The Annals of Statistics
1983, Vol. 11, No. 1, 59–67

QUASI-LIKELIHOOD FUNCTIONS

BY PETER MCCULLAGH

Imperial College, London and University of British Columbia

The connection between quasi-likelihood functions, exponential family models and nonlinear weighted least squares is examined. Consistency and asymptotic normality of the parameter estimates are discussed under second moment assumptions. The parameter estimates are shown to satisfy a property of asymptotic optimality similar in spirit to, but more general than, the corresponding optimal property of Gauss-Markov estimators.

1. Introduction. It is well known that if the likelihood function has the exponential family form, maximum likelihood estimates of regression parameters can often be found using the method of weighted least squares (Nelder and Wedderburn, 1972; Bradley, 1973; Wedderburn, 1974; and Jennrich and Moore, 1975). Here we use the term "weighted least squares" in a fairly general sense: typically the computations involve nonlinear response functions and weights that vary from one iteration to the next. In particular, when the variance is assumed constant the quantity to be minimized is a sum of squared residuals, and the asymptotic results of Jennrich (1969) and Wu (1981) apply. More generally, when the variances are not constant, the estimating equations can be thought of as a generalization of the scoring method used in probit analysis and usually attributed to Fisher.

Modified and conditional likelihoods sometimes have the required exponential form. Thus the method of weighted least squares may be used for partial likelihoods, Cox (1972, 1975), and for conditional likelihoods of the kind that arise in the consideration of doubly conditioned 2×2 or larger contingency tables. In fact the method of weighted least squares can be used to find maximum likelihood estimates even in some cases where the likelihood function does not have the exponential family form (Jorgensen, 1983).

The present paper has two purposes. First we describe a wider class of problems for which the method of weighted least squares may be used to maximize the likelihood function. Secondly, the method of weighted least squares may be used under second moment assumptions thus avoiding the complete specification of the underlying distribution. Following Wedderburn (1974) we refer to the function being maximized as a quasi-likelihood. Here we examine the asymptotic properties of the quasi-likelihood function and we show that the estimators enjoy a certain asymptotic optimality property. There are close connections to the Gauss-Markov theorem and to the asymptotic optimality of maximum likelihood estimators.

2. A class of likelihood functions. Let the observed vector of random variables $\mathbf{Y}$ be of length N and suppose that the log likelihood considered initially as a function of the N-dimensional parameter $\boldsymbol{\theta}$ and an additional parameter σ^2 may be written in the form

$$\sigma^{-2}\{\mathbf{y}^T\boldsymbol{\theta} - b(\boldsymbol{\theta}) - c(\mathbf{y}, \sigma)\} \tag{1}$$

for suitably chosen functions $b(\boldsymbol{\theta})$ and $c(\mathbf{y}, \sigma)$. Differentiating (1) and assuming that the support of the distribution does not depend on $\boldsymbol{\theta}$, we find

$$E(\mathbf{Y}) = \boldsymbol{\mu} = b'(\boldsymbol{\theta}) \quad \text{and} \quad \operatorname{Cov}(\mathbf{Y}) = \sigma^2 b''(\boldsymbol{\theta}) = \sigma^2 \mathbf{V}(\boldsymbol{\mu}), \quad \text{say.} \tag{2}$$

Received October 1981; revised September 1982.
AMS 1980 *subject classifications.* Primary 62J02; secondary 62F12.
Key words and phrases. Exponential family, quasi-likelihood, second moment assumption, weighted least squares.

In fact, the rth order cumulants of $\mathbf{Y}$ are given by $\kappa_r = \sigma^{2r-2}b^{(r)}(\theta)$ so that there is a close connection between (1) and convolutions of natural exponential families (Morris, 1982) where σ^{-2} plays the role of a sample size.

The expression for the cumulants of $\mathbf{Y}$, and in particular equations (2) for the first two cumulants, describe the random component of the model. However, in applications it is usually the systematic or nonrandom variation that is of primary importance. Often the systematic component may be expressed in terms of a regression equation

$$E(\mathbf{Y}) = \mu = \mu(\beta), \tag{3}$$

where β is a p-dimensional parameter vector and $\mu(\cdot)$ is a vector of known functions having bounded third derivatives. In many applications the regression function $\mu(\cdot)$ has a linear component involving a model matrix $\mathbf{X}$ describing the experimental or observational conditions under which the observations were made. In the present paper we assume only identifiability in the sense that distinct β's imply distinct μ's. This would imply that the model matrix $\mathbf{X}$ or, more generally, the derivative matrix $\mathbf{D} = d\mu/d\beta$ have rank p for all β.

The regression equation (3) is an important but special case because it does not involve σ^2. By way of contrast we also consider regression equations that implicitly involve σ^2,

$$E\{\mathbf{h}(\mathbf{Y})\} = \psi(\beta), \tag{4}$$

where $\mathbf{h}(\cdot)$ is a known nonlinear function of the data. We could replace (4) with a more general model that involves σ^2 explicitly but this extra level of generality is not required here.

If σ^2 is known, the expression on the left of (4) is a function of μ alone and, subject to reasonable monotonicity, we can express (4) in form (3). Furthermore, for known σ^2, (1) is an exponential family with the property that the variance and all higher order cumulants of $\mathbf{Y}$ are functions of the mean vector alone. Examples include the exponential, Poisson, multinomial, noncentral hypergeometric and partial likelihoods of the type that arise in survival analysis (Cox, 1975). Weighted least squares may be used to compute the maximum likelihood estimate of β in either (3) or (4), (Bradley, 1973).

In the more general case where σ^2 is an unknown parameter, (1) is not generally an exponential family. Furthermore, it is not possible to express (4) in the form (3). However, weighted least squares may still be used to compute $\hat{\beta}$ provided that the model of interest has the form (3) not involving σ^2.

A further property of the family (1) is that μ and σ^2 are orthogonal in the sense that the mixed second derivative has zero expectation. Furthermore, under (3) but not generally under (4), β and σ^2 are also orthogonal.

Suppose again that σ^2 is unknown and that $c(\mathbf{y}, \sigma) = g(\mathbf{y}) + h(\sigma)$ corresponding to an exponential family model. I know of only three families of distributions, the normal, inverse normal and gamma, that satisfy these requirements. These three families share the remarkable property that $b'(\cdot)$ and $g'(\cdot)$ are functional inverses. This property is closely connected with the existence of a statistic, $W(Y, \mu)$, whose distribution is independent of μ but may depend on σ^2. A further important property of this exponential class is that, for fixed σ, the maximum achievable likelihood is a constant independent of $\mathbf{y}$ and occurs at $\mu = \mathbf{y}$. Furthermore, the estimate of σ^2 is a function of $-2\{\mathbf{y}^T\hat{\theta} - b(\hat{\theta}) - g(\mathbf{y})\}$ which is equivalent to $-2[\mathbf{y}^T\{\hat{\theta} - \theta(\mathbf{y})\} - b(\hat{\theta}) + b(\theta(\mathbf{y}))]$. This quantity is sometimes called the deviance and plays the role of the residual sum of squares. The second form of the expression uses only the functions $\theta(\mu)$ and $b(\theta)$ so that the expression may be used regardless of whether or not (1) is in the exponential family. In fact the support of (1) need not be independent of σ.

The generalized least squares equations for the parameters in (3) can be written

$$\mathbf{D}^T\mathbf{V}^-\{\mathbf{y} - \mu(\hat{\beta})\} = 0 \tag{5}$$

where $\mathbf{D} = d\mu/d\beta$ is of order $N \times p$ and $\mathbf{V}^-$ is a generalized inverse of $\mathbf{V}$. We refer to (5)

as least squares equations primarily because of the geometrical interpretation which involves successive projections of the residual vector $\mathbf{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_0)$ on to the tangent space of the solution locus $\boldsymbol{\mu}(\boldsymbol{\beta})$. Here $\hat{\boldsymbol{\beta}}_0$ is the current estimate of $\boldsymbol{\beta}$. These equations do not depend on σ^2 so that the numerical value of $\hat{\boldsymbol{\beta}}$ is the same whether σ^2 is known or not. The Newton-Raphson method with the second derivative matrix replaced by its expected value, $\mathbf{D}^T\mathbf{V}^{-}\mathbf{D}$, results in an adjustment to $\hat{\boldsymbol{\beta}}_0$ given by

$$\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0 = (\mathbf{D}^T\mathbf{V}^{-}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{V}^{-}(\mathbf{y} - \hat{\boldsymbol{\mu}}_0),$$

where the quantities on the right are computed at $\hat{\boldsymbol{\beta}}_0$.

In the following section the natural order of the assumptions is reversed. Instead of taking the log likelihood to be of the form (1) from which the moments may be derived, we begin with the moments and attempt to reconstruct (1). The reconstituted function is called a quasi-likelihood and its asymptotic properties are investigated.

3. Quasi-likelihood functions. In this and in the following sections the only assumptions on the distribution of the data are those concerning first and second moments and some additional regularity conditions relating to the regression equation (3).

Given the vector of random variables $\mathbf{Y}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\mathbf{V}(\boldsymbol{\mu})$ the log quasi-likelihood, considered as a function of $\boldsymbol{\mu}$, is given by the system of partial differential equations

$$\frac{\partial \ell(\boldsymbol{\mu}; \mathbf{y})}{\partial \boldsymbol{\mu}} = \mathbf{V}^{-}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}) \tag{6}$$

which extends Wedderburn's (1974) definition. For many purposes it is not necessary to solve for $\ell(\boldsymbol{\mu}; \mathbf{y})$. When an explicit solution is required, it can often be found by constructing functions $\boldsymbol{\theta}(\boldsymbol{\mu})$, $b(\boldsymbol{\theta})$ satisfying (2) and writing

$$\ell(\boldsymbol{\mu}; \mathbf{y}) = \mathbf{y}^T\boldsymbol{\theta} - b(\boldsymbol{\theta}) - c(\mathbf{y}, \sigma)$$

where $c(\mathbf{y}, \sigma)$ is entirely arbitrary. Explicit solutions are discussed in Section 6; in any case, the existence of a family of distributions for which $\ell(\cdot\,, \cdot)$ is the log likelihood is not required.

Provided only that the systematic part of the model has the form (3) not involving the unknown σ^2, the generalized least squares equations for $\hat{\boldsymbol{\beta}}$ are given by (5). There is, of course, no guarantee that $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator because the correct likelihood function is generally different from $\ell(\boldsymbol{\mu}; \mathbf{y})$. In particular the higher order cumulants of $\mathbf{Y}$ may not have the required multiplicative form.

We now proceed to outline the properties of quasi-likelihood estimates. It is important to emphasize that the properties of quasi-likelihood estimates are not general properties for all weighted least squares estimates. For example, weighted least squares can be used for parameter estimation when $\operatorname{cov}(\mathbf{Y}) = \mathbf{V}(\sigma^2, \boldsymbol{\rho})$, where $\boldsymbol{\rho}$ is a vector of autoregressive coefficients. Quasi-likelihoods are defined only when the covariance matrix has the simpler multiplicative form $\sigma^2 V(\boldsymbol{\mu})$.

Morton (1981) considers a more general problem involving several nuisance parameters. His estimating equations have the form of (4) where the quantity $\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})$ is replaced by a vector of pivots whose distribution is independent of the nuisance parameters. In the problems considered here where there is a single nuisance parameter σ^2, it is not necessary to construct pivots and in fact the distribution of $\mathbf{Y} - \boldsymbol{\mu}$ or $\mathbf{V}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})/\sigma$ will generally depend on σ^2.

4. Properties of quasi-likelihood functions. The statistical properties of quasi-likelihood functions are very similar to those of ordinary likelihoods except that the nuisance parameter, σ^2, when it is unknown, is treated separately from $\boldsymbol{\beta}$ and is not estimated by weighted least squares. The principal results fall conveniently into three classes; those concerning the score function $\mathbf{U}_\beta = \partial\ell/\partial\boldsymbol{\beta}$, those concerning the estimator $\hat{\boldsymbol{\beta}}$ and those concerning the distribution of the quasi-likelihood-ratio statistic.

The score function, $\mathbf{U}_\beta = \mathbf{u}_\beta(\boldsymbol{\beta}; \mathbf{Y}) = \mathbf{D}^T\mathbf{V}^-(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))$ has zero mean and covariance matrix $\sigma^2\mathbf{i}_\beta = \sigma^2\mathbf{D}^T\mathbf{V}^-\mathbf{D}$ where $-\mathbf{i}_\beta$ is the expected second derivative matrix of $\ell(\boldsymbol{\mu}(\boldsymbol{\beta}); \mathbf{Y})$. When there is no ambiguity we write $\ell(\boldsymbol{\beta}; \mathbf{Y})$ instead of $\ell(\boldsymbol{\mu}(\boldsymbol{\beta}); \mathbf{Y})$.

Under weak conditions on the third derivative of (3) and assuming that $N^{-1}\mathbf{i}_\beta$ has a positive definite limit and that the third moments of $\mathbf{Y}$ are finite, the following asymptotic results apply:

$$N^{-1/2}\mathbf{U}_\beta \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{i}_\beta/N) + O_p(N^{-1/2}), \tag{7}$$

$$E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O(N^{-1}), \tag{8}$$

$$N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{N}_p(\mathbf{0}, N\sigma^2\mathbf{i}_\beta^{-1}) + O_p(N^{-1/2}). \tag{9}$$

If the third moment is infinite the error terms in (7) and (9) are $o_p(1)$. Without altering the order of the approximation in (7) and (9) the observed matrix of second derivatives, $\mathbf{I}_\beta$ can be substituted for $\mathbf{i}_\beta$. In the absence of information concerning higher moments of $\mathbf{Y}$ there is no guarantee that $\mathbf{I}_\beta$ is better than $\mathbf{i}_\beta$ as a measure of the precision of $\hat{\boldsymbol{\beta}}$.

Furthermore it can be shown that among all estimators of $\boldsymbol{\beta}$ for which the influence function is linear, i.e. estimators $\tilde{\boldsymbol{\beta}}$ satisfying

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{L}_\mu(\mathbf{Y} - \boldsymbol{\mu}) + o_p(N^{-1/2}), \tag{10}$$

where $\mathbf{L}_\mu$ is a $p \times N$ matrix of influences, quasi-likelihood estimates have minimum asymptotic variance. The reference class of linear influence estimators is a natural one here because the systematic part of the model (3) is specified in terms of the mean value parameter, $\boldsymbol{\mu}$.

Denote by $\ell(\hat{\boldsymbol{\beta}}_0; \mathbf{y})$ and $\ell(\hat{\boldsymbol{\beta}}; \mathbf{y})$ the maximized values of the log quasi-likelihood under H_0 and under H_A where the corresponding parameter spaces of dimensions q under H_0 and $p > q$ under H_A are appropriately nested. Then, under H_0,

$$2\ell(\hat{\boldsymbol{\beta}}; \mathbf{Y}) - 2\ell(\hat{\boldsymbol{\beta}}_0; \mathbf{Y}) \sim \sigma^2\chi^2_{p-q} + O_p(N^{-1/2}). \tag{11}$$

If $\ell(\boldsymbol{\beta}; \mathbf{y})$ were the correct log likelihood as opposed to an artificially constructed log quasi-likelihood and if $\mathbf{Y}$ were continuously distributed, the error term in (11) would be $O_p(N^{-1})$ (Hayakawa, 1977).

There now follows a brief outline of how the properties described above may be established. The elements of $\mathbf{U}_\beta = \mathbf{D}^T\mathbf{V}^-(\mathbf{Y} - \boldsymbol{\mu})$ involve a sum of N random variables each with zero mean and finite variance. The total variance is $\sigma^2\mathbf{i}_\beta$. If $\mathbf{i}_\beta = O(N)$ no finite set of variances is dominant and the central limit theorem implies (7). For a proof of the remaining results we summarize the order of magnitude of the various quantities involved. The first derivative $\mathbf{U}_\beta$ is $O_p(N^{1/2})$ with zero expectation: all other derivatives of $\ell(\boldsymbol{\beta}; \mathbf{Y})$ are $O_p(N)$. Thus $\mathbf{I}_\beta = -\partial^2\ell(\boldsymbol{\beta}; \mathbf{Y})/\partial\boldsymbol{\beta}^2 = O_p(N)$, $\mathbf{i}_\beta = O(N)$, and $\mathbf{I}_\beta - \mathbf{i}_\beta = O_p(N^{1/2})$ with zero expectation. The equations used for determining $\hat{\boldsymbol{\beta}}$, $\mathbf{u}_\beta(\hat{\boldsymbol{\beta}}; \mathbf{y}) = 0$ can be written as a Taylor series in $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ giving

$$0 = \mathbf{u}_\beta(\boldsymbol{\beta}; \mathbf{y}) - \mathbf{I}_{\beta^*}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{12}$$

where $\mathbf{I}_{\beta^*}$ is the observed information evaluated at a point $\boldsymbol{\beta}^*$ lying on the line segment joining $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. Thus, there exists a $\hat{\boldsymbol{\beta}}$ satisfying $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(N^{-1/2})$ and

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{I}_{\beta^*}^{-1}\mathbf{U}_\beta = \mathbf{i}_{\beta^*}^{-1}\mathbf{U}_\beta + O_p(N^{-1}) = \mathbf{i}_\beta^{-1}\mathbf{U}_\beta + O_p(N^{-1}). \tag{13}$$

Results (8) and (9) follow immediately.

To establish (11) expand $\ell(\boldsymbol{\beta}; \mathbf{Y})$ in a Taylor series about $\hat{\boldsymbol{\beta}}$ giving

$$\ell(\boldsymbol{\beta}; \mathbf{Y}) - \ell(\hat{\boldsymbol{\beta}}; \mathbf{Y}) = -(\tfrac{1}{2})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\mathrm{T}\mathbf{I}_{\beta^*}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Now replace $\mathbf{I}_{\beta^*}$ by $\mathbf{i}_\beta$ and substitute (13) giving

$$2\ell(\hat{\boldsymbol{\beta}}; \mathbf{Y}) - 2\ell(\boldsymbol{\beta}; \mathbf{Y}) = \mathbf{U}\mathrm{T}_\beta\mathbf{i}_\beta^{-1}\mathbf{U}_\beta + O_p(N^{-1/2}). \tag{14}$$

Thus if $\mathbf{U}_\beta$ is asymptotically normal the quasi-likelihood-ratio statistic (14) is asymptotically $\sigma^2\chi^2_p$. For composite null hypotheses (14) can be expressed as the difference of two quadratic forms in $\mathbf{U}_\beta$, one of rank p and the other under H_0 of rank $q < p$. This difference can in turn be expressed as a quadratic form of rank $p - q$ when terms of order $N^{-1/2}$ are ignored. Details are straightforward but tedious: see e.g. Cox and Hinkley (1974, page 323).

Proof of asymptotic optimality follows the same lines as a proof of the Gauss-Markov theorem. The essential idea is that asymptotic unbiasedness, i.e. $E(\tilde{\beta} - \beta) = o(N^{-1/2})$, implies $\mathbf{LD} - \mathbf{I} = o(1)$ where $\mathbf{L}$ is the matrix of influences in (10). Minimizing the variance of $\tilde{\beta}$ subject to this constraint gives $\mathbf{L} = (\mathbf{D}^T\mathbf{V}^-\mathbf{D})^{-1}\mathbf{D}^T\mathbf{V}^-$ when smaller order terms are ignored. This establishes the required result.

The assumptions required here are much weaker than those required for the Gauss-Markov theorem. Global linearity is replaced by the much weaker local linearity (10). Unbiasedness is replaced by the weaker requirement that the bias be $o(N^{-1/2})$: typically the bias is $O(N^{-1})$. The conclusions, while they apply more widely than the Gauss-Markov theorem, are inevitably a little weaker being asymptotic rather than exact. It would be of interest to know whether a more general form of this result applies to weighted least squares in the more general case where the covariance matrix does not have the form $\sigma^2\mathbf{V}(\boldsymbol{\mu})$.

In small samples uniqueness of the maximum cannot be guaranteed except in some special but important cases, Wedderburn (1976), Haberman (1977), Burridge (1981), Pratt (1981). However, in large samples, $\mathbf{i}_\beta$ is $O(N)$ and positive definite for all $\boldsymbol{\beta}$. Also, in any $o(1)$ neighborhood of the true $\boldsymbol{\beta}$, $\mathbf{I}_\beta - \mathbf{i}_\beta = o(N)$. Within this neighborhood of consistent estimators, $\mathbf{I}_\beta$ is positive definite with high probability, implying that there is a single maximum corresponding to a consistent estimator. This analysis assumes that the model is correct and suggests that if the initial estimate of $\boldsymbol{\beta}$ is consistent the iterative equations (4) will converge to the correct maximum with high probability for large N. It is possible that there might also be local maxima, but these local maxima would correspond to inconsistent estimators.

5. Estimation of σ^2. In some applications σ^2 may be known: often $\sigma^2 = 1$. More often σ^2 must be estimated in order to make use of the approximate results (7), (8), (9) and (11). In the absence of information beyond second moments there is little alternative to using

$$\tilde{\sigma}^2 = (\mathbf{y} - \hat{\boldsymbol{\mu}})^T\mathbf{V}^-(\mathbf{y} - \hat{\boldsymbol{\mu}})/(N - p) = X^2/(N - p) \tag{15}$$

where X^2 is a generalized form of Pearson's statistic. The order of magnitude of the error terms in (7), (9) and (11) is unaffected by the insertion of the estimate $\tilde{\sigma}^2$ in place of σ^2.

In certain cases it may be possible to improve on (15). For example if the log likelihood (2) is in the exponential family, then the maximum likelihood estimate of σ^2 is a function of the difference between the likelihood achievable in unrestricted parameter space $\boldsymbol{\mu} \in \mathbb{R}^N$ and that achieved under the model. This difference known as the deviance may be written

$$d(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\ell(\mathbf{y}; \mathbf{y}) - 2\ell(\hat{\boldsymbol{\mu}}; \mathbf{y}).$$

An estimate of σ^2, in general different from the maximum likelihood estimate, can be obtained by equating the observed deviance $d(\mathbf{y}; \hat{\boldsymbol{\mu}})$ to its approximate expected value. This estimator has the slight advantage over $\tilde{\sigma}^2$ that it is asymptotically independent of $\hat{\boldsymbol{\beta}}$. The principal disadvantage is that the expectation of $d(\mathbf{Y}; \hat{\boldsymbol{\mu}})$ is often difficult to compute.

To take an example, suppose that $\mathbf{Y}$ has the gamma distribution with independent components so that σ is the coefficient of variation and

$$d(\mathbf{y}; \boldsymbol{\mu}) = 2\sum\{\log(y_i/\mu_i) - (y_i - \mu_i)/\mu_i\}.$$

The expected value of $d(\mathbf{Y}; \hat{\boldsymbol{\mu}})$ as a function of $\nu = \sigma^{-2}$ is

$$2N\{\psi(\nu) - \log\nu\} - p\nu^{-1} + O(N^{-1}), \tag{16}$$

where $\psi(\cdot)$ is the derivative of the log gamma function. The maximum likelihood estimate of ν is found by equating the leading term of (16) to the observed $d(\mathbf{y}; \hat{\boldsymbol{\mu}})$: it would seem preferable to reduce the bias by including the $O(1)$ term to make an allowance for the number of unknown parameters estimated. There is, unfortunately, no guarantee that the estimate of σ^2 based on equating $d(\mathbf{y}, \hat{\boldsymbol{\mu}})$ to (16) is consistent outside the gamma family.

6. Examples of quasi-likelihood functions.

EXAMPLE 1: *Least squares.* Let $\mathbf{V}(\boldsymbol{\mu}) = \mathbf{V}$, a matrix of known constants. Solving (6) we find $\ell(\boldsymbol{\beta}; \mathbf{y}) = (\frac{1}{2})(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))^T\mathbf{V}^-(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))$, a quadratic form in the residual vector. Consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}$ were proved by Jennrich (1969) and Wu (1981) under weaker conditions on the second derivative matrix than those assumed here.

EXAMPLE 2. *Uncorrelated observations.* Let $\mathbf{V}(\boldsymbol{\mu}) = \operatorname{diag}\{v(\mu_1), v(\mu_2), \cdots, v(\mu_N)\}$. Then the system of equations, (6), becomes uncoupled and the quasi-likelihood has the form $\ell(\boldsymbol{\mu}; \mathbf{y}) = \sum \ell(\mu_i; y_i)$ where the components satisfy

$$\partial\ell(\mu; y)/\partial\mu = (y - \mu)/v(\mu). \tag{17}$$

Equation (17) was given by Wedderburn (1974) as the definition of a log quasi-likelihood function. Explicit solutions for some simple functions are given in Table 1. These show that if the distribution is in the exponential family and if the log likelihood is linear in $\mathbf{y}$ the quasi-likelihood and likelihood are identical apart, possibly, from the unimportant multiplier, σ^2. Thus if $E(Y) = \operatorname{var}(Y) = \mu$ the quasi-likelihood is the same as the Poisson likelihood. However, if the distribution of Y is of the Euler type, $y^{\mu-1}e^{-y}/\Gamma(\mu)$ satisfying $E(Y) = \operatorname{var}(Y) = \mu$, the log likelihood is linear in $\log y$ so that the quasi-likelihood is different from the likelihood. There is a close connection here with natural exponential families (Morris, 1982).

EXAMPLE 3: *Invariance under linear transformations.* Let $\mathbf{Y}_L = \mathbf{LY}$, $\boldsymbol{\mu}_L = \mathbf{L}\boldsymbol{\mu}$ and $\mathbf{V}_L = \mathbf{LVL}^T$ where $\mathbf{L}$ is a nonsingular matrix of order N. The quasi-likelihood $\ell'(\cdot; \cdot)$ based on $\mathbf{Y}_L$ is a solution of

$$\frac{\partial\ell'}{\partial\boldsymbol{\mu}_L} = \mathbf{V}_L^-(\mathbf{y}_L - \boldsymbol{\mu}_L) = (\mathbf{L}^T)^{-1}\mathbf{V}^-(\mathbf{y} - \boldsymbol{\mu}).$$

Thus $\partial\ell'/\partial\boldsymbol{\mu} = \mathbf{L}^T(\partial\ell/\partial\boldsymbol{\mu}_L) = \mathbf{V}^-(\mathbf{y} - \boldsymbol{\mu})$. In other words the quasi-likelihood based on $\mathbf{Y}_L$ is the same as that based on $\mathbf{Y}$. This is weaker than the corresponding result for log likelihoods which are invariant under all invertible transforms.

EXAMPLE 4: *Multinomial response models.* The log likelihood for independent multinomial responses is a sum of terms each having the form (1) with $\sigma^2 = n^{-1}$ and $\theta_j = \log(\pi_j/\pi_k)$, $j = 1, \cdots, k$. Thus the likelihood is equivalent to the quasi-likelihood with $E(Y_j) = n\pi_j$, $\operatorname{var}(Y_j) = n\pi_j(1 - \pi_j)$ and $\operatorname{cov}(Y_i, Y_j) = -n\pi_i\pi_j$. Model construction is particularly simple and conceptually attractive if there is a transformation to linearity elementwise on the mean vector $\boldsymbol{\mu} = E(\mathbf{Y})$. This leads to generalized linear models and in particular to log-linear models. In some circumstances (McCullagh, 1980), it is more appealing first to form cumulative totals $Z_j = Y_1 + \cdots + Y_j$. Linear models are constructed by taking the logistic or other suitable transformation of the elements of $E(\mathbf{Z})$. In either case the usual asymptotic results based on $\hat{\boldsymbol{\beta}}$ and the inverse matrix of second derivatives depend only on the correctness of the second moment assumptions. In particular, to the order of approximation considered here, discreteness is irrelevant.

EXAMPLE 5: *Models with constant coefficient of variation.* Suppose for simplicity that $Y_1, \cdots, Y_N$ are uncorrelated with common mean μ and variance $\sigma^2\mu^2$ where σ is the known coefficient of variation. More realistic examples involve regression or the study of how μ

depends on other variables and σ would generally be unknown. However the broad qualitative conclusions to be drawn from the present example apply quite generally.

As always, for a single sample the quasi-likelihood estimator of μ is $\bar{y}$: the variance $\sigma^2\mu^2/N$ can be obtained directly or via $\ell(\mu; \mathbf{y}) = -\Sigma(y_i/\mu + \log \mu)$. The quantity $N/\sigma^2\mu^2$ can be thought of as the information in the quasi-likelihood available for estimating μ. In particular, no information is derived from the sample variance, in part because its precision cannot be assessed without involving higher order moments.

An alternative estimator, and one that seems appealing if the data are symmetrically distributed or nearly so, is to assume normality and independence and to use maximum likelihood. The resulting estimator, $\hat{\mu}$, does not have a simple closed form but it appears to be more precise: its asymptotic variance is $\sigma^2\mu^2(1 + 2\sigma^2)^{-1}/N$. However the validity of this asymptotic variance depends critically on the assumptions of independence and normality. It is readily shown after a little algebra that the correct asymptotic variance assuming independence up to fourth order but not normality is

$$\frac{\sigma^2\mu^2}{N}\left\{\frac{1 + \sigma^2(2 + \gamma_2) + 2\sigma\gamma_1}{(1 + 2\sigma^2)^2}\right\} + O(N^{-3/2}), \tag{18}$$

where γ_1 and γ_2 are the standardized third and fourth cumulants of Y. By independence up to fourth order we mean that the mixed cumulants of $Y_1, \cdots, Y_N$ up to degree four are all zero. Note that (18) can be arbitrarily large even for symmetrically distributed random variables.

The point here is that if information concerning higher moments is available, inferences can sometimes, though not always, be improved through the use of this information. However, the artificial injection of such information can have a number of undesirable side effects. In particular the true variance (18) may greatly exceed the apparent variance $\sigma^2\mu^2(1 + 2\sigma^2)^{-1}/N$. Even worse, the estimate $\hat{\mu}$ is not consistent unless the assumed moments up to fourth order are correct.

It is of interest to examine the conditions under which, to the order of approximation considered here, no further information concerning μ may be derived from the sample variance, s^2, beyond that already included in $\bar{Y}$. There is a connection here with first order sufficiency and ancillarity (Cox, 1980). The required condition, that the residual from s^2 after linear regression on $\bar{Y}$ should be first order ancillary for μ, implies $\kappa_3 = 2\sigma^4\mu^3$, a property of the gamma family.

EXAMPLE 6: *Over-dispersion.* The dispersion parameter σ^2 arises in a fairly natural way when the data are distributed according to the normal, inverse normal or gamma distributions. The usual models for discrete data including the Poisson, multinomial and non-central hypergeometric families do not include such a parameter. In applications it is well known that the variance often exceeds that predicted by the Poisson or multinomial distribution. Bartlett (1936) used the approximate relation $\text{var}(Y) = \sigma^2\mu(1 + b_1\mu)$ for counts from field trials. Armitage (1957) found $\text{var}(Y) = \sigma^2\mu^{b_2}$ with $1 < b_2 < 2$ and, for the most part, $\sigma^2 > 1$. Finney (1976) suggests estimating b from a large body of data and then, subject to occasional checks, to assume b constant. Fisher (1949), treating continuously distributed data as "pseudo-counts", took $b_2 = 1$ and found $\sigma^2 = 0.2$ approximately. The main point here is that if b is assumed known, quasi-likelihood estimates may be used, thus avoiding the difficulty of constructing a probability model for over-dispersed discrete observations.

Finally, in problems involving random effects it would often be appropriate to examine components of dispersion using the scale defined by σ^2 rather than the more usual components of variance. For example, if effects were multiplicative it might be appropriate to examine components of the coefficient of variation, this, rather than the variance, being constant across groups.

TABLE 1
Quasi-likelihood functions associated with some simple variance functions.

$v(\mu)$	$\ell(\mu; y)$	name	restrictions
σ^2	$\frac{1}{2}(y-\mu)^2$	normal	—
$\sigma^2\mu$	$y \log \mu - \mu$	Poisson	$\mu > 0; y \geq 0$
$\sigma^2\mu^2$	$-y/\mu - \log \mu$	gamma	$\mu > 0; y \geq 0$
$\sigma^2\mu^p$	$\mu^{-p}\left(\frac{\mu y}{1-p} - \frac{\mu^2}{2-p}\right)$	—	$p \neq 0, 1, 2; \mu > 0; y \geq 0$
$\sigma^2 e^{\mu}$	$-(y-\mu)e^{-\mu} + e^{-\mu}$	—	—
$\sigma^2\mu(1-\mu)$	$y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$	binomial	$0 < \mu < 1; 0 \leq y \leq 1$

7. A higher order theory. If further distributional or moment assumptions can be made it may be possible to develop a more refined theory than that given here. For example, if the log likelihood is in the exponential family, exact inference for $\boldsymbol{\beta}$ is possible provided that (1) can be written in the form $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$. On the other hand, if the exponential family is curved, conditioning on ancillary statistics has an effect of order $N^{-1/2}$ and cannot be ignored when improved approximations are required. The presence of the nuisance parameter σ^2 complicates matters: in particular, an appropriate definition of ancillarity is required.

Secondly, if higher order moments are known to be different from those obtained from (2), inferences based on the quasi-likelihood would appear to be inefficient. In simple cases some gains might be made by combining information from several low order moments. Whatever the form of the final estimator, an important consideration is that a good estimate of its precision should be readily available.

Acknowledgments. I wish to thank Professor D. R. Cox, Dr. B. Jorgensen, the associate editor and referees for helpful comments on an earlier version of the paper.

REFERENCES

ARMITAGE, P. (1957). Studies in the variability of pock counts. *J. Hyg. Camb.* **55** 564–581.
BARTLETT, M. S. (1936). Some notes on insecticide tests in the laboratory and in the field. *J. Roy. Statist. Soc.*, Suppl. 3, 185–194.
BRADLEY, E. L. (1973). The equivalence of maximum likelihood and weighted least squares estimates in the exponential family. *J. Amer. Statist. Assoc.* **68** 199–200.
BURRIDGE, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. *J. Roy. Statist. Soc. Ser. B* **43** 41–45.
COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.
COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.
COX, D. R. (1980). Local ancillarity. *Biometrika* **67** 279–286.
COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.
FINNEY, D. J. (1976). Radioligand assay. *Biometrics* **32** 721–740.
FISHER, R. A. (1949). A biological assay of tuberculins. *Biometrics* **5** 300–316.
HABERMAN, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5** 815–841.
HAYAKAWA, T. (1977). The likelihood ratio criterion and the asymptotic expansion of its distribution. *Ann. Inst. Statist. Math.* **29** 359–378.
JENNRICH, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40** 633–643.
JENNRICH, R. I. and MOORE, R. H. (1975). Maximum likelihood estimation by means of non-linear least squares. American Statistical Association; Statistical Computing Section. Proceedings. **1** 57–65.

JORGENSEN, B. (1983). Maximum likelihood estimation and large sample inference for generalized linear and non-linear regression models. *Biometrika* **70** (to appear).

MCCULLAGH, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc. Ser. B* **42** 109–142.

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.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

PRATT, J. W. (1981). Concavity of the log likelihood. *J. Amer. Statist. Assoc.* **76** 103–106.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61** 439–447.

WEDDERBURN, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for generalized linear models. *Biometrika* **63** 27–32.

WU, C. F. (1981). Asymptotic theory of non-linear least squares estimation. *Ann. Statist.* **9** 501–513.

DEPARTMENT OF MATHEMATICS
IMPERIAL COLLEGE
LONDON SW7 2AZ, ENGLAND