde{w}_i=w_i(\tilde{\mu}_i)$ with $w_i(\mu_i)$ as in (1.12). Estimated standard errors are calculated from the diagonal elements of this matrix by taking square roots.

3.1.2 *'Robust' standard errors*

A 'model robust' alternative to the covariance estimate (3.2) is

$$\mathrm{cov}_R(\tilde{\beta}) = L^{-1}\{X^T\mathrm{diag}(c_i^2\,\tilde{w}_i)X\}L^{-1} \quad , \tag{3.3}$$

where $c_i^2=(y_i-\tilde{\mu}_i)^2/V(\tilde{\mu}_i)$ is the contribution of the ith observation to C. Motivating arguments for this estimate, and references to previous work in which the same or similar estimates have been suggested, are given by Royall (1986). Essentially, $\mathrm{cov}_R(\tilde{\beta})$ is designed to provide a consistent estimate, when the $\{Y_i\}$ have general variances $\{\phi_i\}$, of the matrix (1.13), which in the case of a generalized linear model is

$$\{X^T\mathrm{diag}(w_i)X\}^{-1}[X^T\mathrm{diag}\{w_i\phi_i/V(\mu_i)\}X]\{X^T\mathrm{diag}(w_i)X\}^{-1} \quad . \tag{3.4}$$

Thus $\mathrm{cov}_R(\tilde{\beta})$ is consistent for $\mathrm{cov}(\tilde{\beta})$, or more correctly $n\mathrm{cov}_R(\tilde{\beta})$ is consistent for $n\mathrm{cov}(\tilde{\beta})$, under failure of the variance assumption (1.2); this 'robustness' property is not shared by the model-based estimate (3.2).

Extensive study has been made, particularly in the econometrics literature, of the case $\{V(\mu)=1,\ g(\mu)=\mu\}$, i.e. linear models with error variance assumed, tacitly at least, to be constant. Here $\tilde{\beta}$ is the ordinary least squares estimate, and the usual 'model-based' covariance estimate is

$$\mathrm{cov}_M(\tilde{\beta}) = \tilde{\phi}(X^TX)^{-1} \quad . \tag{3.5}$$

The 'robust' alternative is

$$\mathrm{cov}_R(\tilde{\beta}) = (X^TX)^{-1}\{X^T\mathrm{diag}(y_i-\tilde{\mu}_i)^2X\}(X^TX)^{-1} \quad , \tag{3.6}$$

which provides protection against heteroscedasticity of the errors : see Eicker (1963) or White (1980) for details, including regularity conditions.

This chapter explores certain aspects of the 'robust' covariance estimate (3.3) and some variants. First, in §3.2, the calculation and interpretation of robust standard errors are illustrated using a log-linear

model for some data on ship damage; for a further illustration, based on some bioassay data, see Pregibon (1983). Section 3.3 discusses briefly the conflicting considerations of robustness and efficiency, and in §3.4 a 'compromise' is derived by arguments that are 'partially Bayesian' in the sense of Cox (1975). Finally, in §3.5, some remarks are made about finite sample bias, which has been a topic of very recent interest in econometrics.

3.2 Illustrative application : ship damage data

The data in Table 9 on damage to ships by waves were presented by McCullagh & Nelder (1983, p137). Ships are classified by type (A-E), year of construction (1960-64, 1965-69, 1970-74, 1975-79) and period of operation (1960-74, 1975-79). We are given the aggregate months in service, $m_i=m_{rst}$, and the number of damage incidents, $y_i=y_{rst}$, where r represents type, s represents year and t represents period of operation. Of the 40(=5×4×2) conceivable categories, six had no ships and so give no information; five of these are in fact logically impossible. Note that a single ship may have been damaged more than once and some ships will have operated both before and after 1974.

Simple arguments motivate McCullagh & Nelder towards a log-linear model of the form

$$\log \mu_{rst} = \beta_0 + \log m_{rst} + \gamma_r + \delta_s + \epsilon_t \qquad (r=1,...,5;\ s=1,...,4;\ t=1,2) \ , \quad (3.7)$$

i.e. a 'main effects' model including a term to make the expected number of damage incidents in a particular category proportional to the aggregate months in service in that category. The parameters of interest are $\beta=(\beta_0,\gamma_2,...,\gamma_5,\delta_2,...,\delta_4,\epsilon_2)$; note that γ_1, δ_1 and ϵ_1 are set equal to zero to avoid redundancy in the parameterization. Clearly (3.7) is a generalized linear model; in terms of the general notation (1.3) it has $g(\mu)=\log \mu$ and a model matrix X consisting of 'dummy variables', viz.

$$x_{i0} = 1 \ ,$$

$$x_{i1} = \begin{cases} 1 & \text{ship type B} \\ 0 & \text{otherwise} \ , \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{ship type C} \\ 0 & \text{otherwise} \ , \end{cases}$$

etc., up to

$$x_{i9} = \begin{cases} 1 & \text{period of operation 1975-79} \\ 0 & \text{otherwise} \ . \end{cases}$$

TABLE 9. *Number of reported damage incidents and aggregate months of service by ship type, year of construction and period of operation.*

i	*Ship type*	*Year of construction*	*Period of operation*	*Aggregate months service*	*Number of damage incidents*
1	A	1960-64	1960-74	127	0
2	A	1960-64	1975-79	63	0
3	A	1965-69	1960-74	1095	3
4	A	1965-69	1975-79	1095	4
5	A	1970-74	1960-74	1512	6
6	A	1970-74	1975-79	3353	18
7	A	1975-79	1960-74	0	0*
8	A	1975-79	1975-79	2244	11
9	B	1960-64	1960-74	44882	39
10	B	1960-64	1975-79	17176	29
11	B	1965-69	1960-74	28609	58
12	B	1965-69	1975-79	20370	53
13	B	1970-74	1960-74	7064	12
14	B	1970-74	1975-79	13099	44
15	B	1975-79	1960-74	0	0*
16	B	1975-79	1975-79	7117	18
17	C	1960-64	1960-74	1179	1
18	C	1960-64	1975-79	552	1
19	C	1965-69	1960-74	781	0
20	C	1965-69	1975-79	676	1
21	C	1970-74	1960-74	783	6
22	C	1970-74	1975-79	1948	2
23	C	1975-79	1960-74	0	0*
24	C	1975-79	1975-79	274	1
25	D	1960-64	1960-74	251	0
26	D	1960-64	1975-79	105	0
27	D	1965-69	1960-74	288	0
28	D	1965-69	1975-79	192	0
29	D	1970-74	1960-74	349	2
30	D	1970-74	1975-79	1208	11
31	D	1975-79	1960-74	0	0*
32	D	1975-79	1975-79	2051	4
33	E	1960-64	1960-74	45	0
34	E	1960-64	1975-79	0	0*
35	E	1965-69	1960-74	789	7
36	E	1965-69	1975-79	437	7
37	E	1970-74	1960-74	1157	5
38	E	1970-74	1975-79	2161	12
39	E	1975-79	1960-74	0	0*
40	E	1975-79	1975-79	542	1

* Necessarily empty cells.

The known adjustment term, m_i, is easily incorporated as an 'offset'.

In their analysis, McCullagh & Nelder estimate the parameters of (3.7) using quasi-likelihood equations based on the variance function $V(\mu)=\mu$. The motivation for this is clear: the quasi-likelihood equations are precisely the same as maximum likelihood equations derived from the 'usual' assumption, for 'counts' data of this type, of an underlying Poisson distribution. The estimates, $\tilde{\beta}$, are given in column (i) of Table 10.

TABLE 10. *Ship damage: parameter estimates and estimated standard errors for the main effects model*

Parameter	*Estimates*		*Standard errors*			
	(i) *Full data*	(ii) *Obs. no. 21 deleted*	(iii) *Poisson-based* (ϕ=1)	(iv) *'Model-based'* ($\tilde{\phi}$=1.69)	(v) *'Robust'*, (3.12)	(vi) *'Bias-corrected robust'*
β_0	-6.41	-6.41	0.22	0.28	0.12	0.16
γ_2	-0.54	-0.55	0.18	0.23	0.09	0.12
γ_3	-0.69	-1.26	0.33	0.43	0.49	0.67
γ_4	-0.08	-0.08	0.29	0.38	0.37	0.51
γ_5	0.33	0.33	0.24	0.31	0.24	0.32
δ_2	0.70	0.76	0.15	0.19	0.11	0.15
δ_3	0.82	0.76	0.17	0.22	0.14	0.19
δ_4	0.45	0.41	0.23	0.30	0.20	0.27
ϵ_2	0.38	0.44	0.12	0.15	0.10	0.14

The Pearson chi-square statistic for the fitted model is C=42.2 on (34–9)=25 degrees of freedom. This is large enough to cast doubt on the standard Poisson assumption, and suggests in particular that Poisson-based standard errors might be too small. McCullagh & Nelder base their standard errors instead on the 'overdispersed Poisson' assumption

$$\text{var}(Y_i) = \phi\mu_i \qquad (i=1,...,n) \ , \qquad (3.8)$$

with ϕ estimated by $\tilde{\phi}=(42.2/25)=1.69$; this leads to the 'model-based' covariance matrix (3.2), which in the present case is

$$\text{cov}_M(\tilde{\beta}) = \tilde{\phi}\{X^T\text{diag}(\tilde{\mu}_i)X\}^{-1} \ . \qquad (3.9)$$

The corresponding standard errors are given in column (iv) of Table 10; they are simply $\sqrt{(1.69)}=1.30$ times the Poisson-based standard errors, which are given in column (iii).

The covariance estimate (3.9) is not generally consistent under failure of the mean-variance relationship (3.8). With this in mind, it is interesting to study the squared 'standardized Pearson residuals' (McCullagh & Nelder, 1983, p211), defined by

$$r_i^2 = (y_i-\tilde{\mu}_i)^2/\{\tilde{\mu}_i(1-h_{ii})\} \qquad (i=1,...,n) \ , \qquad (3.10)$$

where h_{ii} is the ith diagonal element of the approximate 'hat' matrix

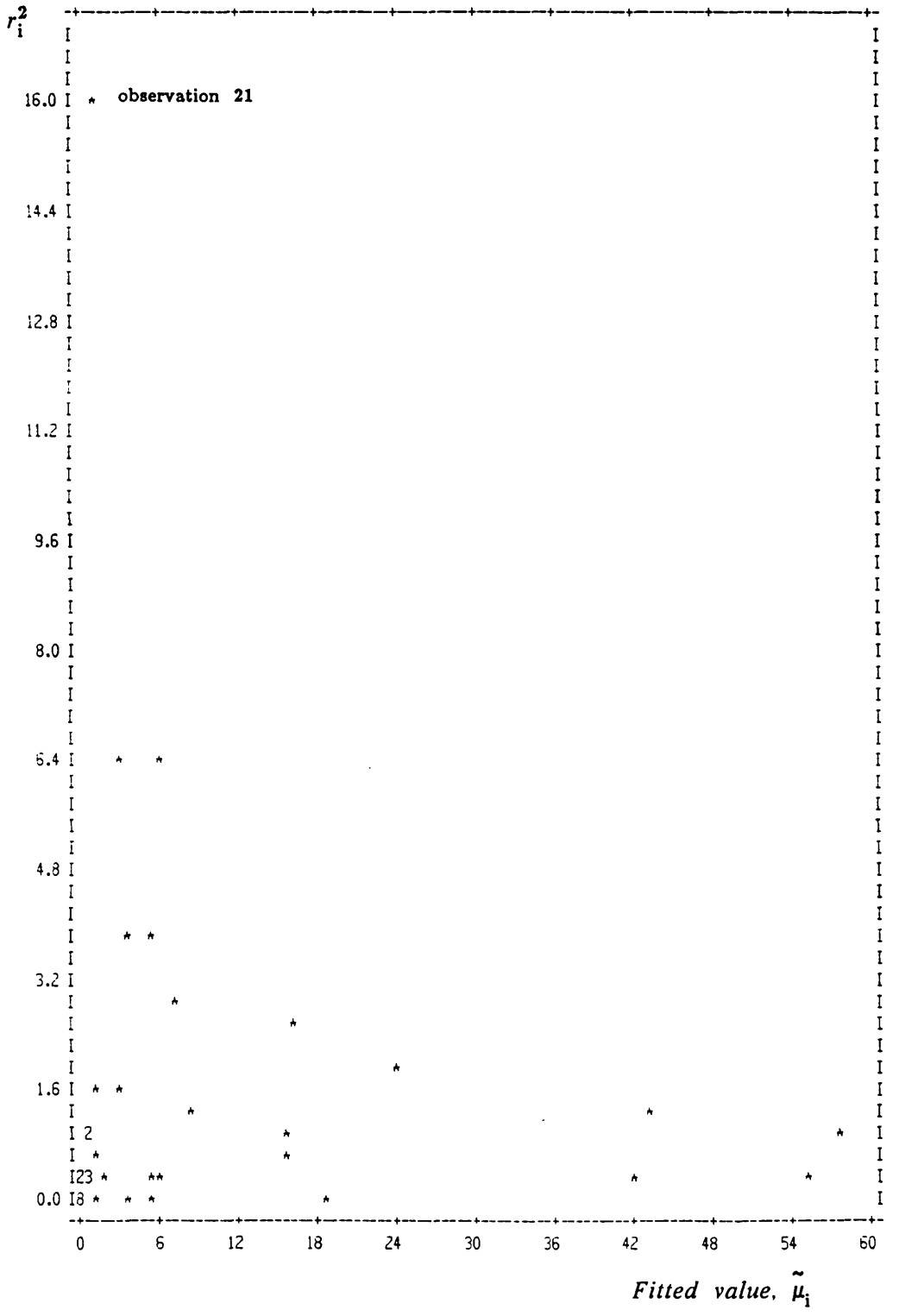
$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2} \ , \qquad (3.11)$$

with $W = \text{diag}(\tilde{w}_i) = \text{diag}(\tilde{\mu}_i)$ in this case. Under (3.8), $E(r_i^2)$ is approximately ϕ, for all i . The $\{r_i^2\}$ are given in Table 11, and plotted in Figure 1 against the fitted values $\{\tilde{\mu}_i\}$. Observation no. 21 stands out, with observed value 6, fitted value 1.47 and squared standardized Pearson residual 16.1 . The plot also suggests a possible relationship between μ_i and $E(r_i^2)$, with large values of r_i^2 being more likely when μ_i is small. While this is far from conclusive evidence against (3.8), it nevertheless casts some doubt on the validity of standard errors based on (3.9).

The proposed robust alternative to (3.9) is, directly from (3.3),

$$\text{cov}_R(\tilde{\beta}) = \{X^T\text{diag}(\tilde{\mu}_i)X\}^{-1}\{X^T\text{diag}(y_i-\tilde{\mu}_i)^2X\}\{X^T\text{diag}(\tilde{\mu}_i)X\}^{-1} \ . \qquad (3.12)$$

FIGURE 1. *Ship damage data, main effects model: squared standardized Pearson residuals versus fitted values*



Standard errors based on this covariance matrix estimate are given in column (v) of Table 10.

TABLE 11. *Ship damage: squared standardized Pearson residuals r_i^2 from the fitted main effects model.*

Period of operation		1960–74				1975–79			
Year of construction		60–64	65–69	70–74	75–79	60–64	65–69	70–74	75–79
Ship type	A	0.21	0.13	0.03	–	0.15	0.42	0.02	1.15
	B	1.26	0.42	1.08	–	1.85	1.06	0.27	0.70
	C	0.00	1.49	16.1	–	0.17	0.31	4.00	0.45
	D	0.40	0.95	0.57	–	0.24	0.93	6.54	3.00
	E	0.10	3.76	0.22	–	–	6.43	2.72	1.47

On comparing columns (iv) and (v) of Table 10 the most prominent feature is that, with only one exception, the robust standard errors are smaller than those from (3.9). There is in fact considerable bias present in the robust covariance estimate (3.12), due to the fact that $(y_i-\tilde{\mu}_i)^2$ is not an unbiased estimate of $\mathrm{var}(Y_i)$. A simple correction, implemented in column (vi) of Table 10, is to multiply $\mathrm{cov}_R(\tilde{\beta})$ by $n/(n-q)$; while this does not completely remove the bias, it has the appeal of making $\mathrm{cov}_R(\tilde{\beta})$ agree exactly with $\mathrm{cov}_M(\tilde{\beta})$ in the single sample problem, in which μ_i is the same for all i and so the form of $V(\mu_i)$ is irrelevant. More generally, the bias correction $n/(n-q)$ should be reasonably effective provided that the $\{h_{ii}\}$ are all close to q/n. When the $\{h_{ii}\}$ vary greatly, unbiasedness may be more nearly achieved by replacing each $(y_i-\tilde{\mu}_i)^2$ in (3.12) with $(y_i-\tilde{\mu}_i)^2/(1-h_{ii})$, although this may have an adverse effect on the stability of $\mathrm{cov}_R(\tilde{\beta})$; this point will be discussed further in §3.5 .

The differences between columns (iv) and (vi) of Table 10 are concentrated mainly in the standard errors for the estimated 'ship type' effects $\tilde{\gamma}_2,\ldots,\tilde{\gamma}_5$; recall that these parameters measure the *differences* in accident-proneness between ship type A and types B, C, D and E. The robust standard errors for $\tilde{\gamma}_3$ and $\tilde{\gamma}_4$ are much larger than their

model-based counterparts, while those for $\tilde{\gamma}_2$ and $\tilde{\beta}_0$ are smaller, even, than the corresponding Poisson-based standard errors. The standard errors for $\tilde{\delta}_2$, $\tilde{\delta}_3$, $\tilde{\delta}_4$ and $\tilde{\epsilon}_2$ do not change much, but are all slightly smaller in column (vi) than in column (iv).

These differences reflect the complex nature of the observed pattern of dispersion, relative to the simple 'constant overdispersion' model (3.8). Evidence of this pattern is to be found in Table 11. Of the large values of r_i^2, say those greater than 3, none is to be found in a position corresponding to ship type A or B, or corresponding to years of construction 1960-1964. The relevant parameter estimates are $\tilde{\beta}_0$ and $\tilde{\gamma}_2$, standard errors for both of which are much smaller in column (vi) than under the 'constant overdispersion' assumption of column (iv). However, unless there is some reason to expect sub-Poisson variation here, the fact that the robust standard errors are also smaller than those in column (iii) is suspicious.

Ship types C and D have the two largest values of r_i^2, and standard errors for the corresponding estimates $\tilde{\gamma}_3$ and $\tilde{\gamma}_4$ are increased accordingly; Table 11 itself does not, however, explain why these increases are so large, and in particular why the standard errors for $\tilde{\gamma}_3$ and $\tilde{\gamma}_4$ are increased while those for $\tilde{\gamma}_5$, $\tilde{\delta}_2$, $\tilde{\delta}_4$, and especially for $\tilde{\delta}_3$, change so little between column (iv) and column (vi) of Table 10.

Clearly observation 21 is very important. Its contribution to C is 13.9, or 33% of the total. To explore its influence a little further, consider re-fitting the same 'main effects' model but with observation 21 deleted completely from the data; the chi-square statistic for the re-fitted model is C=27.5 on 24 degrees on freedom, and parameter estimates are given in column (ii) of Table 10. Comparison of columns (i) and (ii) reveals that the influence of observation 21 is concentrated almost entirely on the determination of $\tilde{\gamma}_3$, the estimated 'ship type C' effect. Note that

it would be surprising to find any major differences between columns (i) and (ii) that were not also represented in the comparison between columns (iv) and (vi) : roughly speaking, differences between columns (i) and (ii) represent uncertainty due to doubt about the applicability of the model to observation 21; if the regression (3.7) is assumed to hold for all observations this means doubt about the applicability of the variance assumption (3.8) to observation 21. The robust covariance matrix estimate (3.12) is designed to allow for failure of (3.8) more generally.

The interpretation of the squared standardized Pearson residual, r_i^2, as a measure of the contribution of observation i to the discrepancy between model-based and robust standard errors may be made more explicit by considering

$$\text{trace}\,[\text{cov}_R(\tilde{\beta})\{\text{cov}_M(\tilde{\beta})\}^{-1}] \ .$$

In terms of the general definitions (3.2) and (3.3) this is

$$(\tilde{\phi})^{-1}\text{trace}\,[\,L^{-1}X^T\text{diag}(c_i^2\tilde{w}_i)X\,] \ ,$$

which may be re-expressed as

$$(\tilde{\phi})^{-1}\sum_{i=1}^{n} c_i^2\tilde{w}_i^{1/2}x_i(X^TWX)^{-1}x_i^T\tilde{w}_i^{1/2} = (\tilde{\phi})^{-1}\sum_{i=1}^{n} c_i^2h_{ii} \ ;$$

here x_i is the ith row of X. Thus the ith observation contributes an amount

$$(\tilde{\phi})^{-1}c_i^2h_{ii} = (\tilde{\phi})^{-1}r_i^2h_{ii}(1-h_{ii})$$

to $\text{trace}\,[\text{cov}_R(\tilde{\beta})\{\text{cov}_M(\tilde{\beta})\}^{-1}]$.

A final remark concerns the computation of $\text{cov}_R(\tilde{\beta})$. While computer packages such as *GLIM* produce $\text{cov}_M(\tilde{\beta})$ as a 'by-product' of the algorithm used to solve the quasi-likelihood equations, a little more calculation is needed for $\text{cov}_R(\tilde{\beta})$. The required matrix manipulations are not possible in *GLIM*. All of the calculations for this section were easily

carried out using the *GENSTAT* package, which incorporates most of the facilities of *GLIM* as well as general matrix handling.

3.3 Robustness versus efficiency

The simple example that will be discussed here will demonstrate the potential loss of efficiency involved in using $\text{cov}_R(\tilde{\beta})$ when the assumed variance function is in fact correct, and will allow the trade-off between robustness and efficiency to be illustrated numerically.

Consider a single sample $Y_1,\ldots,Y_n$, independent and identically distributed, with parameter of interest $\mu=E(Y_i)$ and assumed mean-variance relationship

$$\text{var}(Y_i) = \mu \ . \tag{3.13}$$

The quasi-likelihood estimate corresponding to variance function $V(\mu)=\mu$ is $\tilde{\mu}=n^{-1}\Sigma y_i$, the sample mean. The model-based and robust variance estimates are, from (3.2) and (3.3),

$$\text{cov}_M(\tilde{\mu}) = n^{-1}\tilde{\mu}$$

and

$$\text{cov}_R(\tilde{\mu}) = n^{-2}\Sigma (y_i-\tilde{\mu})^2 \ .$$

We consider the behaviour of these two variance estimates under two particular distributions:

(i) $Y_i \sim$ Poisson(μ)

and (ii) $Y_i \sim$ negative binomial with mean μ and variance μ/a $(0<a<1)$.

The first is probably the most familiar distribution satisfying the variance assumption (3.13). The second, which does not satisfy (3.13), is often used to represent overdispersion relative to the Poisson distribution.

Under the Poisson distribution (i), the two variance estimates are both consistent for $\text{var}(\tilde{\mu})=\mu/n$, and their respective variances are

$$\text{var}\{\text{cov}_M(\tilde{\mu})\} = \mu/n^3$$

and

$$\text{var}\{\text{cov}_R(\tilde{\mu})\} = [(n-1)^2\{\mu(1+3\mu)\}-(n-1)(n-3)\mu^2]/n^5 \ .$$

Thus $\mathrm{cov}_M(\tilde{\mu})$ is more efficient than $\mathrm{cov}_R(\tilde{\mu})$. The asymptotic relative efficiency is $(1+2\mu)^{-1}$, so the loss of efficiency incurred by using the robust variance estimate is severe when μ is large.

Under the negative binomial distribution (ii), we have

$$E\{\mathrm{cov}_M(\tilde{\mu})\} = \mu/n \ ,$$

$$\mathrm{var}\{\mathrm{cov}_M(\tilde{\mu})\} = \mu/(n^3 a) \ ,$$

$$E\{\mathrm{cov}_R(\tilde{\mu})\} = (n-1)\mu/(n^2 a)$$

and $$\mathrm{var}\{\mathrm{cov}_R(\tilde{\mu})\} = [\mu(1+4b+b^2)/(na^3) + 2\mu^2/\{(n-1)a^2\}]\,(n-1)^2/n^4 \ ,$$

where $b=1-a$. The true variance of $\tilde{\mu}$ is $\mu/(na)$, so $\mathrm{cov}_M(\tilde{\mu})$ is inconsistent. Both $\mathrm{cov}_M(\tilde{\mu})$ and $\mathrm{cov}_R(\tilde{\mu})$ have variances that are $O(1/n^3)$; both are biased, but the bias of $\mathrm{cov}_R(\tilde{\mu})$ is $O(1/n^2)$ and therefore unimportant as $n\to\infty$. The bias of $\mathrm{cov}_M(\tilde{\mu})$ is $O(1/n)$. In terms of mean squared error, then, there is some smallest sample size, $n^*(\mu,a)$ say, such that $\mathrm{cov}_R(\tilde{\mu})$ performs better than $\mathrm{cov}_M(\tilde{\mu})$ for all $n>n^*$; a close approximation to n^* may be found by solving

$$\mu/(n^3 a) + b^2\mu^2/(n^2 a^2) = \{\mu(1+4b+b^2)/a^3 + 2\mu^2/a^2\}/n^3$$

to obtain

$$n^*(\mu,a) \cong (2\lambda + 6b)/(b^2\lambda) \ , \qquad (3.14)$$

where $\lambda=\mu a$, and $b=1-a$ as before. Table 12 gives some numerical values.

The values of n^* are not very revealing in isolation: some reference scale is required. A suitable reference value might be the number of observations necessary for the overdispersion, relative to (3.13), to be 'detectible' in some sense. As a rough guide, consider the value, n_D say, that solves

$$E\{\Sigma(Y_i-\tilde{\mu})^2/\tilde{\mu}\} - (n-1) = 2\sqrt{\{2(n-1)\}} \ ,$$

so that for $n>n_D$ the expectation of the Pearson chi-square statistic is more than two approximate standard deviations away from its approximate mean under the null, i.e. Poisson, hypothesis. The further approximation $E\{\Sigma(Y_i-\tilde{\mu})^2/\tilde{\mu}\}\cong(n-1)/a$ now yields

$$n_D \cong 1 + 8/(a^{-1}-1)^2 , \tag{3.15}$$

and values based on (3.15) are given in the last row of Table 12. Note that the approximations made here improve as λ increases.

TABLE 12. *Approximation to* $n^*(\mu,a)$ *based on* (3.14)

$\lambda=\mu a$	a = 0.5	0.6	0.7	0.8	0.9	1.0
0.25	56.0	72.5	102.2	170.0	440.0	∞
1	20.0	27.5	42.2	80.0	260.0	∞
5	10.4	15.5	26.2	56.0	212.0	∞
20	8.6	13.3	23.2	51.5	203.0	∞
1000	8.0	12.5	22.2	50.0	200.1	∞
∞	8.0	12.5	22.2	50.0	200.0	∞
n_D from (3.15)	9.0	19.0	44.6	129.0	649.0	∞

The numerical values in Table 12 indicate that the 'cross-over' sample size n^*, although possibly very large, need not be so large that failure of (3.13) is readily detectible. From this point of view the advantages of $\text{cov}_R(\tilde{\mu})$ under failure of (3.13) are not unimportant in practice.

In choosing between the two variance estimates, then, the non-robustness of $\text{cov}_M(\tilde{\mu})$ must be weighed against the potentially large loss of efficiency involved in using $\text{cov}_R(\tilde{\mu})$. The example given here is not particularly special, and the same considerations will apply more generally. In principle, at least, the above calculations could be extended to a more general setting; it is not clear, however, that mean squared

error is necessarily a good criterion by which to judge variance estimates. Perhaps a more relevant comparison between model-based and robust standard errors would be in terms of the coverage properties of approximate confidence intervals based on them, but such a comparison has not been attempted.

3.4 A 'partially Bayes' approach

3.4.1 *Compromise*

The conflict, illustrated in the previous section, between considerations of robustness and efficiency suggests a natural 'compromise' covariance estimate of the form

$$\mathrm{cov}_\lambda(\tilde{\beta}) = (1-\lambda)\mathrm{cov}_M(\tilde{\beta}) + \lambda\mathrm{cov}_R(\tilde{\beta}) \ , \qquad (3.16)$$

with $\lambda\in[0,1]$. In the context of such a family of estimates, two questions that immediately arise are: (i) what interpretation should be given to different values of λ? and (ii) how might λ be chosen, if a single choice is necessary? These questions are, of course, related.

Here we show how a limited, but nevertheless revealing answer may be derived from a 'partially Bayes' viewpoint, in a spirit similar to that of Cox (1975): the approach is introduced in a fully parametric setting in §3.4.2, and extension to quasi-likelihood models using 'linear Bayes' methods is discussed in §3.4.3.

3.4.2 *'Partially Bayes' derivation : normal errors*

It will be shown here how a covariance estimate of the type (3.16) may be derived from a model in which the error variance varies randomly among observations. A key step will be an application of Bayes' theorem, requiring much stronger distributional assumptions than the second-moment assumptions of quasi-likelihood models.

Consider the following, very specific formulation, which is convenient in allowing explicit calculations. Observations are made on a vector of random variables $Y=(Y_1,\ldots,Y_n)$ having a conditional n-variate normal distribution

$$Y\,\big|\,(\phi_1,\ldots,\phi_n) \sim N_n\{X\beta\,,\,\mathrm{diag}(\phi_i)\} \ , \qquad (3.17)$$

with the conditional variances themselves unobserved and distributed independently and identically as

$$\delta\phi_i^{-1} \sim \chi^2_\nu \qquad (i=1,...,n) \quad , \tag{3.18}$$

where $\delta>0$ and $\nu>2$. The vector of regression parameters, $\beta=(\beta_0,...,\beta_p)$, is regarded as fixed and unknown. Our interest is in the least squares estimate

$$\tilde{\beta} = (X^TX)^{-1}X^TY \quad , \tag{3.19}$$

which is also the quasi-likelihood estimate based on an assumption of constant variance. Consider the 'sampling' properties of $\tilde{\beta}$ in (hypothetical) repeated realizations of the vector $Y=(Y_1,...,Y_n)$ from the conditional distribution (3.17); it is well known that $\tilde{\beta}\,|\,(\phi_1,...,\phi_n)$ is $(p+1)$-variate normally distributed, with

$$E(\tilde{\beta}\,|\,\phi_1,...,\phi_n) = \beta$$

and $$\mathrm{cov}(\tilde{\beta}\,|\,\phi_1,...,\phi_n) = (X^TX)^{-1}X^T\mathrm{diag}(\phi_i)X(X^TX)^{-1} \quad . \tag{3.20}$$

This is of no immediate use because the conditional variances $\{\phi_1,...,\phi_n\}$ are unknown. However, some information about these variances is available, both from the sample $y=(y_1,...,y_n)$ and from the 'prior' (3.18); a simple application of Bayes' theorem shows that, *a posteriori*, the variances $\{\phi_1,...,\phi_n\}$ are distributed independently as

$$(\phi_i^{-1}\,|\,Y=y) \sim \chi^2_{\nu+1}/\{\delta+(y_i-\mu_i)^2\} \qquad (i=1,...,n) \quad , \tag{3.21}$$

where $\mu_i=\Sigma x_{ir}\beta_r$. In particular, this means that

$$E(\phi_i\,|\,Y=y) = \{\delta+(y_i-\mu_i)^2\}/(\nu-1) \qquad (i=1,...,n) \quad , \tag{3.22}$$

so that $\{\phi_1,...,\phi_n\}$ may be eliminated from (3.20) by taking the posterior expectation

$$E_{(\phi_1,\ldots,\phi_n \mid Y=y)}\{\mathrm{cov}(\tilde{\beta} \mid \phi_1,\ldots,\phi_n)\} = (X^TX)^{-1}X^T\mathrm{diag}[\{\delta+(y_i-\mu_i)^2\}/(\nu-1)]X(X^TX)^{-1}. \quad \ldots(3.23)$$

To derive an estimate from (3.23), make the substitutions

$$\beta = \tilde{\beta} \qquad (3.24)$$

and

$$\delta/(\nu-2) = \tilde{\phi} \qquad (3.25)$$

so that (3.23) becomes

$$(X^TX)^{-1}X^T\mathrm{diag}[\{\tilde{\phi}(\nu-2)+(y_i-\tilde{\mu}_i)^2\}/(\nu-1)]X(X^TX)^{-1} .$$

Then the 'compromise' estimate (3.16) is formally recovered by putting $\lambda=1/(\nu-1)$.

The motivation for (3.24) is obvious, and (3.25) arises naturally from an 'empirical Bayes' type of argument. For it is easily shown that

$$E\{(Y_i-\tilde{\mu}_i)^2 \mid \phi_1,\ldots,\phi_n\} = \phi_i - 2\phi_i h_{ii} + h_i\Phi h_i^T \qquad (3.26)$$

where h_i is the ith row and $h_{ii}=h_ih_i^T$ the ith diagonal element of the 'hat' matrix $H=X(X^TX)^{-1}X^T$, and Φ is the diagonal matrix with entries $\{\phi_i\}$. Hence, unconditionally,

$$E\{(Y_i-\tilde{\mu}_i)^2\} = (1-h_{ii})E(\phi_i) = (1-h_{ii})\delta/(\nu-2)$$

and so, since $\Sigma(1-h_{ii})=n-q$, $\tilde{\phi}=(n-q)^{-1}\Sigma(y_i-\tilde{\mu}_i)^2$ is an unbiased estimate of the 'prior mean' $\delta/(\nu-2)$.

Thus the indexing parameter λ of the family of 'compromise' estimates (3.16) has an interpretation in terms of the shape parameter of an assumed underlying distribution for the variances $\{\phi_1,\ldots,\phi_n\}$. The 'model-based' estimate $\mathrm{cov}_M(\tilde{\beta})$ and the 'robust' estimate $\mathrm{cov}_R(\tilde{\beta})$ are both extreme limiting cases: the value $\lambda=0$ corresponds to

$$\phi_i^{-1} \sim \lim_{\nu\to\infty} \chi^2_\nu/\delta \qquad \{\delta/(\nu-2) \text{ fixed}\}$$

which has zero variance; and $\lambda=1$ corresponds to

$$\phi_i^{-1} \sim \lim_{\nu \to 2} \chi^2_\nu / \delta \qquad \{\delta/(\nu-2) \text{ fixed}\},$$

which is improper. Intermediate values of λ correspond to 'degrees of freedom' $\nu=\lambda^{-1}+1$ in the mixing distribution (3.18).

An alternative calibration, which will be useful in generalizing this approach, is in terms of the *coefficient of variation* of the distribution of $\{\phi_i\}$; as a function of ν, the squared coefficient of variation is $2/(\nu-4)=\vartheta_\phi$, say, provided $\nu>4$. Thus a given value of λ corresponds to $\vartheta_\phi=2\lambda/(1-3\lambda)$, a particular implication being that only values of λ less than $\frac{1}{3}$ correspond to a finite variance for $\{\phi_i\}$.

If a single choice of λ is required, the interpretation in terms of the distribution of $\{\phi_i\}$ again suggests estimation based on the distribution of the squared residuals, whose first moment gave (3.25). An estimate of ν, and hence a value for λ, may be derived by equating the sample variance of the $\{(y_i-\tilde{\mu}_i)^2\}$ to its theoretical value. A point to note in this connection is that the variance of ϕ_i exists only for $\nu>4$, so this procedure never gives a λ-value greater than or equal to $\frac{1}{3}$. An important practical limitation is likely to be imposed by finite sample size; while a detailed study has not been attempted, preliminary considerations suggest that a value of λ calculated from the sample fourth moment of the residuals would be poorly determined in all but very large samples, and it would usually be preferable, if a choice has to be made, to choose λ on other grounds.

3.4.3 *Generalization via 'linear Bayes' approximation*

In the previous section it was shown how, in the case of the unweighted least squares estimate in a linear model with normal errors, the 'compromise' estimate (3.16) may be derived from a 'partially empirical

partially Bayes' approach based on a particular distributional form for the error variances. The crucial feature was the expression, in (3.22), of the 'posterior expectation' of ϕ_i as a linear function of the ith squared error.

It is immediately apparent that the same approach cannot be directly extended to quasi-likelihood models, specified only in terms of the first two moments, because such a specification is insufficient for application of Bayes' theorem. Moreover, even with a fully specified parametric form for the conditional distribution of Y given $(\phi_1,...,\phi_n)$, the existence of a 'prior' that gives rise to a linear form of posterior expectation like (3.22) is exceptional: the normal/inverse-χ^2 combination of (3.17) and (3.18) is rather special in this respect.

However, a degree of generalization is possible using the 'linear Bayes' methods of Hartigan (1969), also discussed by Mouchart & Simar (1982). The key idea is that of the *linear expectation*, defined for scalar random variables X and Θ as

$$\hat{E}(\Theta \,|\, X) = E(\Theta) + \frac{\mathrm{cov}(\Theta,X)}{\mathrm{var}(X)}\{X-E(X)\} \; . \tag{3.27}$$

A possible interpretation of $\hat{E}(\Theta \,|\, X)$ is as a best, in the sense of least squares, approximation of $E(\Theta \,|\, X)$ by a linear function of X.

Consider first the formulation of §3.4.2, but without the restriction to normal errors. That is, (3.17) is replaced by the moment specification

$$\left.\begin{aligned} E(Y \,|\, \phi_1,...,\phi_n) &= X\beta \\ \mathrm{cov}(Y \,|\, \phi_1,...,\phi_n) &= \mathrm{diag}(\phi_i) \; . \end{aligned}\right\} \tag{3.28}$$

In place of the fully parametric specification (3.18) for the distribution of $\{\phi_i\}$, write

$$\left.\begin{aligned} E(\phi_i) &= \mu_\phi \\ \mathrm{var}(\phi_i) &= \vartheta_\phi \mu_\phi^2 \; . \end{aligned}\right\} \tag{3.29}$$

The conditional covariance matrix of the least squares estimate, $\tilde{\beta}$, is still given by (3.20); its exact posterior expectation can no longer be calculated. However, if we note that

$$\begin{aligned}\mathrm{cov}\{\phi_i, (Y_i-\mu_i)^2\} &= \mathrm{cov}[\phi_i, E\{(Y_i-\mu_i)^2 \,|\, \phi_i\}] \\ &= \mathrm{var}(\phi_i) = \vartheta_\phi \mu_\phi^2\end{aligned}$$

and $$E\{(Y_i-\mu_i)^2\} = E[E\{(Y_i-\mu_i)^2 \,|\, \phi_i\}] = \mu_\phi ,$$

and assume that the conditional standardized fourth moment of Y_i,

$$\rho_4 = E\{(Y_i-\mu_i)^4 \,|\, \phi_i\}/\phi_i^2 ,$$

exists and is the same for all i, then

$$\begin{aligned}\mathrm{var}\{(Y_i-\mu_i)^2\} &= E[\mathrm{var}\{(Y_i-\mu_i)^2 \,|\, \phi_i\}] + \mathrm{var}[E\{(Y_i-\mu_i)^2 \,|\, \phi_i\}] \\ &= (\rho_4-1)E(\phi_i^2) + \mathrm{var}(\phi_i) \\ &= (\rho_4\vartheta_\phi + \rho_4 - 1)\,\mu_\phi^2 , \qquad (3.30)\end{aligned}$$

and so, from (3.27),

$$\begin{aligned}\hat{E}\{\phi_i \,|\, (Y_i-\mu_i)^2=(y_i-\mu_i)^2\} &= \mu_\phi + \vartheta_\phi\{(y_i-\mu_i)^2-\mu_\phi\}/(\rho_4\vartheta_\phi+\rho_4-1) \\ &= \{\mu_\phi(\rho_4-1)(\vartheta_\phi+1) + \vartheta_\phi(y_i-\mu_i)^2\}/(\rho_4\vartheta_\phi+\rho_4-1) . \qquad (3.31)\end{aligned}$$

The normal/inverse-χ^2 combination considered in §3.4.2 has $\rho_4=3$ and $\vartheta_\phi=2/(\nu-4)$ (provided $\nu>4$), and it is easily verified that the exact posterior expectation (3.22) and the linear version (3.31) are identical in that case.

A more general version of (3.23) now follows in an obvious way, by replacing each ϕ_i in (3.20) with its linear expectation (3.31). An estimate of $\mathrm{cov}(\tilde{\beta})$ is derived by making the natural substitutions $\beta=\tilde{\beta}$ and $\mu_\phi=\tilde{\phi}$, and the 'compromise' estimate (3.16) is formally recovered by putting

$$\lambda = \vartheta_\phi/(\rho_4\vartheta_\phi+\rho_4-1) \qquad (\vartheta_\phi>0\,,\ \rho_4>1)\ . \tag{3.32}$$

The form of (3.32) is interesting. As a function of ϑ_ϕ, at fixed ρ_4, λ is strictly increasing and converges to $1/\rho_4$ as $\vartheta_\phi\to\infty$; recall that this upper limit, a consequence of assuming finite second moment in the distribution of $\{\phi_i\}$, was $\frac{1}{3}$ under normality. As a function of ρ_4, at fixed ϑ_ϕ, λ is strictly decreasing to a limit value (as $\rho_4\to\infty$) of zero. Thus increasing kurtosis and increasing heteroscedasticity have conflicting effects on the determination of λ, and kurtosis is dominant in the sense that, no matter how large ϑ_ϕ is, λ is bounded above by $1/\rho_4$.

If a sample-based choice of λ is required, the sample variance of the $\{(y_i-\tilde{\mu}_i)^2\}$ may still be used provided at least one of ρ_4 and ϑ_ϕ is known; the same reservations expressed at the end of §3.4.2 apply also here. If ρ_4 and ϑ_ϕ are both unknown there is a problem in that unless, for example, there exist identifiable groups of observations within which ϕ_i can be assumed constant, kurtosis and heteroscedasticity are not separately identifiable; from (3.30) the estimable quantity is essentially $\rho_4\vartheta_\phi+\rho_4$, which does not determine a value for λ.

We have shown, then, how the approach of §3.4.2 may be extended, using the idea of a linear expectation, from the normality (3.17) to a wider family of error distributions indexed by the fourth moment. To conclude this section, we indicate how the development may be generalized to situations where $\mu_i(\beta)$ is a non-linear model and $\tilde{\beta}$ is the quasi-likelihood estimate based on some more general variance function $V(\mu)$. By analogy with (3.28), assume

$$E(Y_i\,|\,\phi_1,\ldots,\phi_n) = \mu_i(\beta)$$

and

$$\operatorname{cov}(Y_i\,|\,\phi_1,\ldots,\phi_n) = \operatorname{diag}\{\phi_i V(\mu_i)\}\ .$$

An exact expression for $\operatorname{cov}(\tilde{\beta}\,|\,\phi_1,\ldots,\phi_n)$, corresponding to (3.20), is not now

generally available, but the same arguments as before apply directly to the asymptotic approximation (1.13). The only change necessary is to replace $(Y_i-\mu_i)$ by $(Y_i-\mu_i)/\{V(\mu_i)\}^{\frac{1}{2}}$ throughout. In particular, the 'compromise' covariance estimate (3.16) is still derived by making the substitutions $\beta=\tilde{\beta}$, $\mu_\phi=\tilde{\phi}$ and $\lambda=\vartheta_\phi/(\rho_4\vartheta_\phi+\rho_4-1)$ in the linear expectation of $\text{cov}(\tilde{\beta}\,|\,\phi_1,\ldots,\phi_n)$, but now with

$$\rho_4 = E[(Y_i-\mu_i)^4/\{V(\mu_i)\}^2\,|\,\phi_i]/\phi_i^2 \ .$$

Provided that ρ_4 exists, and is constant over observations, the generalization is complete.

3.5 A remark about bias correction

As noted in §3.2 , the robust covariance estimate $\mathrm{cov}_R(\tilde{\beta})$ is biased in finite samples. A major component of the bias arises from the fact that $(y_i-\tilde{\mu}_i)^2$ is not an unbiased estimate of $\mathrm{var}(Y_i)$.

Consider the case $\{V(\mu)=1\,,\ g(\mu)=\mu\}$, i.e. unweighted least squares estimation in linear models. In addition to being the most common application of quasi-likelihood estimation, this case has the advantage of allowing exact bias-correction; here the tendency to underestimate $\mathrm{var}(Y_i)$ is the *only* source of bias. Explicitly, suppose that

$$E(Y) = X\beta$$

and

$$\mathrm{cov}(Y) = \Phi\ ,$$

where $\Phi=\mathrm{diag}(\phi_i)$ as before. The least squares estimator is

$$\tilde{\beta} = (X^TX)^{-1}X^TY$$

with exact covariance matrix

$$\mathrm{cov}(\tilde{\beta}) = (X^TX)^{-1}X^T\Phi X(X^TX)^{-1}\ . \tag{3.33}$$

The robust covariance estimate (3.6) simply replaces each ϕ_i in (3.33) with $(y_i-\tilde{\mu}_i)^2$. Its expectation under homoscedasticity, with $\phi_i=\phi$ for all i, is

$$E\{\mathrm{cov}_R(\tilde{\beta})\} = \phi(X^TX)^{-1}X^T\mathrm{diag}(1-h_{ii})X(X^TX)^{-1}\ , \tag{3.34}$$

where h_{ii} is the ith diagonal element of $H=X(X^TX)X^T$; consistency is based on standard assumptions that include $h_{ii}\to 0$ for all i as $n\to\infty$. In finite samples h_{ii} need not be small for all i and the bias in $\mathrm{cov}_R(\tilde{\beta})$, even under homoscedasticity, can be severe. Chesher & Jewitt (1984, 1986) give examples and calculate bounds, based on $\max\{h_{ii}\}$, for the bias.

Two bias-corrected versions that have been suggested in the literature are

$$A = \{n/(n-q)\}\,\mathrm{cov}_R(\tilde{\beta})$$

and $$B = (X^TX)^{-1}X^T\mathrm{diag}\{(y_i-\tilde{\mu}_i)^2/(1-h_{ii})\}X(X^TX)^{-1} .$$

The 'degrees of freedom' corrected estimate A was derived as a jackknife estimate by Hinkley (1977); it is exactly unbiased when ϕ_i and h_{ii} are both constant, i.e. for balanced designs under homoscedasticity. The estimate B has been suggested by Chesher & Jewitt (1984) and by MacKinnon & White (1985); by (3.26) it is unbiased under homoscedasticity, regardless of $\{h_{ii}\}$. An implicit assumption here is that $\max\{h_{ii}\}<1$.

A third possibility is the 'direct' bias-correction

$$C = \mathrm{cov}_R(\tilde{\beta}) \div E\{\mathrm{cov}_R(\tilde{\beta})\} \times \mathrm{cov}(\tilde{\beta}) ,$$

which is well-defined, and unbiased, under homoscedasticity; here '$\div$' and '$\times$' act element-by-element on the matrices concerned.

The matrices A, B and C coincide in the balanced case when the $\{h_{ii}\}$ are all equal to q/n.

The point to be made here is simply that division of $(y_i-\tilde{\mu}_i)^2$ by $(1-h_{ii})$, although bias-correcting, is also variance-inflating. As a result the variability of estimate B may be rather large, particularly when $\max\{h_{ii}\}$ is close to 1.

To illustrate, consider a highly unbalanced $n\times 2$ model matrix X with

$$x_{i0} = 1 \qquad (i=1,\ldots,n) ,$$

$$x_{i1} = \begin{cases} 0 & (i=1,\ldots,n-2) \\ 6 & (i=n-1) \\ 1 & (i=n) \end{cases} ,$$

and $n\geq 3$. The diagonal elements of the 'hat' matrix in this case are

$$h_{ii} = \begin{cases} (\delta^2+1)/d & (i=1,\ldots,n-2) \\ \{(n-1)\delta^2-2\delta+1\}/d & (i=n-1) \\ \{\delta^2-2\delta+(n-1)\}/d & (i=n) \ , \end{cases}$$

where $d = \det(X^TX) = n(\delta^2+1)-(\delta+1)^2$. When δ is small, h_{nn} is close to 1.

In this example the 'slope' estimate $\tilde{\beta}_1$ has true variance, under homoscedasticity,

$$\mathrm{var}(\tilde{\beta}_1) = \phi n/d \ , \tag{3.35}$$

which has a limit value of $\phi n/(n-1)$ as $\delta \to 0$. The corresponding element of $\mathrm{cov}_R(\tilde{\beta})$ is

$$[(\delta+1)^2 \sum_{i=1}^{n-2} (y_i-\tilde{\mu}_i)^2 + \{\delta(n-1)-1\}^2(y_{n-1}-\tilde{\mu}_{n-1})^2 + (\delta-n+1)^2(y_n-\tilde{\mu}_n)^2\,]/d^2 \ , \tag{3.36}$$

with expectation

$$\phi(n-2)[\,(\delta+1)^2\{(\delta^2+1)(n-2)-2\delta\} + \{\delta(n-1)-1\}^2 + \delta^2(\delta-n+1)^2\,]/d^3 \ , \tag{3.37}$$

which has a limit value of $\phi(n-2)/(n-1)^2$ as $\delta \to 0$. So $\mathrm{cov}_R(\tilde{\beta})$ is subject to substantial bias when δ is small.

Bias-correction, under homoscedasticity, is provided by either B or C above. The element of B corresponding to $\mathrm{var}(\tilde{\beta}_1)$, call it b_{11}, is derived by dividing each $(y_i-\tilde{\mu}_i)^2$ in (3.36) by $(1-h_{ii})$. The corresponding element of C, call it c_{11}, is calculated by multiplying (3.36) by the ratio of (3.35) to (3.37). Both b_{11} and c_{11} are quadratic in the observations, so their variances depend on the shape of the error distribution. Some indication of the general behaviour comes from calculation under an assumption of normality for the $\{Y_i-\mu_i\}$: the main feature is that, whereas the variance of $(Y_n-\tilde{\mu}_n)^2$ is $O(\delta^4)$ as $\delta \to 0$, $(Y_n-\tilde{\mu}_n)^2/(1-h_{ii})$ has a constant variance of $2\phi^2$, and as a consequence it may be shown that

$$\mathrm{var}(b_{11}) = 2\phi^2[1 + 1/\{(n-1)^2(n-2)\}] + O(\delta)$$

and

$$\mathrm{var}(c_{11}) = 2\phi^2n^2/\{(n-1)^2(n-2)\} + O(\delta) \ .$$

For small δ, then, b_{11} is much more variable than c_{11}, even when n is quite small.

This rather extreme example serves to illustrate the general point. While B and C may both be thought of as generalizations of A to the unbalanced case, their behaviour can differ greatly; in particular, the stability of C when $\max\{h_{ii}\}$ is close to 1 is not shared by B. It seems preferable, in unbalanced situations, to apply bias-correction *after* the squared residuals have been 'smoothed' into $\mathrm{cov}_R(\tilde{\beta})$.

While the focus here has been on the case $\{V(\mu)=1, g(\mu)=\mu\}$, it is immediately apparent that 'bias-corrected' estimates analogous to A, B and C may be defined more widely. Although exact unbiasedness is not usually possible in the context of non-linear models or estimates other than least squares, the same general considerations apply.

CHAPTER 4

Extended quasi-likelihood and double exponential families

4.1 Introduction

This short chapter will explore and make explicit the relationship between two recently proposed constructions, these being the *extended quasi-likelihood* of Nelder & Pregibon (1983) and the *double exponential families* of Efron (1986). The motivation for, and use of, these ideas will be mentioned only briefly, for they are extensively described in the aforementioned references. Rather, the aim here will be to give some specific details of the similarities, mentioned in the rejoinder of Diaconis & Efron (1985) and again in Efron (1986), between the two constructions.

4.2 Extended quasi-likelihood

An important function, related to a given variance function $V(\mu)$, is the *deviance*, defined for a single observation as

$$D(y;\mu) = -2\int_{y}^{\mu} \frac{y-u}{V(u)}\,du \quad . \tag{4.1}$$

The deviance measures the discrepancy between an observation y and its expected value μ, and is useful in comparing different regression models for the same data; in particular, if H_r and H_s are two hypotheses of dimension $r<s$, H_r nested within H_s, then under H_r

$$\sum_{i=1}^{n} D(\tilde{\mu}_i^{(s)};\tilde{\mu}_i^{(r)}) = \sum_{i=1}^{n} \{D(y_i;\tilde{\mu}_i^{(r)}) - D(y_i;\tilde{\mu}_i^{(s)})\}$$

has an asymptotic χ^2_{s-r} distribution (McCullagh, 1983).

Nelder & Pregibon (1983), see also McCullagh & Nelder (1983, pp212-14), consider the problem of comparing different *variance*

functions on the same set of data, pointing out that the variance function determines the scale on which $D(y;\mu)$ is measured, so that differencing across variance functions is unhelpful. With the aim of making the deviance behave more like a log-likelihood, Nelder & Pregibon (1983) define an *extended quasi-likelihood*

$$a(y;\mu,\phi) = \exp[-\tfrac{1}{2}\log\{2\pi\phi V(y)\} - \tfrac{1}{2}D(y;\mu)/\phi] \ ; \qquad (4.2)$$

Nelder & Pregibon (1983) use the notation $Q^+(y;\mu)$, and McCullagh & Nelder (1983, equation 11.2) use l', in both cases to stand for $\log\{a(y;\mu,\phi)\}$. Problems of definition when $V(y)=0$ are discussed by Nelder & Pregibon (1983) and by McCullagh & Nelder (1983, p214).

Nelder & Pregibon (1983) and McCullagh & Nelder (1983, p213) use (4.2) as an approximate likelihood for inference in the context of a parametric family of variance functions, e.g. $\{V(\mu)=\mu^\lambda: -\infty<\lambda<\infty\}$, and give examples of its application.

4.3 Double exponential families

Diaconis & Efron (1985,§5) introduce a generalization of natural exponential families, discussed also by Efron (1986) under the name *double exponential families.* In Efron's (1986) notation our interest here is in $\tilde{f}_{\mu,\phi^{-1},1}(y)$; we shall write this as

$$b^*(y;\mu,\phi) = B(\mu,\phi)\phi^{-\frac{1}{2}}f(y;y)\exp\{-\tfrac{1}{2}D(y;\mu)/\phi\} \qquad (4.3)$$

$$= B(\mu,\phi)b(y;\mu,\phi) \ , \text{ say.}$$

Here

$$f(y;\mu) = p(y)\exp\{y\theta - u(\theta)\} \qquad (4.4)$$

is a one-parameter natural exponential family with respect to measure $F(y)$, taken here to be either Lebesgue or counting measure. The mean and

variance are respectively $\mu=u'(\theta)$ and $V(\mu)=u''(\theta)$. The deviance $D(y;\mu)$ is defined as in (4.1), and $B(\mu,\phi)$ is a normalizing constant. Efron's (1986) $g_{\mu,1}(y)$, $[dG_1(y)]$ and $c(\mu,\phi^{-1},1)$ are respectively $f(y;\mu)/p(y)$, $p(y)[dF(y)]$ and $B(\mu,\phi)$ in our notation.

Diaconis & Efron's (1985) original motivation for (4.3) was to provide a 'random effects' interpretation, analogous to variance components in the normal case, for heterogeneity among means in the context of other exponential families. Efron (1986) gives examples of the use of $b(y_i;\mu_i,\phi_i)$ as an approximate likelihood in regression models for both mean and dispersion, i.e. in models with

$$E(Y_i) = \mu_i(\beta) \qquad (i=1,...,n) \qquad (4.5)$$

and

$$\text{var}(Y_i) = \phi_i(\gamma)V(\mu_i) \qquad (i=1,...,n)\ , \qquad (4.6)$$

with $V(.)$ known, and $\mu_i(.)$, $\phi_i(.)$ known functions of a few unknown parameters. West (1985) considers a similar type of model, but with the 'deterministic' specification of the $\{\phi_i\}$ in (4.6) replaced by an assumption that $\{\phi_1,...,\phi_n\}$ is a random sample from a distribution with a few unknown parameters; the function $b(y;\mu,\phi)$ is precisely the approximate likelihood, his (2.4), used by West as the basis of a Bayesian approach.

4.4 Comparison

4.4.1 *Existence*

The extended quasi-likelihood (4.2) is defined, apart from possible problems when $V(y)=0$, for any given variance function $V(\mu)$. The same is not true of the double exponential family (4.3), which requires the existence of a natural exponential family $f(y;\mu)$, as in (4.4), having mean μ

and variance $V(\mu)$. The existence of such families has been a topic of considerable recent interest: Morris (1982) finds the natural exponential families corresponding to all quadratic variance functions, while Tweedie (1984) and Jorgensen (1987) consider variance functions of the form $V(\mu)=\mu^{\lambda}$, a particular result being that there is no natural exponential family corresponding to powers $\lambda\in(0,1)$.

For a general $V(\mu)$, then, the existence of a corresponding natural exponential family, and hence a double exponential family, is not guaranteed. In the remainder of this chapter attention is restricted to variance functions for which both extended quasi-likelihood and double exponential family are defined; some examples are given in §4.4.3.

4.4.2 *Suggested use*

The original motivation for $a(y;\mu,\phi)$, to enable comparison of different variance functions, and that for $b^*(y;\mu,\phi)$ as a 'random effects' model, are seemingly unconnected. However the use by West (1985) and Efron (1986) of the unnormalized form $b(y;\mu,\phi)$ as an approximate likelihood in models like {(4.5),(4.6)} has much in common with Nelder & Pregibon's (1983) use of $a(y;\mu,\phi)$ in models with a parametric family of variance functions. Indeed, the Nelder-Pregibon type of variance specification

$$\text{var}(Y_i) = \phi V^*(\mu_i;\lambda)$$

is formally included in the general form (4.6) by writing, for example, $V(\mu_i)=V^*(\mu_i;0)$ and $\phi_i(\gamma)=\phi V^*(\mu_i;\lambda)/V^*(\mu_i;0)$ so that $\gamma=\{\beta,\lambda,\phi\}$. Pregibon (1984) has suggested that $a(y_i;\mu_i,\phi_i)$ be used as an approximate likelihood for models with the general type of variance specification (4.6).

4.4.3 *Comparison of approximate likelihoods*

Here we consider, in cases where both are defined, the functions

$$a(y;\mu,\phi) = \exp[-\tfrac{1}{2}\log\{2\pi\phi V(y)\} - \tfrac{1}{2}D(y;\mu)/\phi]$$

and $$b(y;\mu,\phi) = \phi^{-\frac{1}{2}} f(y;y)\exp\{-\tfrac{1}{2}D(y;\mu)/\phi\} ,$$

whose ratio is

$$b(y;\mu,\phi)/a(y;\mu,\phi) = f(y;y)\{2\pi V(y)\}^{\frac{1}{2}} , \qquad (4.7)$$

a function of y only. An immediate conclusion to be drawn is that the functions $a(y_i;\mu_i,\phi_i)$ and $b(y_i;\mu_i,\phi_i)$, considered as approximate likelihood functions, lead to the same maximum likelihood estimates.

It remains to investigate the behaviour of (4.7) as a function of y. Efron (1986) points out that an assumption of asymptotic normality for $f(y;\mu)$ as some notional parameter, N say, tends to infinity, implies in particular that

$$f(y;y) \to \{2\pi V(y)\}^{\frac{1}{2}} \quad \text{for all } y, \quad \text{as } N\to\infty ,$$

which suggests that the ratio (4.7) may be approximately equal to 1. Consider the nature of this approximation in some particular cases:

(i) $V_k(\mu)=k$, ($k>0$, not depending on μ)

$$f_k(y;\mu) = (2\pi k)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(y-\mu)^2/k\} \quad (-\infty<y,\mu<\infty),$$

$$D_k(y;\mu) = (y-\mu)^2/k,$$

$$a_k(y;\mu,\phi) = b_k(y;\mu,\phi) = (2\pi\phi k)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(y-\mu)^2/(\phi k)\};$$

$$= f_{\phi k}(y;\mu);$$

(ii) $V(\mu)=\mu$, $$f(y;\mu) = \mu^y e^{-\mu}/y! \qquad (\mu>0,\ y=0,1,2,...),$$

$$D(y;\mu) = 2\{y\log(y/\mu)-(y-\mu)\},$$

$$a(y;\mu,\phi) = (2\pi\phi y)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}D(y;\mu)/\phi\},$$

$$b(y;\mu,\phi) = \phi^{-\frac{1}{2}}(y^y e^{-y}/y!)\exp\{-\tfrac{1}{2}D(y;\mu)/\phi\},$$

so $b(y;\mu,\phi)/a(y;\mu,\phi) = y^y e^{-y}(2\pi y)^{\frac{1}{2}}/y!$;

(iii) $V(\mu)=\mu(N-\mu)/N \qquad (N=1,2,3,...)$,

$$f(y;\mu) = \binom{N}{y}(\mu/N)^y(1-\mu/N)^{N-y} \quad (0<\mu<N,\ y=0,1,...,N),$$

$$D(y;\mu) = 2[y\log(y/\mu)+(N-y)\log\{(N-y)/(N-\mu)\}],$$

$$a(y;\mu,\phi) = \{2\pi\phi y(N-y)/N\}^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}D(y;\mu)/\phi\},$$

$$b(y;\mu,\phi) = \phi^{-\frac{1}{2}}[N!(y/N)^y(1-y/N)^{N-y}/\{y!(N-y)!\}]\exp\{-\tfrac{1}{2}D(y;\mu)/\phi\},$$

so $b/a = \{2\pi y(N-y)/N\}^{\frac{1}{2}}N!(y/N)^y(1-y/N)^{N-y}/\{y!(N-y)!\}$;

(iv) $V_k(\mu)=k\mu^2 \qquad (k>0)$,

$$f_k(y;\mu) = y^{-1}(y/\mu k)^{1/k}\exp(-y/\mu k)/\Gamma(1/k) \qquad (y,\mu>0),$$

$$D_k(y;\mu) = 2\{(y-\mu)/\mu-\log(y/\mu)\}/k,$$

$$a_k(y;\mu,\phi) = (2\pi\phi k y^2)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}D_k(y;\mu)/\phi\},$$

$$b_k(y;\mu,\phi) = \phi^{-\frac{1}{2}}y^{-1}(ek)^{-1/k}\{\Gamma(1/k)\}^{-1}\exp\{-\tfrac{1}{2}D_k(y;\mu)/\phi\},$$

so $b/a = (2\pi k)^{\frac{1}{2}}(ek)^{-1/k}\{\Gamma(1/k)\}^{-1}$; and

(v) $V_k(\mu)=k\mu^3 \qquad (k>0)$,

$$f_k(y;\mu) = (2\pi k y^3)^{-\frac{1}{2}}\exp\{-(y-\mu)^2/(2k\mu^2 y)\} \qquad (y,\mu>0),$$

$$D_k(y;\mu) = (y-\mu)^2/(k\mu^2 y),$$

$$a_k(y;\mu,\phi) = b_k(y;\mu,\phi) = f_{\phi k}(y;\mu)\ .$$

Some points to note are:

(*a*) $b(y;\mu,1) = f(y;\mu)$ exactly, always;

(b) in the 'normal' and 'inverse Gaussian' cases (i) and (v), $a_k(y;\mu,\phi)=b_k(y;\mu,\phi)=f_{\phi k}(y;\mu)$ exactly, for all k, ϕ, y and μ;

(*c*) in both of the discrete examples (ii) and (iii), substitution of the Stirling approximation

$$x! \sim (2\pi x)^{\frac{1}{2}}x^x e^{-x}$$

for all factorials makes $b(y;\mu,\phi)$ equal to $a(y;\mu,\phi)$; and

(*d*) in the 'gamma' case (iv), b/a depends only on k and is approximately equal to 1, at least for small k; the approximation here is given by the slightly different Stirling formula

$$\Gamma(x) \sim (2\pi x)^{\frac{1}{2}}x^{x-1}e^{-x} \qquad (4.8)$$

applied at $x=1/k$. Moreover, use of the approximation (4.8) at $x=1/(\phi k)$ gives $f_{\phi k}(y;\mu)/a_k(y;\mu,\phi)\cong 1$. In this sense a_k and b_k are both close to the gamma likelihood $f_{\phi k}$. Note, however, that while $f_{\phi k}(y;\mu)$ and $a_k(y;\mu,\phi)$ both depend on ϕ and k only through the product ϕk, this property is not shared by $b_k(y;\mu,\phi)$; a consequence is that the approximate likelihoods $\{b_k(y;\mu,\phi/k)\colon k>0\}$, all of which represent the same variance specification $\mathrm{var}(Y_i)=\phi\mu_i^2$, are not the same for all k. Indeed, the ratio

$$b_1(y;\mu,\phi)/b_k(y;\mu,\phi/k)$$

is unbounded as $k\to\infty$. However this ambiguity, although a rather curious feature, is not a problem in practice since 'approximate likelihood ratios' relevant to the comparison of two hypothesised coefficients of variation, $\sqrt{\phi}$ and $\sqrt{\psi}$ say, have the form

$$b_k(y;\mu,\phi/k)/b_k(y;\mu,\psi/k),$$

which does *not* depend on the choice of k.

Nelder & Pregibon (1983) point out that in all of the above examples, and more generally, $a(y;\mu,\phi)$ is the (unnormalized) saddlepoint approximation to the natural exponential family distribution with $\mathrm{var}(Y)=\phi V(\mu)$, when such a family exists. In the notation used above, then, $a(y;\mu,\phi)$ is the unnormalized saddlepoint approximation to $b_\phi(y;\mu,1)$ when $b_\phi(y;\mu,1)$ is defined.

4.4.4 *A remark about estimation*

In the previous section it was found that approximate likelihoods based on $a(y_i;\mu_i,\phi_i)$ and on $b(y_i;\mu_i,\phi_i)$ are equivalent as far as estimation is concerned. Estimating equations, e.g. for β and γ in models like {(4.5),(4.6)}, may be based on the approximate score functions

$$\partial(\log a)/\partial\mu = \partial(\log b)/\partial\mu = (y-\mu)/\{\phi V(\mu)\} \qquad (4.9)$$

and

$$\partial(\log a)/\partial\phi = \partial(\log b)/\partial\phi = \tfrac{1}{2}\{D(y;\mu)-\phi\}/\phi^2. \qquad (4.10)$$

If $\beta=(\beta_0,...,\beta_p)$ and $\gamma=(\gamma_0,...,\gamma_t)$, say, are distinct sets of parameters, the estimating equations are

$$\sum_{i=1}^{n} \frac{y_i-\mu_i}{\phi_i V(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \beta_r} = 0 \qquad (r=0,...,p) \qquad (4.11)$$

and

$$\sum_{i=1}^{n} \phi_i^{-1} \left\{ \frac{D(y_i;\mu_i)}{\phi_i} - 1 \right\} \frac{\partial \phi_i}{\partial \gamma_s} \qquad (s=0,...,t) \ . \qquad (4.12)$$

Here we note that, while (4.11) are unbiased estimating equations on account of (4.9) having zero expectation, the equations (4.12) are not, in general, unbiased under {(4.5),(4.6)}. From (4.10) we see that the equations (4.12) will generally yield estimates that are consistent for γ in the specification

$$E[D(Y_i;\mu_i)] = \phi_i(\gamma) \qquad (i=1,...,n) \ , \qquad (4.13)$$

which is generally different from (4.6).

A possible basis for assuming the two specifications (4.6) and (4.13) to be approximately the same is the Taylor series approximation

$$E[D(Y;\mu)] \cong D(\mu;\mu) + \tfrac{1}{2}D''(\mu,\mu)\mathrm{var}(Y) \ ,$$

where $D''(y;\mu)$ is the second derivative of $D(y;\mu)$ with respect to y; since $D(\mu,\mu)=0$ and $D'(y;\mu)=-2\int_y^{\mu}\{1/V(u)\}du$, so that $D''(\mu,\mu)=2/V(\mu)$, this becomes

$$E[D(y;\mu)] \cong \mathrm{var}(Y)/V(\mu) \, .$$

However, this approximation may be poor in particular instances. As a simple illustration, take $V(\mu)=\mu^2$ and $\phi_i(\gamma)=\phi$, an unknown constant, and consider behaviour under the lognormal distribution, which was one of the examples of §2.3.1. The deviance for a single observation is

$$D(y;\mu) = 2\{(y-\mu)/\mu - \log(y/\mu)\}$$

with expectation, under the lognormal,

$$E[D(Y;\mu)] = 2\{\log\mu - E(\log Y)\} = \log(1+\phi) \ .$$

Thus $\tilde{\phi}$ given by solving (4.12) is consistent not for ϕ but for $\log(1+\phi)$; clearly the approximation deteriorates as ϕ increases, e.g. $\phi=1$ has $\log(1+\phi)=0.69$. The lognormal here was not chosen as an extreme case, rather for algebraic simplicity: behaviour in the other examples of §2.3.1 is qualitatively the same.

Often the parameters γ are nuisance parameters, β being the object of primary interest. When the estimating equations are 'separate' as in (4.11) and (4.12), β is still consistently estimated; however, inconsistent estimation of γ will usually affect the 'weights' in (4.11), and so reduce the efficiency with which β is estimated.

More generally, when β and γ do not represent distinct sets of parameters, the relationship of consistently-estimated quantities to parameters of interest is ill-defined without further assumptions.

4.4.5 *Normalized versions*

Consider now the normalized densities

$$b^*(y;\mu,\phi) = B(\mu,\phi)b(y;\mu,\phi)$$

and, by analogy,

$$a^*(y;\mu,\phi) = A(\mu,\phi)a(y;\mu,\phi),$$

the normalization in both cases being with respect to the same dominating measure, $F(y)$, used in defining b^*. It is immediately apparent that $B(\mu,1)=1$ for all μ since, as pointed out in §4.4.3, $b(y;\mu,1)=f(y;\mu)$ exactly. More generally, though, the normalizing constants A and B are not unity and depend on both μ and ϕ.

Efron (1986) argues that $b(y;\mu,\phi)$ is appropriate for use as an approximate likelihood, in models like {(4.5),(4.6)}, on the grounds that

(1) $B(\mu,\phi) \cong 1$,

(2) $E(Y) \cong \mu$,

and (3) $\text{var}(Y) \cong \phi V(\mu)$,

the expectation and variance here being with respect to $b^*(y;\mu,\phi)$. The approximations are justified by asymptotic arguments based on an assumption that is essentially limiting normality of $b^*(y;\mu,\phi)$ as some parameter, N say, tends to infinity. In non-normal situations, however, these approximations may be poor.

For a specific, and practically important, example, consider the 'Poisson' case, $V(\mu)=\mu$, for which $a(y;\mu,\phi)$ and $b(y;\mu,\phi)$ are given in §4.4.3. Numerical normalization with respect to counting measure on {0,1,2,...} yields the results given in Table 13; in calculating $a(y;\mu,\phi)$ we follow the suggestion of McCullagh & Nelder (1983, p214) to use $V(y)=y+\frac{1}{6}$ rather than $V(y)=y$, thereby avoiding difficulty at $y=0$. Three values of μ are considered (μ=10.0, 1.0, 0.1) and three values of ϕ (1.0, 2.0 and 3.0).

From Table 13 we see that the normalizing constant, mean and standard deviation for a^* are close, at every value of (μ,ϕ), to the corresponding values for b^*. Moreover, at μ=10 there is also close agreement with the 'theoretical' values given by the approximations (1)-(3) above; in fact the values for b^* in part (i) of Table 13 are given by Efron (1986, Table 2) in support of these approximations. However, the quality of all three approximations clearly deteriorates as μ decreases and as ϕ increases; for example with μ=0.1 and ϕ=2 the means of a^* and b^* are respectively 0.238 and 0.233, more than twice their common 'theoretical' value of 0.1.

Parallel calculations, not reported in detail here, for the 'binomial' variance function $V(\mu)=\mu(N-\mu)/N$ yield qualitatively similar results: a^* and b^* are close to each other but not necessarily to the 'theoretical' mean and variance. Here the approximations (1)-(3) are good when ϕ is close to 1

TABLE 13. *Normalizing constant, mean and standard deviation for* $a^*(y;\mu,\phi)$ *and* $b^*(y;\mu,\phi)$ *in the case* $V(\mu)=\mu$

(i) μ=10.0 ϕ =	1.0	2.0	3.0
$\{A(\mu,\phi)\}^{-1}$	1.000	1.012	1.025
$\{B(\mu,\phi)\}^{-1}$	*1.000*	*1.012*	*1.026*
Mean	10.000	9.969	9.919
	10.000	*9.968*	*9.916*
Standard deviation	3.162	4.481	5.479
	3.162	*4.482*	*5.480*
'Theoretical' s.d. = $\sqrt{(10\phi)}$	3.162	4.472	5.477
(ii) μ=1.0			
$\{A(\mu,\phi)\}^{-1}$	0.993	0.986	0.940
$\{B(\mu,\phi)\}^{-1}$	*1.000*	*0.995*	*0.948*
Mean	1.009	1.143	1.326
	1.000	*1.132*	*1.313*
Standard deviation	0.999	1.354	1.659
	1.000	*1.353*	*1.657*
'Theoretical' s.d. = $\sqrt{\phi}$	1.000	1.414	1.732
(iii) μ=0.1			
$\{A(\mu,\phi)\}^{-1}$	0.980	0.816	0.735
$\{B(\mu,\phi)\}^{-1}$	*1.000*	*0.831*	*0.747*
Mean	0.102	0.238	0.361
	0.100	*0.233*	*0.354*
Standard deviation	0.320	0.533	0.717
	0.316	*0.528*	*0.712*
'Theoretical' s.d. = $\sqrt{(\phi/10)}$	0.316	0.447	0.548

(In each position the upper figure relates to a^* and the lower (italic) figure to b^*.)

and μ and $N-\mu$ are both large, but deteriorate as ϕ increases and as μ approaches the extremes, 0 and N.

4.5 Remarks

The calculations of §4.4.5 show that the 'approximate likelihood' functions $a(y;\mu,\phi)$ and $b(y;\mu,\phi)$ do not necessarily correspond, even approximately, to a distribution having the required mean, μ, and variance, $\phi V(\mu)$. In particular, the models

$$Y_i \sim a^*(y_i;\mu_i(\beta),\phi_i(\gamma)) \qquad (i=1,\ldots,n)$$

and

$$Y_i \sim b^*(y_i;\mu_i(\beta),\phi_i(\gamma)) \qquad (i=1,\ldots,n),$$

although close to one another, may be very different from the mean and variance specification {(4.5),(4.6)}. The practical implications of this are not clear, though, since when $a(y;\mu,\phi)$, or equivalently $b(y;\mu,\phi)$, is used as an approximate likelihood the normalizing factor is ignored.

The remarks of §4.4.4, on the other hand, are more readily interpreted: use of $a(y;\mu,\phi)$, or equivalently $b(y;\mu,\phi)$, as an approximate likelihood seems more appropriate to a model specification of the form

$$E(Y_i) = \mu_i(\beta) \qquad (i=1,\ldots,n)$$

and

$$E[D(Y_i;\mu_i)] = \phi_i(\gamma) \qquad (i=1,\ldots,n), \quad \text{as in (4.13)},$$

than to a model specified in the manner of {(4.5),(4.6)}. As shown in §4.4.4, sometimes the two specifications (4.6) and (4.13) are approximately the same. In fact there exist at least two particular families of distributions, namely the normal and inverse Gaussian examples (i) and (v) of §4.4.3, that allow both specifications to be met exactly. In general, though, as illustrated in §4.4.4, the two specifications cannot be considered equivalent, or even approximately so. An important question that arises,

then, is whether specification of a model for the expected deviance, which apparently has theoretical advantages, can reasonably be made in practice. The alternative, a model for the ratio of the variance to some known function of the mean, seems more appealing because of the familiarity of the ideas involved. However it is not clear which, if either, of the two types of specification is the more 'natural'.

APPENDIX 1

Details of the calculation leading to approximation (2.9)

The source of the error in expression (20) of Cox & Hinkley (1968) is not clear. Their (19) contains a misplaced bracket: at the beginning of the third line, ' $+\frac{1}{720}\{\gamma_4$ ' should read ' $+\{\frac{1}{720}\gamma_4$ ', but this could be a printer's error. The major part of the calculation, after expanding the logarithm of the Edgeworth series and differentiating twice, is to evaluate expectations of Hermite polynomials and their products. Those expectations relevant to the order of (2.9) are given here, with just H_r standing for $H_r(\epsilon)$:

$E(H_3) = \gamma_1;$

$E(H_4) = \gamma_2;$

$E(H_1H_3) = \gamma_2;$

$E(H_2H_3) = 6\gamma_1 + \gamma_3;$

$E(H_2^2) = 2 + \gamma_2;$

$E(H_3H_4) = 36\gamma_1 + 35\gamma_2\gamma_1 + 12\gamma_3 + o(N^{-2});$

$E(H_3^2) = 6 + 10\gamma_1^2 + 9\gamma_2 + \gamma_4;$

$E(H_3H_5) = 150\gamma_1^2 + 60\gamma_2 + 35\gamma_2^2 + 56\gamma_3\gamma_1 + 15\gamma_4 + o(N^{-2});$

$E(\mathrm{H}_1H_4) = 4\gamma_1 + \gamma_3;$

$E(H_2H_4) = 10\gamma_1^2 + 8\gamma_2 + \gamma_4;$

$E(H_1H_6) = 35\gamma_2\gamma_1 + 6\gamma_3 + o(N^{-2});$

$E(H_4^2) = 24 + o(1);$

$E(H_5^2) = 120 + o(1);$

$E(H_2H_5) = 20\gamma_1 + 35\gamma_2\gamma_1 + 10\gamma_3 + o(N^{-2});$

$E(H_1H_3^2) = 54\gamma_1 + 35\gamma_2\gamma_1 + 15\gamma_3 + o(N^{-2});$

$E(H_2H_3^2) = 36 + o(1);$

$E(H_3^2H_4) = 216 + o(1);$

$E(H_1H_3H_4) = 24 + o(1);$

$E(H_2^2H_3) = 62\gamma_1 + o(N^{-1/2})$;

$E(H_2H_3H_5) = 120 + o(1)$;

$E(H_1H_3^3) = 324 + o(1)$;

$E(H_2^2H_3^2) = 372 + o(1)$;

$E(H_2^2H_4) = 24 + o(1)$;

$E(H_1H_3H_6) = E(H_2^2H_6) = o(1)$.

Terms are collected to give (2.9); to $O(N^{-1})$ only one term, that involving $E(H_2^2)$, is required.

APPENDIX 2

A property of a family of unbiased estimating equations, and a connection with parameter orthogonality

Cox & Reid (1987, §2.2) consider some consequences of orthogonality,

$$E_{\psi,\lambda}\{\partial^2 l(\psi,\lambda)/\partial\psi\partial\lambda\} = 0 \qquad \text{for all } \psi,\lambda\ , \tag{A2.1}$$

in a regular parametric family of distributions $\{f_Y(y;\psi,\lambda)\}$ for a $n\times 1$ vector Y, where $l(\psi,\lambda)=\log f_Y(y;\psi,\lambda)$; here '$E_{\psi,\lambda}$' means expectation with respect to the density $f_Y(y;\psi,\lambda)$. Particularly important is their property (iv), stated as ' $\hat\psi_\lambda$, the maximum likelihood estimate of ψ when λ is given, varies only slowly with λ.' Specifically it is shown that $\hat\psi_\lambda-\hat\psi=O_p(n^{-1})$, where $\hat\psi$ is the unrestricted maximum likelihood estimate; the result is 'local' in that if λ here is fixed it must be the true value, otherwise it must be within $O(n^{-\frac12})$ of the true value. The proof given by Cox & Reid proceeds via an expansion of $l(\psi,\lambda)$ near $(\hat\psi,\hat\lambda)$; a slightly more direct approach, based on an expansion of the *score function* $u(\psi,\lambda)=\partial l(\psi,\lambda)/\partial\psi$, has applications also when the likelihood function is not available.

Consider, then, the local linearization

$$u(\psi,\lambda) = u(\hat\psi,\hat\lambda) + (\psi-\hat\psi)\left.\frac{\partial u}{\partial\psi}\right|_{(\hat\psi,\hat\lambda)} + (\lambda-\hat\lambda)\left.\frac{\partial u}{\partial\lambda}\right|_{(\hat\psi,\hat\lambda)} + O_p(\|\theta-\hat\theta\|^2)$$

where $\theta=(\psi,\lambda)$, etc.; substituting $\psi=\hat\psi_\lambda$ gives

$$(\hat\psi_\lambda-\hat\psi)\left.\frac{\partial u}{\partial\psi}\right|_{(\hat\psi,\hat\lambda)} = (\hat\lambda-\lambda)\left.\frac{\partial u}{\partial\lambda}\right|_{(\hat\psi,\hat\lambda)} + O_p(\|\theta-\hat\theta\|^2)\ , \tag{A2.2}$$

and the rest of the argument is then much as in Cox & Reid (1987). Under orthogonality, $(\partial u/\partial\lambda)\big|_{(\hat\psi,\hat\lambda)}=O_p(n^{\frac12})$ as $n\to\infty$. Quite generally $(\hat\lambda-\lambda)=O_p(n^{-\frac12})$, $(\hat\psi_\lambda-\hat\psi)=O_p(n^{-\frac12})$ and the remainder term in (A2.2) is $O_p(1)$; also $\partial u/\partial\psi$ is typically a sum of n terms of non-zero expectation, so

$n\{(\partial u/\partial\psi)\big|_{(\hat\psi,\hat\lambda)}\}^{-1}=O_p(1)$. Hence $\hat\psi_\lambda-\hat\psi=O_p(n^{-1})$.

The result may be extended in two stages. First it may be made less 'local' by restricting attention to likelihoods that satisfy

$$E_{\psi,\xi}\{\partial^2 l(\psi,\lambda)/\partial\psi\partial\lambda\} = 0 \quad \textit{for all } \psi,\ \xi,\ \lambda,$$

which is a much stronger condition than the orthogonality (A2.1); it implies, in particular, that

$$E_{\psi,\xi}\{\partial l(\psi,\lambda)/\partial\psi\} = 0 \quad \textit{for all } \psi,\ \xi,\ \lambda,$$

i.e. the score equation $\partial l(\psi,\lambda)/\partial\psi=0$ is an unbiased estimating equation for ψ at every λ. A simple example is the bivariate normal family

$$\begin{bmatrix}y_1\\y_2\end{bmatrix} \sim N_2\left[\begin{bmatrix}\psi\\0\end{bmatrix},\begin{bmatrix}1 & \xi\\ \xi & 1\end{bmatrix}\right]$$

with $$f_Y(y;\psi,\xi) = (2\pi)^{-1}(1-\xi^2)^{-\frac12}\exp[-\tfrac12\{(y_1-\psi)^2-2\xi(y_1-\psi)y_2+y_2^2\}/(1-\xi^2)]\ ;$$

for any given $\lambda\in(-1,1)$, the score equation

$$\partial l(\psi,\lambda)/\partial\psi = \{(y_1-\psi)-\lambda y_2\}/(1-\lambda^2) = 0$$

is an unbiased estimating equation for ψ, regardless of the true value of ξ.

Now take λ to be an arbitrary value, rather than the true value as before; and suppose that $\hat\lambda$, rather than being the maximum likelihood estimate, is such that $\hat\lambda-\lambda=O_p(n^{-\frac12})$. Then, with $\hat\psi=\hat\psi_{\hat\lambda}$, the behaviour of all quantities in (A2.2) is as before, and in particular $\hat\psi_\lambda-\hat\psi=O_p(n^{-1})$.

An immediate further extension is to the situation where $\{u(\psi,\lambda) : \lambda\in\mathbb{R}\}$ is a more general family of estimating functions for ψ, not necessarily likelihood-based score functions; the required property is still

$$E_\psi\{u(\psi,\lambda)\} = 0 \qquad \textit{for all } \psi,\ \lambda,$$

i.e. $u(\psi,\lambda)=0$ is an unbiased estimating equation for ψ at every value of λ.

The result $\hat{\psi}_\lambda - \hat{\psi} = O_p(n^{-1})$ implies in particular that the asymptotic (normal) distribution of a solution based on any fixed value λ is the same as that of a solution based on a data-dependent value $\hat{\lambda}$, provided $\hat{\lambda} - \lambda = O_p(n^{-\frac{1}{2}})$.

For a specific example, consider the family of estimating equations (2.17) from §2.2.2: the asymptotically optimal value $\lambda^0 = (2+\gamma_2)/\{2+\gamma_2-(\gamma_1/\sqrt{\phi})\}$ is unknown under only second-moment assumptions but may be $\sqrt{n}$-consistently estimated, by $\hat{\lambda}$ say, from sample third and fourth moments. Thus solution of (2.17) with $\lambda = \hat{\lambda}$ yields an estimate for μ, $\hat{\mu}_{\hat{\lambda}}$ say, that has the same first order asymptotic efficiency as would $\hat{\mu}_{\lambda^0}$.

Similar arguments hold throughout when ψ and λ are vectors, and application to the other families of 'refined' estimating equations of §§2.2.2, 2.3.2 is straightforward.

APPENDIX 3

Details of the calculation leading to approximation (2.29)

The second derivative of the logarithm of (2.28) is

$$\partial^2 l/\partial\mu^2 = \partial^2 \log f_G/\partial\mu^2 + \frac{(1+\Sigma c_r L_r)\{\partial^2(\Sigma c_r L_r)/\partial\mu^2\} - \{\partial(\Sigma c_r L_r)/\partial\mu\}^2}{(1+\Sigma c_r L_r)^2}$$

where here, and from now on, we write just L_r for $L_r^{(\alpha)}(z)$. The numerator of the second term here has, for $r,s \geq 3$ $(r \neq s)$,

coefficient of c_r: $\{r(r+1)L_r - 2r(r+\alpha)L_{r-1} + (r+\alpha-1)(r+\alpha)L_{r-2}\}$

coefficient of c_r^2: $\{rL_r^2 + (r+\alpha-1)(r+\alpha)L_{r-2}L_r - (r+\alpha)^2 L_{r-1}^2\}$

coefficient of $c_r c_s$: $[\{r(r+1)+s(s+1)-2rs\}L_r L_s + 2(s-r)(r+\alpha)L_{r-1}L_s$
$+2(r-s)(s+\alpha)L_{s-1}L_r + (r+\alpha-1)(r+\alpha)L_{r-2}L_s + (s+\alpha-1)(s+\alpha)L_{s-2}L_r$
$-2(r+\alpha)(s+\alpha)L_{r-1}L_{s-1}]$

The well known differential relation between generalized Laguerre polynomials (Abramowitz & Stegun, 1965, p783) proves useful here. This numerator must be multiplied by

$$(1+\Sigma c_r L_r)^{-2} = 1 - 2\Sigma c_r L_r + 3(\Sigma c_r L_r)^2 - 4(\Sigma c_r L_r)^3 + \ldots$$

The rest of the calculation consists mostly of evaluation of expectations of the form

$$E(L_{r_1} L_{r_2} \ldots L_{r_k}) \quad . \tag{A3.1}$$

One helpful fact, when discounting terms of high order in ϕ, is that

$$E(L_{r_1} L_{r_2} \ldots L_{r_k}) = o(\phi^{-(r_1+r_2+\ldots+r_k+1)}) \ .$$

Also, on account of the fact that $E(L_1^r) = (-1)^r \mu_r$, where μ_r is the rth central moment of z, the following expressions prove very useful:

$L_2 = \frac{1}{2}\{L_1^2 + 2L_1 - (\alpha+1)\}$;

$L_3 = \frac{1}{6}\{L_1^3 + 6L_1^2 + 3(1-\alpha)L_1 - 4(1+\alpha)\}$;

$$L_4 = \tfrac{1}{24}\{L_1^4 + 12L_1^3 + (30-6\alpha)L_1^2 - (4+28\alpha)L_1 - (15-3\alpha)(\alpha+1)\};$$

$$L_5 = \tfrac{1}{120}\{L_1^5 + 20L_1^4 + (110-10\alpha)L_1^3 + (140-100\alpha)L_1^2 - (95+200\alpha-15\alpha^2)L_1 - (56-40\alpha)(\alpha+1)\};$$

$$L_6 = \tfrac{1}{720}\{L_1^6 + 30L_1^5 + (285-15\alpha)L_1^4 + (940-260\alpha)L_1^3 + (555-1200\alpha+45\alpha^2)L_1^2 - (906+1296\alpha-330\alpha^2)L_1 - (185-400\alpha+15\alpha^2)(\alpha+1)\};$$

$$L_7 = \tfrac{1}{71}\{L_1^7 + 42L_1^6 + (609-21\alpha)L_1^5 + (3640-560\alpha)L_1^4 + (7875-4620\alpha+105\alpha^2)L_1^3 + (714-12936\alpha+1470\alpha^2)L_1^2 - (7637+7567\alpha-5005\alpha^2+105\alpha^3)L_1 - (204-3696\alpha+420\alpha^2)(\alpha+1)\} .$$

In calculating $E(\partial^2 l/\partial\mu^2)$ to $O(1)$ only one of the expectations (A3.1) makes a contribution, namely

$$E(L_2^2) = \tfrac{1}{12}\phi^{-2}\{6 + 3\phi(\tau_4-4\tau_3+2) + O(\phi^2)\} .$$

To $O(\phi)$, the following make a contribution:

$$E(L_3) = \tfrac{1}{6}c_3(\alpha+1)(\alpha+2)(\alpha+3)$$

$$E(L_4) = \tfrac{1}{24}c_4(\alpha+1)(\alpha+2)(\alpha+3)(\alpha+4)$$

$$E(L_1L_3) = \tfrac{1}{12}\phi^{-1}\{2\tau_4-12\tau_3+O(\phi)\}$$

$$E(L_1L_4) = \tfrac{1}{12}\phi^{-2}\{-2\tau_3+O(\phi)\}$$

$$E(L_2L_3) = \tfrac{1}{12}\phi^{-2}\{-6\tau_3+O(\phi)\}$$

$$E(L_3^2) = \tfrac{1}{12}\phi^{-3}\{2+O(\phi)\}$$

$$E(L_3L_4) = \tfrac{1}{12}\phi^{-3}\{-3\tau_3+O(\phi)\}$$

$$E(L_4^2) = \tfrac{1}{24}\phi^{-4}\{1+O(\phi)\}$$

$$E(L_2L_5) = \tfrac{1}{12}\phi^{-3}\{-\tau_3+O(\phi)\}$$

$$E(L_5^2) = \tfrac{1}{120}\phi^{-5}\{1+O(\phi)\}$$

$$E(L_1L_3^2) = \tfrac{1}{12}\phi^{-3}\{-18\tau_3-12+O(\phi)\}$$

$$E(L_2^2L_3) = \tfrac{1}{12}\phi^{-3}\{-31\tau_3-24+O(\phi)\}$$

$$E(L_2L_3^2) = \tfrac{1}{12}\phi^{-4}\{6+O(\phi)\}$$

$$E(L_2^2L_4) = \tfrac{1}{12}\phi^{-4}\{3+O(\phi)\}$$

$$E(L_2^2L_3^2) = \tfrac{1}{12}\phi^{-5}\{31+O(\phi)\}$$

$$E(L_3^2L_4) = \tfrac{1}{12}\phi^{-5}\{3+O(\phi)\}$$

$$E(L_1L_3^3) = \tfrac{1}{12}\phi^{-5}\{18+O(\phi)\}$$

$$E(L_1L_3L_4) = \tfrac{1}{12}\phi^{-4}\{2+O(\phi)\}$$

$$E(L_2L_3L_5) = \tfrac{1}{12}\phi^{-5}\{1+O(\phi)\} .$$

The approximation (2.29) follows, at length, by collecting terms; an interesting feature is that terms involving c_5 all cancel out.

APPENDIX 4

Details of the calculation leading to approximation (2.36)

The coefficients $\{c_r\}$ of (2.35) are given here in terms of the cumulants of the mixing distribution, along with their order of magnitude as $\psi \to 0$ under both (*a*) 'limiting normality' and (*b*) 'constant shape':

		(*a*)	(*b*)
c_2	$\frac{1}{2}\psi$	$O(\psi)$	$O(\psi)$
c_3	$-\kappa_3$	$O(\psi^2)$	$O(\psi^{3/2})$
c_4	$\kappa_4 + 3\kappa_2^2$	$O(\psi^2)$	$O(\psi^2)$
c_5	$-\kappa_5 - 10\kappa_3\kappa_2$	$O(\psi^3)$	$O(\psi^{5/2})$
c_6	$\kappa_6 + 15\kappa_4\kappa_2 + 10\kappa_3^2 + 15\kappa_2^3$	$O(\psi^3)$	$O(\psi^3)$

etc. Now $E(\partial^2 l/\partial\mu^2)$ is as in Appendix 3; on taking expectations, terms involving c_5 and higher are $o(\psi^3)$ and, to the order of (2.36), only the following expectations are needed:

$$\left.\begin{aligned} E(L_2) &= \tfrac{1}{2}\psi\,; \\ E(L_2^2) &= 1 + 5\psi + o(\psi)\,; \end{aligned}\right\} \text{ only these two contribute to the } \psi^2 \text{ term}$$

$$E(L_1^2 L_2) = 2 + o(1);$$

$$E(L_2^2 L_1) = -4 + o(1);$$

$$E(L_2^3) = 10 + o(1);$$

$$E(L_1^2) = 1 + \psi + o(\psi);$$

$$E(L_1 L_2) = -2\psi + o(\psi)\ .$$

On collecting terms, those involving c_4 cancel out to this order and (2.36) results.

APPENDIX 5

Details of the calculation leading to approximation (2.41)

To the order given by (2.41), terms in (2.40) involving c_3 and higher may be disregarded. If we write

$$l^*(y;\mu,\phi) = \log f_0(y\,|\,\mu) + \log\{1 + c_2 P_2(y,\mu)/V(\mu)\}$$

then, using the differential relation between the polynomials $\{P_r\}$ (Morris, 1982), it is easily shown that

$$\partial^2 l^*/\partial\mu^2 = \partial^2 \log f_0/\partial\mu^2 + c_2[(1+c_2P_2/V)\{V^2(2(1+c)V - V''P_2) + 2VV'(2(1+c)VP_1 + V'P_2)\}$$
$$- c_2\{2(1+c)VP_1 + V'P_2\}^2] / [V^2(1+c_2P_2/V)]^2$$

where we have written V, V' and P_r for $V(\mu)$, $V'(\mu)$ and $P_r(y,\mu)$ respectively. After expanding the denominator of the second term here, only the following expectations make a contribution to terms of the required order:

$E(P_2) = (\phi - 1)V;$

$E(P_2^2) = 2(1+c)V^2 + o(1)$.

On collection of terms, (2.41) follows easily.

REFERENCES

ABRAMOWITZ, M. & STEGUN, I.A. (1965), eds. *Handbook of Mathematical Functions* , National Bureau of Standards, U.S. Government Printing Office, Washington, D.C.

ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H., & TUKEY, J.W. (1972). *Robust estimates of location: survey and advances.* Princeton, N.J.: Princeton University Press.

CHESHER, A.D. & JEWITT, I. (1984). Finite sample properties of least squares covariance matrix estimators. Discussion Paper 84/163, Dept. Economics, Bristol University.

CHESHER, A.D. & JEWITT, I. (1986). The bias of the heteroskedasticity consistent covariance matrix estimator. Discussion Paper 86/176, Dept. Economics, Bristol University. To appear in *Econometrica.*

COX, D.R. (1975). A note on partially Bayes inference and the linear model. *Biometrika* **62** , 651-54.

COX, D.R. (1983). Some remarks on overdispersion. *Biometrika* **70** , 269-74.

COX, D.R. & HINKLEY, D.V. (1968). A note on the efficiency of least-squares estimates. *J. R. Statist. Soc.* B **30** , 284-89.

COX, D.R. & OAKES, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall

COX, D.R. & REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J.R. Statist. Soc.* B , to appear.

CROWDER, M.J. (1982). On weighted least-squares and some variants. Technical Report No. 13, Dept. Mathematics, Surrey University.

CROWDER, M.J. (1986a). On linear and quadratic estimating functions. Unpublished report, Surrey University.

CROWDER, M.J. (1986b). On consistency and inconsistency of estimating equations. *Econometric Theory* 3 (to appear).

DIACONIS, P. & EFRON, B. (1985). Testing for independence in a two-way table: new interpretations of the chi-square statistic (with discussion). *Ann. Statist.* 13, 845-913.

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* 3 , 1189-1242.

EFRON, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* 81, 709-21.

EICKER, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann. Math. Statist.* 34 , 447-56.

GODAMBE, V.P. & THOMPSON, M.E. (1978). Some aspects of the theory of estimating equations. *J. Statist. Plan. Infer.* 2 , 95-104.

GOURIEROUX, C., MONFORT, A & TROGNON, A. (1984a). Pseudo maximum likelihood methods: theory. *Econometrica* 52 , 681-700.

GOURIEROUX, C., MONFORT, A & TROGNON, A. (1984b). Pseudo maximum likelihood methods: application to Poisson models. *Econometrica* 52, 701-20.

GREEN, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J.R. Statist. Soc.* B 46 , 149-192.

HARTIGAN, J.A. (1969). Linear Bayesian methods. *J.R. Statist. Soc.* B **31** , 446-54.

HINKLEY, D.V. (1977). Jackknifing in unbalanced situations. *Technometrics* **19** , 285-292.

INAGAKI, N. (1973). Asymptotic relations between the likelihood estimating functions and the maximum likelihood estimator. *Ann. Inst. Statist. Math.* **25**, 1-26.

JORGENSEN, B. (1987). Exponential dispersion models. *J. R. Statist. Soc.* B (to appear).

MACKINNON, J.G. & WHITE, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* **29** , 305-25.

MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11** , 59-67.

MCCULLAGH, P. (1984). Generalized linear models. *Eur. J. Op. Res.* **16** , 285-292.

MCCULLAGH, P. & NELDER, J.A. (1983). *Generalized Linear Models.* London: Chapman and Hall.

MORRIS, C.N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10** , 65-80.

MORTON, R. (1981). Efficiency of estimating equations and the use of pivots. *Biometrika* **68** , 227-233.

MOUCHART, M. & SIMAR, L. (1982). Theory and applications of least squares approximation in Bayesian analysis. In *Specifying Statistical Models* , ed. J.P. Florens, M. Mouchart, J.P. Raoult, L. Simar & A.F.M. Smith : Springer-Verlag Lecture Notes in Statistics, no. 16.

NELDER, J.A. & PREGIBON, D. (1983). Quasi-likelihood models and data analysis. Technical report, AT&T Bell Laboratories, Murray Hill NJ.

NELDER, J.A. & WEDDERBURN, R.W.M. (1972). Generalized linear models. *J.R. Statist. Soc.* A **135** , 370-84.

PREGIBON, D. (1983). An alternative covariance estimated for generalised linear models. *GLIM Newsletter* No. 6 , 51-55.

PREGIBON, D. (1984). Review of *Generalized Linear Models*, by P.McCullagh & J.A.Nelder. *Ann. Statist.* **12**, 1589-96.

ROYALL, R.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *Int. Statist. Review* **54**, 221-26.

SMITH, R.H. (1888). True average of observations? *Nature* **37** , 464.

STIGLER, S.M. (1980). R.H. Smith, a Victorian interested in robustness. *Biometrika* **67** , 217-21.

TWEEDIE, M.C.K. (1984). An index which distinguishes between some important exponential families. In *Proceedings of the Indian Statistical Institute Golden Jubilee Conference on Statistics: Applications and New Directions*, eds. J.K.Ghosh & J.Roy, pp579-604. Calcutta: Indian Statistical Institute.

WEDDERBURN, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61** , 439-47.

WEDDERBURN, R.W.M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63** , 27-32.

WEST, M. (1985). Generalized linear models: scale parameters, outlier accommodation and prior distributions. *Bayesian Statistics 2*, 531-58. Amsterdam: North-Holland.

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** , 817-38.

WHITTLE, P. (1961). Gaussian estimation in stationary time series. *Bull. Int. Stat. Inst.* **39** , 1-26.