RECENT DEVELOPMENTS IN QUASI-LIKELIHOOD METHODS

David Firth
Department of Mathematics
University of Southampton
Southampton SO9 5NH
England

Please cite as: Firth, D (1993). Recent developments in quasi-likelihood methods. *Bull. Int. Stat. Inst.* **55**, 341–358. http://warwick.ac.uk/dfirth/papers/Florence1993.pdf

1. *Introduction: Quasi-likelihood estimating equations*

The term 'quasi-likelihood' has been used recently to describe a fairly wide variety of techniques for estimation and inference. This paper focuses on the method introduced by Wedderburn (1974), further developed by McCullagh (1983), but which, as pointed out by Crowder (1987), has roots at least as far back as Williams (1959, §4.5). A brief introduction and overview are given by McCullagh (1986), and more comprehensive treatments with examples of application may be found in McCullagh & Nelder (1989) and McCullagh (1991).

The quasi-likelihood method of estimation is probably best viewed as a straightforward extension of generalized least squares. Suppose that $\mathbf{y}$ is a $n \times 1$ response vector, assumed to be a realization of a random vector $\mathbf{Y}$ with

$$E(\mathbf{Y}) = \boldsymbol{\mu}(\boldsymbol{\beta}), \qquad \mathrm{cov}(\mathbf{Y}) = \phi V(\boldsymbol{\mu}), \tag{1}$$

where $\mu_i(\beta)$ $(i = 1, \ldots, n)$ are regression functions depending on a $p \times 1$ vector of unknown parameters $\boldsymbol{\beta}$, ϕ is a scalar dispersion parameter and $V(.)$ is a symmetric, positive-definite matrix of known functions of the unknown means $\boldsymbol{\mu}$. The functions $\mu_i(.)$ typically express dependence on explanatory variables, often but not necessarily *via* a linear model $\boldsymbol{\mu} = X\boldsymbol{\beta}$ or a generalized linear model $g(\boldsymbol{\mu}) = X\boldsymbol{\beta}$ for some specified link function $g(.)$; the parameters $\boldsymbol{\beta}$ are taken to be of interest. If the elements of V are known constants, not depending on $\boldsymbol{\mu}$, a standard generalized least-squares approach minimizing $(\mathbf{y} - \boldsymbol{\mu})^T V^{-1} (\mathbf{y} - \boldsymbol{\mu})$ yields the vector equation

$$D^T V^{-1} (\mathbf{y} - \boldsymbol{\mu}) = 0, \tag{2}$$

to be solved for $\boldsymbol{\beta}$, where D is the $n \times p$ matrix of derivatives $\partial \mu_i / \partial \beta_r$. When the covariance matrix V is functionally dependent on $\boldsymbol{\mu}$, the equations (2) are called the *quasi-likelihood* (estimating) *equations* and their solution vector $\hat{\boldsymbol{\beta}}$, assuming it exists, the *quasi-likelihood estimate*. Quasi-likelihood estimation may therefore be viewed as extending the domain of application of generalized least squares *via* the estimating equations (2). McCullagh (1991) argues strongly that this, rather

than the possibly more obvious route of minimizing $(\mathbf{y}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})$, is the appropriate extension of least-squares ideas to the general setting (1).

Quasi-likelihood estimates have an optimality property that may be regarded as extending the familiar Gauss-Markov optimality of least-squares estimates. McCullagh (1983) shows that, among all estimators obtained as solutions to linear (*i.e.*, linear in $\mathbf{y}$) unbiased estimating equations, the quasi-likelihood estimator $\hat{\boldsymbol{\beta}}$ is best in the sense of having greatest asymptotic precision; if $\tilde{\boldsymbol{\beta}}$ is another estimator obtained from linear estimating equations, then for any constant vector $\mathbf{a}$ the asymptotic variance of $\mathbf{a}^T\tilde{\boldsymbol{\beta}}$ is at least as great as that of $\mathbf{a}^T\hat{\boldsymbol{\beta}}$. See also McCullagh & Nelder (1989, §9.5) and Morton (1981). An alternative, non-asymptotic extension of the Gauss-Markov theorem holds for the estimating equation (2) itself (*e.g.*, Godambe & Heyde, 1987; Godambe & Thompson, 1989), but it appears that such finite-sample optimality does not in general extend to the estimate derived as its solution.

The name 'quasi-likelihood' was coined by Wedderburn (1974) largely because of similarities between the behaviour of the vector

$$\mathbf{U} = D^T V^{-1}(\mathbf{Y}-\boldsymbol{\mu})/\phi \tag{3}$$

and that of a likelihood-based score vector. Specifically, the familiar identities

$$E(\mathbf{U}) = 0, \qquad \mathrm{cov}(\mathbf{U}) = -E(\partial\mathbf{U}/\partial\boldsymbol{\beta}) \tag{4}$$

that hold if $\mathbf{U}$ is the score vector in a regular likelihood-based model continue to hold for the 'quasi-score' vector $\mathbf{U}$ in (3), under only the second-moment assumptions (1). Since the identities (4) form the basis of standard arguments for the asymptotic properties of maximum likelihood estimates and associated inference procedures, similar properties hold also for quasi-likelihood estimates under the considerably weaker assumptions (1), subject of course to certain regularity and other technical conditions.

A further connection between quasi-likelihood and maximum likelihood estimates is that, if there exists a linear exponential-family model for $\mathbf{y}$ that has $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $\mathrm{cov}(\mathbf{Y}) \propto V(\boldsymbol{\mu})$, then solution of (2) for $\boldsymbol{\beta}$ is equivalent to solving the maximum likelihood equations for that exponential family. The simplest example is the case $V(\boldsymbol{\mu}) = I$, the identity matrix, corresponding to constant-variance regression with no correlation. The unique exponential-family model with this mean-variance structure is $\mathbf{Y} \sim N(\boldsymbol{\mu}, \phi I)$ (Morris, 1982). The equivalence in this case between solution of (2), the least-squares equations, and maximization of the normal-model likelihood for $\boldsymbol{\beta}$, is well known. Other examples of this equivalence are:

Covariance function $V(\boldsymbol{\mu})$	*Exponential family model*
$\mathrm{diag}(\mu_i)$	Independent Poisson
$\mathrm{diag}\{\mu_i(1-\mu_i/m_i)\}$	Independent binomials, indices m_i
$\mathrm{diag}(\mu_i^2)$	Independent exponential/gamma
V (no dependence on $\boldsymbol{\mu}$)	Multivariate normal

In these and other such instances, quasi-likelihood estimation of the parameters $\boldsymbol{\beta}$ is operationally equivalent to maximum likelihood estimation based on a standard model. This equivalence is often exploited at a practical level by making use of software designed for exponential-family maximum likelihood calculations, such as `GLIM`, to obtain quasi-likelihood estimates.

This paper reviews some of the main developments and applications of quasi-likelihood methods as introduced above. The aim will be broadly to indicate strands of development and to provide appropriate references, rather than to explore any aspect in detail. Sections 2-4 discuss general

issues such as the estimation of standard errors, procedures for inference, efficiency and robustness. Sections 5-7 mention some of the major applications of quasi-likelihood methods to date, including models for overdispersed data and the 'generalized estimating equations' approach to modelling longitudinal and clustered data. Sections 8 and 9 introduce the ideas of 'extended' quasi-likelihood and 'local' quasi-likelihood, designed to increase flexibility in the variance specification and in the regression model, respectively.

A notable omission from this review is discussion of the somewhat broader, martingale-based theory of quasi-likelihood estimation and inference in stochastic processes (*e.g.*, Hutton & Nelson, 1986; Godambe & Heyde, 1987; Heyde, 1989; Sørensen, 1990) and the theory of estimating functions generally. These topics are surveyed in Godambe (1991); see also McCullagh (1991) or McCullagh & Nelder (1989, §9.4).

2. *Standard errors*

Under the model assumptions (1), and subject to certain regularity and other conditions, the quasi-likelihood estimate $\hat{\boldsymbol{\beta}}$ that solves (2) is consistent and asymptotically normal, with asymptotic variance-covariance matrix $[i(\boldsymbol{\beta})]^{-1} = \phi(D^T V^{-1} D)^{-1}$. This depends in general on $\boldsymbol{\beta}$ through the matrices D and V, and on the dispersion parameter ϕ. When computing estimated standard errors it is usual to substitute $\hat{\boldsymbol{\beta}}$ for the unknown $\boldsymbol{\beta}$, and if ϕ is unknown to estimate it using an appropriately chosen quadratic form. Since

$$E\left[(\mathbf{Y}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})\right] = n\phi,$$

an unbiased estimate of ϕ would be $(\mathbf{y}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})/n$; the 'degrees of freedom corrected' (but not, in general, unbiased) estimate

$$\tilde{\phi} = (\mathbf{y}-\hat{\boldsymbol{\mu}})^T V^{-1}(\hat{\boldsymbol{\mu}})(\mathbf{y}-\hat{\boldsymbol{\mu}})/(n-p)$$

takes account of substituting $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})$ for $\boldsymbol{\mu}$, by analogy with the corresponding procedure for linear models. Estimated standard errors for components of $\hat{\boldsymbol{\beta}}$ are then computed as square roots of diagonal elements of

$$\mathrm{cov}_M(\hat{\boldsymbol{\beta}}) = \tilde{\phi}(\hat{D}^T \hat{V}^{-1} \hat{D})^{-1}, \tag{5}$$

where $\hat{D}$ denotes $D(\hat{\boldsymbol{\beta}})$, *etc.*, and the subscript M indicates a 'model-based' estimate.

The estimated covariance matrix (5) is constructed on the assumption that the model specification (1) is correct. In many applications the assumption $E(\mathbf{Y}) = \boldsymbol{\mu}(\boldsymbol{\beta})$ defines the quantities of interest, while the second-moment specification $\mathrm{cov}(\mathbf{Y}) = \phi V(\boldsymbol{\mu})$ is a working 'guess' at the true covariance structure, made in the hope of obtaining increased efficiency when estimating $\boldsymbol{\beta}$. In this respect the quasi-likelihood approach enjoys a useful robustness property. Since the quasi-likelihood estimating equation (2) is linear in $\mathbf{y}$, it is an unbiased estimating equation (*i.e.*, the left hand side has mean zero) under only the first-moment specification $E(\mathbf{Y}) = \boldsymbol{\mu}(\boldsymbol{\beta})$; unbiasedness of the estimating equation, and hence (subject to certain conditions) consistency of $\hat{\boldsymbol{\beta}}$, is robust to failure of the working covariance structure $V(\boldsymbol{\mu})$. Unfortunately, the same is not true for the estimated covariance matrix in (5). If $\mathrm{cov}(\mathbf{Y})$ is not as specified in (1), the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is the 'information sandwich'

$$(D^T V^{-1} D)^{-1} D^T V^{-1} \,\mathrm{cov}(\mathbf{Y}) V^{-1} D (D^T V^{-1} D)^{-1}$$

(*e.g.*, Cox, 1961), an obvious estimate of which is

$$\mathrm{cov}_R(\hat{\boldsymbol{\beta}}) = (\hat{D}^T\hat{V}^{-1}\hat{D})^{-1}\hat{D}^T\hat{V}^{-1}(\mathbf{y}-\hat{\boldsymbol{\mu}})(\mathbf{y}-\hat{\boldsymbol{\mu}})^T\hat{V}^{-1}\hat{D}(\hat{D}^T\hat{V}^{-1}\hat{D})^{-1}. \qquad (6)$$

For a general discussion motivating this type of 'robust' covariance matrix estimator, see Royall (1986); see also Pregibon (1983).

The use of a working covariance function $V(\boldsymbol{\mu})$ to 'tune' the quasi-likelihood procedure in the hope of near-optimal estimation of $\boldsymbol{\beta}$, together with robust standard errors derived from $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$, is the essence of the 'generalized estimating equations' approach to the analysis of longitudinal or clustered data (Liang & Zeger, 1986; see also §7).

The price paid for robustness is typically some loss of efficiency under ideal conditions, so it is to be expected that $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$ will be less efficient than the model-based estimator $\mathrm{cov}_M(\hat{\boldsymbol{\beta}})$ when the working $V(\boldsymbol{\mu})$ is in fact correct. As a simple example to illustrate this, consider (as in Royall, 1986) the one-parameter model $\mu_i(\beta) = \beta$ $(i = 1, \ldots, n)$, $V(\boldsymbol{\mu}) = \mathrm{diag}(\boldsymbol{\mu})$ in which ϕ is assumed to be 1. Then $\hat{\beta} = \bar{y}$, the sample mean, and the model-based and robust variance estimates for $\hat{\beta}$ are

$$\mathrm{cov}_M(\hat{\beta}) = \bar{y}/n, \qquad \mathrm{cov}_R(\hat{\beta}) = \sum_{i=1}^{n}(y_i-\bar{y})^2/n^2.$$

The asymptotic relative efficiency of the two variance estimators, under a working assumption that $Y_i \sim \mathrm{Poisson}(\beta)$, is $1/(1+2\beta)$: the loss of efficiency incurred when using $\mathrm{cov}_R(\hat{\beta})$ can be severe if β is large.

The example just given is unrealistically simple but serves to illustrate the general point. In a more complex setting involving overdispersed counts, Breslow (1990b) confirms by simulation that $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$ can be much less precise than $\mathrm{cov}_M(\hat{\boldsymbol{\beta}})$. However, the picture is not entirely clear. In further simulations, made in the context of correlated binary data, Sharples & Breslow (1992) find that $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$ is almost as efficient as $\mathrm{cov}_M(\hat{\boldsymbol{\beta}})$ when the assumed form of $V(\boldsymbol{\mu})$ is correct. A systematic study, to identify the conditions under which $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$ has reasonably high efficiency, would be useful.

In situations where loss of efficiency can be serious, there is a trade-off between precision under the assumed $V(\boldsymbol{\mu})$ and robustness to departures from that working model. A 'compromise' estimator of the form $\lambda\,\mathrm{cov}_R(\hat{\boldsymbol{\beta}})+(1-\lambda)\,\mathrm{cov}_M(\hat{\boldsymbol{\beta}})$, where $0 \le \lambda \le 1$, is motivated by Firth (1987a) using a partially Bayesian argument; see also Efron (1986b). The tuning constant λ may be chosen either on the basis of previous experience, or derived, in an empirical Bayes fashion, from the observed distribution of the residuals. More work is needed on this 'adaptive' type of estimator.

A further problem with the robust estimator of $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ is its bias. The true standard deviations of elements of $\hat{\boldsymbol{\beta}}$ are typically underestimated by standard errors derived from $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$, as is confirmed in simulation studies by Breslow (1990b) in the context of overdispersed Poisson models, and by Sharples & Breslow (1992) and Lee, Scott and Soo (1993) in models for correlated binary data. The problem is most easily understood in terms of a simple linear model in which $\boldsymbol{\mu} = X\boldsymbol{\beta}$, $V(\boldsymbol{\mu}) = I$ and $\hat{\boldsymbol{\beta}}$ is the ordinary least-squares estimate $(X^TX)^{-1}X^T\mathbf{y}$. The robust estimator of $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ in this case is

$$\mathrm{cov}_R(\hat{\boldsymbol{\beta}}) = (X^TX)^{-1}X^T\,\mathrm{diag}(y_i-\hat{\mu}_i)^2X(X^TX)^{-1},$$

which is biased since the squared residuals $(y_i-\hat{\mu}_i)^2$ tend to underestimate the corresponding variances $\mathrm{var}(Y_i)$; for example, if $\mathrm{var}(Y_i) = \phi$ (constant) as in the working model, then $E\left[(Y_i-\hat{\mu}_i)^2\right] = \phi(1-h_{ii})$ where $h_{ii} \in (0,1)$ is the ith 'leverage', *i.e.*, the ith diagonal element of the 'hat' matrix $H = X(X^TX)^{-1}X^T$. While consistency of $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$ for the true $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ 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 $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$ can be severe. Chesher & Jewitt (1986) give examples and calculate bounds, based on $\max\{h_{ii}\}$, for the bias.

Various bias-corrected versions of $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$ have been suggested in the linear-model case. These include (i) a version corrected by replacing each squared residual $(y_i - \hat{\mu}_i)^2$ by $(y_i - \hat{\mu}_i)^2/(1 - h_{ii})$ (MacKinnon & White, 1985), and (ii) a 'degrees of freedom' correction in which $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$ is simply multiplied by $n/(n-p)$ (Hinkley, 1977). See also Wu (1986). The former correction is exactly unbiased under the working, constant-variance model; the latter requires also that $h_{ii} =$ constant $= p/n$ $(i = 1, \ldots, n)$, *i.e.*, a balanced design, to achieve exact unbiasedness. A potential problem with approach (i) occurs if there are observations with high leverage, *i.e.*, if h_{ii} is close to 1 for some i. In such a case, division by $1 - h_{ii}$ is not only bias-reducing but is also severely variance-inflating, so the bias-corrected version of $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$ may be rather unstable; an extreme example appears in Firth (1987a). Further work is needed to establish a generally-reliable form of bias correction for $\mathrm{cov}_R(\hat{\boldsymbol{\beta}})$, and to extend these ideas drawn from the literature on linear models to the wider context of quasi-likelihood models.

3. *Tests and confidence regions*

As in likelihood-based theory, standard methods for approximate inference on $\boldsymbol{\beta}$ in (1) include the 'Wald'-type test, the (quasi-)score test and the (quasi-)likelihood ratio test, and procedures based on these for the construction of confidence regions. McCullagh (1991, §11.6) provides a good summary. The methods are as in the familiar likelihood-based theory, but with two main differences.

First, although the quasi-likelihood approach outlined in §1 provides the necessary ingredients $\hat{\boldsymbol{\beta}}$, $\mathbf{U}$ and $i(\boldsymbol{\beta})$ for Wald-based and score-based inferences, no 'likelihood-like' function is immediately available for tests and confidence regions of the likelihood-ratio (LR) type. In the 'independence' case with $V(\boldsymbol{\mu}) = \mathrm{diag}\{V_i(\mu_i)\}$, this is remedied by defining a quasi-loglikelihood function

$$Q(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V_i(t)} dt \,, \tag{7}$$

which is like a log-likelihood in the sense that its derivative vector with respect to $\boldsymbol{\beta}$ is the quasi-score function,

$$\nabla Q = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V_i(\mu_i)} \times \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = D^T V^{-1} (\mathbf{y} - \boldsymbol{\mu})/\phi = \mathbf{U}.$$

More generally, if $V(\boldsymbol{\mu})$ is non-diagonal, it is not usually possible to find a function Q such that $\nabla Q = \mathbf{U}$; a notable exception is if V does not depend on $\boldsymbol{\mu}$, in which case Q is the multivariate normal $N(\boldsymbol{\mu}, \phi V)$ log-likelihood. Some recent theoretical work, *e.g.*, Li (1992), has focused on the construction of a 'likelihood-like' function when no Q satisfies $\nabla Q = \mathbf{U}$ exactly; see also McCullagh & Nelder (1989, §9.3) and McCullagh (1991, §11.7). The results of this work seem quite promising, but it is too early to assess their practical value. At present, then, likelihood-ratio type methods are available when $\nabla Q = \mathbf{U}$ can be satisfied by appropriate choice of the quasi-loglikelihood function Q, but not in general.

The second main difference is the presence of the dispersion parameter ϕ. If ϕ is unknown it must be estimated, *e.g.*, by $\tilde{\phi}$ as in (5). Allowance for estimation of ϕ may be made by using F

instead of χ^2 approximations for the null distributions of test statistics. For example, an approximate $100(1-\alpha)\%$ confidence region of the form

$$\{\boldsymbol{\beta}:\ W(\boldsymbol{\beta}) \leq \phi\chi^2_{p,\alpha}\},$$

for ϕ known, becomes

$$\{\boldsymbol{\beta}:\ W(\boldsymbol{\beta})/p \leq \tilde{\phi}F_{p,n-p,\alpha}\}$$

if $\tilde{\phi}$ is an estimate on $n-p$ degrees of freedom; here $W(\boldsymbol{\beta})$ is $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\hat{D}^T\hat{V}^{-1}\hat{D})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ (estimate-based), or $\mathbf{U}^T(D^TV^{-1}D)^{-1}\mathbf{U}$ (score-based), or $2[Q(\hat{\boldsymbol{\beta}};\mathbf{y})-Q(\boldsymbol{\beta};\mathbf{y})]$ (LR-based, if Q is defined).

The relative merits of estimate-based, score-based and LR-based methods are much as in standard likelihood theory. The main considerations are outlined by McCullagh (1991, §11.6), who argues that LR-based methods are preferable when available; this justifies the ongoing search for a suitable function to use as a quasi-loglikelihood in cases where no Q satisfies $\nabla Q = \mathbf{U}$. When no such Q exists, or if the computational expense of LR-based confidence regions cannot be met, score-based methods are preferred to the approach based on asymptotic normality of $\hat{\boldsymbol{\beta}}$, on the grounds of invariance to reparameterization.

If the 'working' covariance matrix $V(\boldsymbol{\mu})$ is incorrect, the asymptotic theory underlying the above χ^2 approximations is invalid. A remedy, at least asymptotically, in the case of the estimate-based and score-based methods, is to make appropriate use of the 'information sandwich' of §2 when constructing test statistics, *etc.* Details are given by Rotnitzky & Jewell (1991), for example. However, as observed by Rotnitzky & Jewell and by others (*e.g.*, Breslow, 1990b), the resulting 'robust' methods can be rather unstable in other than large samples, and are computationally quite cumbersome. As an alternative, Rotnitzky & Jewell propose adjustments like those of Rao & Scott (1981) to the more stable, model-based methods described above, designed to restore the validity of the standard χ^2 approximations.

4. *Efficiency and robustness*

While quasi-likelihood provides estimators that are optimal among those derived from linear estimating equations, an obvious question is to what extent efficiency is lost relative to estimators outside that class. It is natural to make comparisons with the 'globally' optimum method, maximum likelihood. Cox & Hinkley (1968) consider the case of ordinary least-squares estimation in a linear model $\{\boldsymbol{\mu}(\boldsymbol{\beta}) = X\boldsymbol{\beta},\ V(\boldsymbol{\mu}) = I\}$, and show how the asymptotic efficiency of least squares relative to maximum likelihood depends principally on the skewness of the true error distribution. Firth (1987b) extends these calculations to the context of models with constant coefficient of variation $\{V(\boldsymbol{\mu}) = \mathrm{diag}(\mu_i^2)$; see §6$\}$, and to quasi-likelihood models for overdispersed binomial, Poisson and exponential data (see §5). It is found that quasi-likelihood estimates have high efficiency under modest overdispersion, confirming a conclusion of Cox (1983).

Under only the moment assumptions (1), the method of maximum likelihood is not available and so the calculations just referred to are mainly of theoretical interest. An alternative comparison is with an estimator derived from the optimum *quadratic* estimating equation (Crowder, 1987; Firth, 1987b) based on (1). Such calculations also are mainly of theoretical interest, since the optimum relative weight given to linear and quadratic terms in the estimating equation depends on the 3rd and 4th moments of $\mathbf{Y}$, which are not specified in (1). Crowder (1987) shows that in some situations even a sub-optimal quadratic estimating equation, not requiring knowledge of 3rd and 4th

moments, yields an estimator with markedly better properties than the quasi-likelihood estimator; but such situations do not seem typical.

Godambe & Thompson (1989) advocate the general use of optimum quadratic estimating equations under assumptions such as (1), and present a unified theoretical framework, but as noted in Firth (1987b) there are some practical difficulties. Quadratic estimating equations rely for their unbiasedness on the correctness of $V(\boldsymbol{\mu})$, so the corresponding estimates lack the robustness of quasi-likelihood estimates, mentioned in §2, to failure of the 'working' covariance function. The problem of needing to know 3rd and 4th moments to determine optimum weights has been mentioned already; if these are to be estimated from the data, highly unstable results are to be expected in all but very large samples. Finally, and more seriously from a practical viewpoint, valid standard errors for estimates derived from quadratic estimating equations also depend on 3rd and 4th moments, and may therefore be difficult to determine reliably.

Quadratic estimating equations have been found useful in situations where the working covariance function $V(\boldsymbol{\mu})$ is not fully specified but depends on a further unknown parameter or parameters (see §§7,8).

The 'robustness' of quasi-likelihood estimates, in the sense of consistency under misspecification of $V(\boldsymbol{\mu})$, should not be confused with the more usual meaning of robustness in connection with regression methods, *i.e.*, robustness to contamination of the data by spurious observations or gross errors. In common with the method of least squares, quasi-likelihood estimators have a linear influence function and therefore exhibit similar sensitivity to outliers. Robust alternatives to least squares are well-developed in the literature; Morgenthaler (1992) explores the generalization of one such alternative, the method of 'least absolute deviations', to the more general quasi-likelihood setting.

5. *Overdispersion*

Quasi-likelihood methods are now routinely used in regression problems, especially generalized linear models, to deal with data that are overdispersed relative to a standard probability model, *e.g.*, a binomial or Poisson model. In the case of overdispersed 'Poisson' data, for example, variance assumptions that may be made in order to accommodate extra variation are

$$\mathrm{var}(Y_i) = \phi\mu_i \tag{8}$$

or

$$\mathrm{var}(Y_i) = \mu_i + \lambda\mu_i^2. \tag{9}$$

Of these, specification (8) fits directly into the quasi-likelihood framework; overdispersion is represented by $\phi > 1$, and the rarer phenomenon of underdispersion by $\phi < 1$. The quasi-likelihood equations for $\boldsymbol{\beta}$ are the same as maximum likelihood equations based on a Poisson model, but the standard likelihood analysis is modified by estimation of ϕ as in (5). That is, estimated standard errors derived from the standard likelihood analysis are multiplied by $\sqrt{\tilde{\phi}}$ to allow for overdispersion, or underdispersion, whichever is the case.

The alternative specification (9) is more complicated, but is often regarded as more realistic. One simple justification is that if $\lambda > 0$ it corresponds to a 'random effects' model in which $Y_i|Z_i \sim \mathrm{Poisson}(\mu_i Z_i)$, and $Z_1, \ldots, Z_n$ are i.i.d. random effects with $E(Z_i) = 1$, $\mathrm{var}(Z_i) = \lambda$; here Z_i might be thought of as representing the effect of unobserved covariates in a log-linear model, for example. If λ is known, (9) is also amenable to the quasi-likelihood approach, in this

case with $\phi = 1$. This suggests a natural 'see-saw' algorithm in which quasi-likelihood estimation of $\boldsymbol{\beta}$ is alternated with estimation of λ by some other method.

For overdispersed binomial data, the appropriate variance functions are different but the basic idea is the same. Discussion of suitable variance specifications in the binomial case can be found in Cox & Snell (1989, §3.2.4) and McCullagh & Nelder (1989, §4.5.1); see also Morton (1991).

The details of quasi-likelihood estimation, including moment-based methods for parameters in the variance function, such as λ above, are given by Williams (1982) for overdispersed binomial models and by Breslow (1984) for the Poisson case. Moore (1986) provides a unified asymptotic treatment. Methods for criticism and elaboration of the assumed variance function are developed by Moore (1987) and Ganio & Schafer (1992). Breslow (1989,1990a,b) studies quasi-score tests in the overdispersed-Poisson context.

6. *Multiplicative errors*

Multiplicative-error regression models, in which $Y_1, \ldots, Y_n$ are assumed independent with

$$Y_i = \mu_i(\boldsymbol{\beta}) Z_i \qquad (i = 1, \ldots, n) \tag{10}$$

where the errors Z_i have $E(Z_i) = 1$, $\mathrm{var}(Z_i) = \phi$, are natural alternatives to the standard additive-error model if the response is a positive measurement, for example. The variance is $\mathrm{var}(Y_i) = \phi\mu_i^2$, which may adequately describe the commonly-found phenomenon of greater imprecision in larger measurements.

A standard approach in such a model is to log-transform the data so that the errors become additive, and then to perform a least-squares analysis, *i.e.*, a maximum-likelihood analysis based on a lognormal distribution for the $\{Z_i\}$. Alternatively, (10) may be treated directly by quasi-likelihood, which in the case $V(\boldsymbol{\mu}) = \mathrm{diag}\{\mu_i^2\}$ corresponds to maximum likelihood based on the assumption that the $\{Z_i\}$ have a gamma distribution. There are thus two natural quasi-likelihood approaches to (10), let us call them the lognormal and gamma methods, and it is of some interest to consider briefly their relative merits. Asymptotic efficiency calculations (Firth, 1988; Hill & Tsai, 1988) are inconclusive. For example, there is little loss of efficiency if the gamma method is used when in fact the $\{Z_i\}$ are lognormally distributed, and *vice versa*. From a practical viewpoint also it appears that there is usually little difference between the two methods of analysis, except that the lognormal approach is highly sensitive to potential rounding error in measurements close to zero, and breaks down completely if any observed y_i is zero or negtive; such problems are avoided if the original scale of the data is used throughout, as in the gamma method.

7. *Longitudinal, clustered and other correlated data*

The most active area of recent research on quasi-likelihood has been application and further development of the methods in modelling non-independent responses, often in the form of counts or binary outcomes, for which standard normal-theory approaches are inappropriate. Problems involving longitudinal or repeated-measures data, cluster-sampled data, data with hierarchical error structure or 'random effects', and time series, have been successfully analysed by quasi-likelihood methods. The relevant literature is large, and only a brief survey of some of the main developments is attempted here.

Regression problems involving successive responses from each of K independent subjects, or from subjects sampled in K independent clusters, may be considered as a special case of (1) in

which $V(\boldsymbol{\mu})$ is block-diagonal with covariance function matrices $V_k(\boldsymbol{\mu}_k)$ $(k = 1, \ldots, K)$ for each 'cluster', or for each subject in the repeated-measures context. A simple framework for imposing a common structure upon $V_1, \ldots, V_K$ is provided by Liang & Zeger (1986), who write

$$V_k(\boldsymbol{\mu}_k) = A_k^{1/2}(\boldsymbol{\mu}_k) R_k(\boldsymbol{\alpha}) A_k^{1/2}(\boldsymbol{\mu}_k)$$

in which A_k is a diagonal matrix of variance functions and $R_k(\boldsymbol{\alpha})$ a working correlation matrix, possibly depending on unknown parameters $\boldsymbol{\alpha}$ that are assumed equal for all k. As a simple example, if the responses are counts then A_k might be specified as $A_k = \mathrm{diag}\{\boldsymbol{\mu}_k\}$, mimicking the Poisson variance function. The simplest specification of $R_k(\boldsymbol{\alpha})$ is $R_k = I$ (Scott & Holt, 1982; Binder, 1983), which simply 'ignores' any correlation among the responses. More complex specifications of $R_k(\boldsymbol{\alpha})$, based on a 'guess' at the underlying correlation structure, might be made in the hope of improved efficiency of estimation for the regression parameters $\boldsymbol{\beta}$ in $\boldsymbol{\mu}(\boldsymbol{\beta})$. In the case of cluster sampling, an obvious specification has all the off-diagonal elements equal to α, a single intra-cluster correlation parameter. For longitudinal data, $R(\boldsymbol{\alpha})$ may be chosen as in Liang & Zeger (1986) or Zeger & Liang (1986) to describe a pattern of serial correlation among repeated observations on the same subject. If $\boldsymbol{\alpha}$ is known, quasi-likelihood estimation of $\boldsymbol{\beta}$ can be applied directly. In practice an estimate of $\boldsymbol{\alpha}$ must be used; Liang & Zeger (1986) develop a simple method based on residuals.

Prentice (1988) and Prentice & Zhao (1991) develop these ideas further, and in particular suggest that the quasi-likelihood estimating equations for $\boldsymbol{\beta}$ be supplemented by quadratic estimating equations, then solved for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ jointly. Zeger, Liang & Albert (1988) make a useful distinction between 'population averaged', or marginal, regression models as above, and 'subject specific' models involving the estimation of random effects. Thall & Vail (1990) discuss alternative forms of working covariance function $V_k(\boldsymbol{\mu}_k)$ for longitudinal count data. For correlated binary data, Lipsitz, Laird & Harrington (1991) show that specification in terms of odds ratios rather than correlations has certain advantages; Liang, Zeger & Qaqish (1992) and the accompanying discussion synthesize much of the development up to mid-1991; Sharples & Breslow (1992) and Lee, Scott & Soo (1993) explore efficiency relative to maximum likelihood, and present valuable empirical findings on the finite-sample performance of estimates and standard errors. Ashby *et al.* (1992) give other relevant references, including several on applications of these methods.

Morton (1987, 1988) uses quasi-likelihood to model count data with a more complex, hierarchical variance-covariance structure, and develops quasilikelihood-ratio tests that mimic the standard analysis of variance procedures for 'split-plot' type experiments. Engel (1990) and Firth & Harris (1991) find that the same approach is particularly simple in the case of measurements subject to several sources of multiplicative error; as in §6, quasi-likelihood then provides an alternative to standard least-squares analysis of log-transformed data. Breslow & Clayton (1993) elegantly synthesize a number of aspects of the quasi-likelihood approach to generalized linear models with random effects, and present a 'penalized' quasi-likelihood method for use when subject-specific rather than marginal effects are of interest; this is closely related to the general, algorithmic approach of Schall (1991).

Quasi-likelihood methods have also been applied recently in the analysis of non-Gaussian time series. For example, in the terminology used by Cox (1981) to classify time-series models, Zeger (1988) uses quasi-likelihood estimation in the context of a 'parameter-driven' model for a time series of counts, while Zeger & Qaqish (1988) show how to construct 'observation-driven' quasi-likelihood models for counts and other non-Gaussian series in which Markov structure can be assumed.

8. *Joint modelling of mean and dispersion: 'extended' quasi-likelihood*

The 'overdispersed Poisson' variance function (9), involving an unknown parameter λ, falls outside the standard quasi-likelihood framework (1). The methods of §§1-3 do not yield an estimate of λ. The same difficulty arises if the dispersion parameter ϕ is not assumed constant as in (1) but is allowed to depend on covariates, as suggested by Pregibon (1984), according to some parametric model. In simple cases such as (9), moment-based estimators can be constructed for parameters in the variance specification (*e.g.*, Williams, 1982; Breslow, 1984). The need for a systematic approach for more general use prompted Pregibon (1984) and Nelder & Pregibon (1987) to develop a 'likelihood-like' method; the quasi-loglikelihood defined in (7), although 'likelihood-like' with regard to $\boldsymbol{\beta}$, cannot be used directly for inference on a parameter such as λ.

A fairly flexible generalization of the standard variance assumption $\mathrm{var}(Y_i) = \phi V_i(\mu_i)$ is

$$\mathrm{var}(Y_i) = \phi_i(\boldsymbol{\gamma}) V_i(\mu_i; \boldsymbol{\lambda}), \tag{11}$$

in which the functions $\phi_i(.)$ are determined by known covariates or factors, *e.g.*, *via* a generalized linear model, and $\boldsymbol{\lambda}$ provides some flexibility in the specification of $V_i(.)$. In most applications, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are of low dimension and are not both present in the same model. The overdispersed-Poisson variance (9) is an example with scalar λ and no $\boldsymbol{\gamma}$. In overdispersed-binomial problems, a commonly used variance assumption (*e.g.*, Cox & Snell, 1989, §3.2.4) having γ but not $\boldsymbol{\lambda}$, is

$$\mathrm{var}(Y_i) = \{1 + \gamma(m_i - 1)\} V_i(\mu_i),$$

where $V_i(\mu_i)$ is the binomial variance and m_i the binomial 'index' or 'number of trials'. Many other instances of variance assumptions in the form (11) have appeared in the literature. Of particular current interest is the use of this type of model in connection with quality-improvement experiments (*e.g.*, Nelder & Lee (1991); Engel, 1992), where $\phi_i(\boldsymbol{\gamma})$ is used to model dispersion separately from the mean in order that process conditions can be determined under which the mean is close to a desired target while the dispersion is minimized.

Pregibon (1984) and Nelder & Pregibon (1987) suggest, for joint inference on $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$, an 'extended' quasi-loglikelihood (EQL) function

$$Q^+(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}; \mathbf{y}) = \sum_{i=1}^{n} \left\{ -\frac{1}{2} \log[2\pi \phi_i(\boldsymbol{\gamma}) V_i(y_i; \boldsymbol{\lambda})] - \frac{1}{2} \frac{D_\lambda(y_i, \mu_i)}{\phi_i(\boldsymbol{\gamma})} \right\},$$

in which D_λ is the *deviance* function,

$$D_\lambda(y_i, \mu_i) = -2 \int_{y_i}^{\mu_i} \frac{y_i - t}{V_i(t; \boldsymbol{\lambda})} dt,$$

corresponding to the variance function $V_i(\mu_i; \boldsymbol{\lambda})$. Similar constructions are suggested also by West (1985) and, particularly, Efron (1986). Motivation for Q^+ is provided by the fact that, if there exists a linear exponential-family model with variance function $V_i(\mu_i, \boldsymbol{\lambda})$, Q^+ is the log-likelihood function based on a saddlepoint approximation to that family. The overdispersed-Poisson variance function $\mu_i + \lambda \mu_i^2$, for example, is the variance function for a negative binomial family; use of Q^+ for inference in that case is approximately the same as use of the negative-binomial likelihood. The function Q^+ 'extends' Q in (7) in the sense that, in the absence of $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$, the two functions differ only by a constant, and so for purposes of inference on $\boldsymbol{\beta}$ are equivalent. McCullagh & Nelder (1989, §9.6) show how the 'likelihood-like' properties of Q for inference on $\boldsymbol{\beta}$, mentioned in §1, extend in an approximate fashion to Q^+ for inference about $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ jointly.

Estimating equations derived from Q^+ by differentiation with respect to $\boldsymbol{\beta}$ are

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i V_i(\mu_i, \boldsymbol{\lambda})} \times \frac{\partial \mu_i}{\partial \beta_r} = 0 \quad (r = 1, \ldots, p),$$

and are, as before, unbiased estimating equations even if the variance specification is incorrect. However, as pointed out by Davidian & Carroll (1988), the corresponding equations obtained by differentiation with respect to $\boldsymbol{\gamma}$ or $\boldsymbol{\lambda}$ are not, in general, unbiased, even if the model assumptions are correct. For example, in the simple case where $\phi_i(\boldsymbol{\gamma}) = \gamma$, a constant, differentiation with respect to γ yields the equation

$$\sum_{i=1}^{n} \{D_\lambda(y_i, \mu_i) - \gamma\} = 0. \tag{12}$$

Typically $E[D_\lambda(Y_i, \mu_i)]$ is approximately, but not exactly, equal to γ, and so each term of the sum in (12) contributes a constant bias to the equation. As a result, γ is not consistently estimated. This example sets the general pattern: except in some special cases, maximization of Q^+ yields inconsistent estimators of parameters $\boldsymbol{\gamma}$ or $\boldsymbol{\lambda}$ in the variance formula. Since consistency is usually regarded as a minimum requirement of estimation procedures, this might be considered a rather serious shortcoming.

An alternative to (12) is to use the so-called *pseudo-likelihood* procedure (Carroll & Ruppert, 1982, 1988; Davidian & Carroll, 1987) in which $D_\lambda(y_i, \mu_i)$ in equations such as (12) is replaced by $(y_i - \mu_i)^2 / V_i(\mu_i; \boldsymbol{\lambda})$. The quadratic estimating equations that result are unbiased, and hence the estimate consistent, provided that the assumed variance specification is correct. Mainly for this reason, pseudo-likelihood has found wider use than EQL in practice.

Lee & Nelder (1992) compare EQL and pseudo-likelihood in simulation experiments. While EQL is inconsistent for the variance parameters, and therefore unsatisfactory as $n \to \infty$, it is found that estimates based on EQL can actually perform better in finite samples than a consistent alternative such as pseudo-likelihood; the effects of bias are outweighed by differences in variance between the two approaches. Further investigation is needed to identify more precisely the conditions under which one or other method is to be preferred on grounds of efficiency, and at what sample size the asymptotic bias of EQL begins to have an appreciable effect.

9. *'Local' quasi-likelihood*

Much attention has recently been given to techniques for nonparametric regression, or 'scatterplot smoothing'. One strand of development has been in algorithmic methods such as kernel averaging, running means, running lines, *etc.*, and the notion of local quasi-likelihood may be seen as a straightforward extension of such methods.

For simplicity, consider the case with a single explanatory variable x and response Y. The aim is to model $E(Y|x)$. A parametric model as in (1) may be fitted to a sample of independent points $\{(x_i, Y_i),\ i = 1, \ldots, n\}$ by solving the quasi-likelihood equations

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \times \frac{\partial \mu_i}{\partial \beta_r} = 0 \qquad (r = 1, \ldots, p) \tag{13}$$

for the unknown parameters $\boldsymbol{\beta}$ in $\mu_i(\boldsymbol{\beta}) = E(Y_i|x_i)$. For simplicity, we re-express (13) as $\sum \mathbf{u}_i(\boldsymbol{\beta}) = 0$. For any specified value of x, a 'local' model with parameter vector $\boldsymbol{\beta}(x)$ may be estimated by solving for $\boldsymbol{\beta}(x)$ the equations

$$\sum_{i=1}^{n} w_i(x)\mathbf{u}_i(\boldsymbol{\beta}(x)) = 0, \tag{14}$$

where the weights $\{w_i(x)\}$ are chosen to be close to zero if x_i is distant from x. The regression function $E(Y|x)$ is then estimated by $\mu(\hat{\boldsymbol{\beta}}(x))$.

A simple example has $\mu(\beta) = x\beta$, $V(\mu) = \mu$, in which the scalar parameter β is the slope of a straight line regression through the origin. Solution of the 'global' quasi-likelihood equation

$$\sum_{i=1}^{n} \frac{y_i - x_i\beta}{x_i\beta} x_i = 0$$

yields $\hat{\beta} = \sum y_i / \sum x_i$, the well-known ratio estimate. The introduction of weights w_i as in (14) results in the 'local' ratio estimate $\hat{\beta}(x) = \{\sum w_i(x)y_i\}/\{\sum w_i(x)x_i\}$ at each value of x, and the corresponding fitted regression curve is $\hat{E}(Y|x) = x\hat{\beta}(x)$.

The weights $\{w_i(x)\}$ may be chosen in a variety of ways. One possibility is to use a kernel function, such as the Gaussian kernel $w_i(x) = f((x_i - x)/b)$ where $f(z) = \exp(-z^2/2)$, in which b is an adjustable bandwidth. Several other forms of weighting are discussed in Hastie & Tibshirani (1990, §2).

Tibshirani & Hastie (1987) introduced the idea of a locally-weighted score equation as in (14). Firth, Glosup & Hinkley (1991) use local quasi-likelihood fitting to test the adequacy of a parametric regression model $\mu(\boldsymbol{\beta})$; the parametric fit $\mu(\hat{\boldsymbol{\beta}})$ is the limit as the bandwidth b tends to infinity of the local regression curves $\mu(\hat{\boldsymbol{\beta}}_b(x))$, and this embedding provides a suitable framework for criticism of $\mu(\boldsymbol{\beta})$ itself. Fan, Heckman & Wand (1992) develop asymptotic theory and discuss bandwidth selection for the class of locally-fitted generalized polynomial models, in which $\mu(\boldsymbol{\beta}) = g^{-1}(\beta_1 + \beta_2 x + \ldots + \beta_p x^p)$ for some specified link function g; local fits based on polynomials of odd degree (linear, cubic,...) are found to have particularly good properties, including reduced bias near the ends of the range of sample x-values. This last finding generalizes the well known result that the edge-effect bias of a kernel-averaging smoother may be improved by using instead kernel-weighted least squares to fit local straight lines. In Severini & Staniswalis (1992), local quasi-likelihood fitting is used in a 'semiparametric' model involving both parametric and nonparametric terms; 'locally constant' fitting of the nonparametric part, based effectively on a polynomial of degree zero, is considered in detail.

Work on these ideas is in its infancy, but it appears that local quasi-likelihood fitting as a generalization of smoothing methods such as kernel averaging and local least squares has promise, and merits further research.

10. *Concluding remarks*

This paper surveys a large body of recent literature on quasi-likelihood methods, but is far from comprehensive. Topics left out, in addition to those already mentioned at the end of §1, include applications in models with covariate measurement error (*e.g.*, Whittemore & Keller, 1988; Carroll & Stefanski, 1990) and work on computational algorithms (*e.g.*, Gay & Welsch, 1988; Osborne, 1992).

Acknowledgements

The author is grateful to A. J. Lee and J. W. McDonald for providing some helpful references.

BIBLIOGRAPHY

Ashby, M., Neuhaus, J.M., Hauck, W.W., Bacchetti, P., Heilbron, D.C., Jewell, N.P., Segal, M.R. and Fusaro, R.E. (1992). An annotated bibliography of methods for analysing correlated categorical data. *Statistics in Medicine* **11**, 67-99.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Review* **51**, 279-292.

Breslow, N.E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38-44.

Breslow, N.E. (1989). Score tests in overdispersed GLMs. In *Statistical Modelling*, eds. A. Decarli, B.J. Francis, R. Gilchrist, G.U.H. Seeber, 64-74. Springer, London.

Breslow, N. (1990a). Further studies in the variability of pock counts. *Statistics in Medicine* **9**, 615-626.

Breslow, N.E. (1990b). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *J. Amer. Statist. Assoc.* **85**, 565-571.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Carroll, R.J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* **10**, 429-441.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman & Hall, London.

Carroll, R.J. and Stefanski, L.A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *J. Amer. Statist. Assoc.* **85**, 652-663.

Chesher, A.D. and Jewitt, I. (1987). The bias of a heteroskedasticity consistent covariance-matrix estimator. *Econometrica* **55**, 1217-1222.

Cox, D.R. (1961). Tests of separate families of hypotheses. *Proc. 4th Berkeley Symposium*, 105-123.

Cox, D.R. (1981). Statistical analysis of time series: some recent developments. *Scand. J. Statist.* **8**, 93-115.

Cox, D.R. (1983). Some remarks on overdispersion. *Biometrika* **70**, 269-274.

Cox, D.R. and Hinkley, D.V. (1968). A note on the efficiency of least-squares estimates. *J. R. Statist. Soc.* B **30**, 284-289.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data* (2nd ed.) Chapman and Hall, London.

Crowder, M.J. (1987). On linear and quadratic estimating functions. *Biometrika* **74**, 591-597.

Davidian, M. and Carroll, R.J. (1987). Variance function estimation. *J. Amer. Statist. Assoc.* **82**, 1079-1091.

Davidian, M. and Carroll, R.J. (1988). A note on extended quasi-likelihood. *J. R. Statist. Soc.* B **50**, 74-82.

Efron, B. (1986a). Double exponential families and their use in generalized linear models. *J. Amer. Statist. Assoc.* **81**, 709-721.

Efron, B. (1986b). Discussion of 'Jackknife, bootstrap and other resampling methods in regression analysis' by C.F.J. Wu. *Ann. Statist.* **14**, 1301-1304.

Engel, J. (1990). Quasi-likelihood inference in a generalized linear mixed model for balanced data. *Statistica Neerlandica* **44**, 221-239.

Engel, J. (1992). Modelling variation in industrial experiments. *Applied Statistics* **41**, 579-593.

Fan, J., Heckman, N.E. and Wand, M.P. (1992). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. Unpublished manuscript.

Firth, D. (1987a). *Quasi-likelihood Estimation: Efficiency and Other Aspects.* Unpublished PhD thesis, University of London.

Firth, D. (1987b). On the efficiency of quasi-likelihood estimation. *Biometrika* **74**, 233-245.

Firth, D. (1988). Multiplicative errors: log-normal or gamma? *J. R. Statist. Soc.* B **50**, 266-268.

Firth, D., Glosup, J. and Hinkley, D.V. (1991). Model checking with nonparametric curves. *Biometrika* **78**, 245-252.

Firth, D. and Harris, I.R. (1991). Quasi-likelihood for multiplicative random effects. *Biometrika* **78**, 545-556.

Ganio, L.M. and Schafer, D.W. (1992). Diagnostics for overdispersion. *J. Amer. Statist. Assoc.* **87**, 795-804.

Gay, D.M. and Welsch, R.E. (1988). Maximum likelihood and quasi-likelihood for nonlinear exponential family regression models. *J. Amer. Statist. Assoc.* **83**, 990-998.

Godambe, V.P. (1991) (ed.) *Estimating Functions.* Oxford University Press.

Godambe, V.P. and Heyde, C.C. (1987). Quasi-likelihood and optimal estimation. *Int. Stat. Rev.* **55**, 231-244.

Godambe, V.P. and Thompson, M.E. (1989). An extension of quasilikelihood estimation (with discussion). *J. Statist. Plan. Inf.* **22**, 137-172.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman & Hall, London.

Heyde, C.C. (1989). Quasi-likelihood and optimality of estimating functions: some current unifying themes. *Bull. Int. Stat. Inst.* **53** (Book 1), 19-29.

Hill, J.R. and Tsai, C.-L. (1988). Calculating the efficiency of maximum quasilikelihood estimation. *Applied Statistics* **37**, 219-230.

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

Hutton, J.E. and Nelson, P.I. (1986). Quasi-likelihood estimation for semimartingales. *Stoch. Processes Appl.* **22**, 633-643.

Lee, A.J., Scott, A.J. and Soo, S.C. (1993). Comparing Liang-Zeger estimates with maximum likelihood in bivariate logistic regression. *J. Statist. Comput. Simul.* **44**, 133-148.

Li, B. (1992). A deviance function in the quasi-likelihood method. Unpublished manuscript, Dept. Statistics, Pennsylvania State University.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Liang, K.-Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *J. R. Statist. Soc.* B **54**, 3-40.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* **78**, 153-160.

MacKinnon, J.G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J. Econometrics* **29**, 305-325.

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

McCullagh, P. (1986). Quasi-likelihood. In *Encyclopedia of Statistical Sciences* **7**, eds. S. Kotz and N.L. Johnson, 464-467. Wiley, New York.

McCullagh, P. (1991). Quasi-likelihood and estimating functions. In *Statistical Theory and Modelling*, eds. D.V. Hinkley, N. Reid & E.J. Snell, 265-286. Chapman and Hall, London.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (2nd Edition). Chapman and Hall, London.

Moore, D.F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika* **73**, 583-588.

Moore, D.F. (1987). Modelling the extraneous variance in the presence of extra-binomial variation. *Applied Statistics* **36**, 8-14.

Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika* **79**, 747-754.

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.

Morton, R. (1987). A generalized linear model with nested strata of extra-Poisson variation. *Biometrika* **74**, 247-257.

Morton, R. (1988). Analysis of generalized linear models with nested strata of variation. *Austral. J. Statist.* **30**A, 215-224.

Morton, R. (1991). Analysis of extra-multinomial data derived from extra-Poisson variables conditional on their total. *Biometrika* **78**, 1-6.

Nelder, J.A. and Lee, Y. (1991). Generalized linear models for the analysis of Taguchi-type experiments. *Appl. Stoch. Models and Data Anal.* **7**, 107-120.

Nelder, J.A. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudo-likelihood: some comparisons. *J. R. Statist. Soc.* B **54**, 273-284.

Nelder, J.A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74**, 221-232.

Osborne, M.R. (1992). Fisher's method of scoring. *Int. Statist. Review* **60**, 99-117.

Pregibon, D. (1983). An alternative covariance estimate for generalised linear models. *GLIM Newsletter* **6**, 51-55.

Pregibon, D. (1984). Review of *Generalized Linear Models*, by P. McCullagh & J. A. Nelder. *Ann. Statist.* **12**, 1589-1596.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1048.

Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825-839.

Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data for complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *J. Amer. Statist. Assoc.* **76**, 221-230.

Rotnitzky, A. and Jewell, N.P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**, 485-497.

Royall, R.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *Int. Statist. Rev.* **54**, 221-226.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719-727.

Scott, A.J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *J. Amer. Statist. Assoc.* **77**, 848-854.

Severini, T.A. and Staniswalis, J.G. (1992). Quasi-likelihood estimation in semiparametric models. Unpublished manuscript.

Sharples, K. and Breslow, N.E. (1992). Regression analysis of correlated binary data: some small sample results for the estimating equation approach. *J. Statist. Comput. Simul.* **42**, 1-20.

Sørensen, M. (1990). On quasi-likelihood for semimartingales. *Stoch. Processes Appl.* **35**, 331-346.

Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657-671.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82**, 559-568.

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

West, M. (1985). Generalized linear models: scale parameters, outlier accommodation and prior distributions. *Bayesian Statistics* **2**, 531-558.

Whittemore, A.S. and Keller, J.B. (1988). Approximations for regression with covariate measurement error. *J. Amer. Statist. Assoc.* **83**, 1057-1066.

Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics* **31**, 144-148.

Williams, E.J. (1959). *Regression Analysis.* Wiley, New York.

Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Ann. Statist.* **14**, 1261-1350.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621-629.

Zeger, S.L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.

Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049-1060.

Zeger, S.L. and Qaqish, B. (1988). Markov regression models for time-series: a quasi-likelihood approach. *Biometrics* **44** , 1019-1031.

SUMMARY

This paper reviews some recent developments and applications of quasi-likelihood functions, as introduced by Wedderburn (*Biometrika* **61**, 1974, 439-47) and McCullagh (*Ann. Statist.* **11**, 1983, 59-67). Topics discussed include optimality, efficiency and robustness of estimation; calculation of estimated standard errors; applications such as overdispersion, multiplicative errors, longitudinal data and other correlated data; and the notions of 'extended' quasi-likelihood and 'local' quasi-likelihood.

RESUME

Cet article fait un exposé des applications et développements récents en matière de fonctions de quasi-vraisemblance, introduites par Wedderburn (*Biometrika* **61**, 1974, 439-47) et McCullagh (*Ann. Statist.* **11**, 1983, 59-67). Nous discutons, entre autres, des sujets suivants: robustesse, efficacité et optimalité de l'estimation de paramètres et calcul d'écart-types estimés. Plusieurs applications sont également présentées pour illustrer les problèmes de données trop dispersées, de données à erreurs multiplicatives, de données longitudinales ou corrélées d'une autre façon. Nous abordons finalement les notions de quasi-vraisemblance étendue et de quasi-vraisemblance locale.