ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Statistics & Probability Letters ▮ (▮▮▮▮) ▮▮▮–▮▮▮

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

A note on quasi-likelihood for exponential families

David H. Annis*,[1]

VP, Forecasting and Pricing, Wachovia Treasury, 201 South College Street NC0572, Charlotte Plaza 17th Floor, Charlotte, NC 28244, USA

Received 24 August 2005; received in revised form 13 July 2006; accepted 7 August 2006

Abstract

Maximum likelihood estimation for exponential families depends exclusively on the first two moments of the data. Recognizing this, Wedderburn [1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika 61, 439–447] proposed estimating regression parameters based on a quasi-likelihood function requiring only the relationship between the mean and variance. We extend quasi-likelihood to situations in which there exists vague prior information on the mean parameters. It is shown when data are exponential family with quadratic variance functions, maximum a posteriori inference under a conjugate prior relies solely on two moments of the data and the prior distribution. This result suggests a Bayesian analog of quasi-likelihood for which only two moments of the data and two moments of the prior need be specified.

Keywords: Quasi-likelihood; Bayesian; Conjugate prior

1. Introduction

Let Y be distributed according to $F = F(\cdot|\theta)$, which belongs to a one-parameter exponential family given by

$$\mathrm{d}F(y|\theta) = f(y|\theta) = h(y)\exp\{y\theta - \psi(\theta)\}, \tag{1}$$

where θ is the *natural parameter* and $\psi(\theta)$ is the cumulant function, from which the central moments of Y can be calculated, in particular $\psi'(\theta) = E(Y) = \mu$ and $\psi''(\theta) = \mathrm{Var}(Y) = V(\mu)$, where the primes denote differentiation with respect to θ. It is well known that for members of the exponential family, the score equations for maximum likelihood estimation rely solely on the first two moments,

$$\frac{\partial l(\theta)}{\partial \mu} = \left\{\left(\frac{\partial \mu}{\partial \theta}\right)^{-1}[y - \psi'(\theta)]\right\} = \{V(\mu)^{-1}(y-\mu)\} = 0.$$

*Fax: +1 704 383 3878.

E-mail address: david.annis@wachovia.com.

[1] A portion of this work was completed while the author was on the Operations Research faculty at the Naval Postgraduate School in Monterey, CA, USA.

0167-7152/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2006.08.011

This observation prompted Wedderburn (1974) to propose quasi-score equations, which require only specification of the mean and variance function,

$$\frac{\partial K}{\partial \mu} = \frac{y-\mu}{V(\mu)} = 0; \quad V(\mu) > 0. \tag{2}$$

He notes that K has properties similar to those of a log-likelihood function and is, in fact, a log-likelihood when Y belongs to an exponential family. We propose a similar set of equations for maximum quasi-posterior estimation when there is vague prior information about the mean parameter, μ. By analogy, define

$$\frac{\partial Q}{\partial \mu} = \frac{z-\mu}{V(\mu)}; \quad V(\mu) > 0. \tag{3}$$

It will be shown that for suitably chosen z, Q has properties similar to those of a log-posterior.

2. Conjugate analysis for exponential families

Diaconis and Ylvisaker (1979) characterize a proper conjugate prior for the natural parameter of an exponential family,

$$\pi(\theta|s,m) \propto \exp\{s\theta - m\psi(\theta)\}, \quad \theta \in \Theta; \ m \in \mathbb{R}_+; \ s \in \Omega_m, \tag{4}$$

where Ω is the support of the mean parameter, μ, and $\Omega_m = \{s : (s/m) \in \Omega\}$. It is assumed that $\pi(\cdot|s,m)$ is a continuous, proper density on Θ for all pairs $(s,m) \in (\Omega_m \times \mathbb{R}_+)$. Since there exists a one-to-one correspondence between $\theta \in \Theta$ and $\mu \in \Omega$, the prior distribution $\pi(\theta)$ induces a prior distribution $\pi^*(\cdot)$ on μ,

$$\pi^*(\mu|s,m) \propto \exp\{s\theta(\mu) - m\psi(\theta(\mu))\}\left|\frac{\partial\mu}{\partial\theta}\right|^{-1}, \quad \theta \in \Theta; \ m \in \mathbb{R}_+; \ s \in \Omega_m. \tag{5}$$

In general, π^* is not conjugate to the likelihood (1). However, Consonni and Veronese (1992) demonstrate that when Y has a quadratic variance function (QVF), $V(\mu) = a\mu^2 + b\mu + c$, $\partial\mu/\partial\theta \propto \exp\{b\theta(\mu) + 2a\psi(\theta(\mu))\}$, and therefore $\pi^*(\mu)$ is also conjugate. Morris (1982) characterizes QVF exponential families, and shows this class of distributions to be quite broad, encompassing, among others, the normal, poisson and binomial distributions.

Diaconis and Ylvisaker (1979) show that for conjugate priors of the form (4), the prior mean of μ is s/m. Their proof relies on the fact that $\int_\Theta \pi'(\theta)\,\mathrm{d}\theta = 0$. For Y belonging to a QVF exponential family, Morris (1983) shows

$$E_\pi[h(\mu)(\mu - (s/m))] = \frac{1}{m} E_\pi[h'(\mu)V(\mu)]$$

for $h(\mu)$ with continuous first derivative $h'(\mu)$. Setting $h(\mu) = [\mu - (s/m)]$, $\mathrm{Var}_\pi(\mu) = V(s/m)/(m-a)$, resulting in an expression for the prior variance. Thus, when Y follows a QVF exponential family, any combination of prior mean ($\lambda \in \Omega$) and variance ($\tau^2 \in \mathbb{R}_+$) can be achieved by selecting $m = a + V(\lambda)/\tau^2$ and $s = m\lambda$.

Since $\pi(\cdot)$ is conjugate to F, closed form expressions for the prior mean and variance imply the existence of closed form expressions for the same posterior quantities. Consider a simple random sample $Y_1, Y_2, \ldots, Y_n$ from a QVF exponential family whose realizations are $y_1, y_2, \ldots, y_n$, respectively. Then letting $\bar{y}$ denote the arithmetic mean of the observations, the posterior distribution of θ can be written, trivially, as $\pi(\theta|s,m,y_1,\ldots,y_n) \propto \exp\{(s+n\bar{y})\theta - (m+n)\psi(\theta)\}$, which is maximized as a function of μ when

$$\begin{aligned} \frac{\partial}{\partial\mu}\log \pi(\theta|s,m,y_1,\ldots,y_n) &= \sum\left\{\left(\frac{\partial\mu}{\partial\theta}\right)^{-1}\left[\left(\frac{s}{n} + y_i\right) - \left(\frac{m}{n} + 1\right)\psi'(\theta)\right]\right\} \\ &= \left(\frac{m+n}{n}\right)V(\mu)^{-1}\sum\{(z_i - \mu)\} = 0 \end{aligned} \tag{6}$$

and is proportional to (3) with $z_i = (s + ny_i)/(m+n)$. Maximizing (6) requires only knowledge of the first two moments of the data, μ and $V(\mu)$, and the first two moments of the prior, which are known functions of the hyperparameters s and m. Note that unlike its quasi-likelihood analog, (3) does *not* have expectation zero (with respect to either the conditional distribution of Y given μ or the posterior of μ given Y).

Jackson et al. (1970) show that for exponential families, the conjugate prior is least favorable with respect to quadratic loss among all priors for a specified mean and variance. Thus, the resulting posterior mean is the Bayes rule and, therefore, minimax under squared error loss. Therefore, using the posterior distribution of θ as the objective function of μ is appealing since it shrinks the data toward the posterior *mean* of μ. Maximizing the posterior distribution of μ directly, by updating (5), shrinks toward the posterior *mode*.

3. Application to generalized linear models

In this section, we extend the conjugate analysis of the previous section to the weighted generalized linear model (WGLM) framework in which the mean parameters of the observations are functions of known covariates. To this end, suppose $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ are independent, but no longer identically distributed, and for each $i \in 1, 2, \ldots, n$, let

$$f(y_i|\theta_i, k_i) = [h(y_i)\exp\{y_i\theta_i - \psi(\theta_i)\}]^{k_i} = f(y_i|\theta_i)^{k_i},$$

where $g(\mu) = g(\psi'(\theta)) = X\beta = \eta$ for a known monotonic function, $g(\cdot)$, and $k_1, k_2, \ldots, k_n$ are known weights associated with the observations $y_1, y_2, \ldots, y_n$, respectively. Often the weights are taken to be $k_1 = k_2 = \cdots = k_n = 1$, however, this is not necessary.

In the generalized linear model framework, a natural extension of the conjugate prior (4) is one which places an independent prior (depending on s_i and m_i) on each θ_i and views their product as a function of the regression parameters β,

$$p(\beta|s, m) = \prod_{i=1}^{n} \exp\{s_i\theta_i(\beta) - m_i\psi(\theta_i(\beta))\}, \quad \theta \in \Theta^n; \;\; m \in \mathbb{R}_+^n; \;\; s \in \Omega_m^n. \tag{7}$$

Because (7) is conjugate to the exponential family, the hyperparameters can be chosen to achieve certain coarse characteristics. Since m can be interpreted as a prior precision, large values of m cause (7) to be concentrated about the prior mode while small ones result in a more diffuse set of prior beliefs. Further, for each term in the product m and s are chosen to satisfy $E(\mu_i) = s_i/m_i$ and $\mathrm{Var}(\mu_i) = V(s_i/m_i)/(m_i - a)$, where $V(\mu) = a\mu^2 + b\mu + c$ is the QVF for y_i. Furthermore, certain cases—notably $V(\mu) = c$ (corresponding to the normal distribution)—terms in (7) are symmetric, though in general they are not. Because (7) is least favorable (Jackson et al., 1970), the Bayes estimate is minimax, and, therefore, (7) is a conservative choice when only two prior moments are known.

Chen and Ibrahim (2003) propose a similar prior where $m_1 = m_2 = \cdots = m_n$. The Chen–Ibrahim prior is appropriate when the prior confidence in the mean response is roughly constant for all values of the covariates. By allowing for separate m_i, however, (7) can accommodate situations when there is more certainty for some regions in the design space than others. The log-posterior distribution of θ (viewed as an objective function of β) is maximized when

$$\begin{aligned}\frac{\partial}{\partial\beta}\log\pi(\theta(\beta)|s, m, y, k) &= \frac{\partial}{\partial\beta}\sum \log\pi(\theta_i(\beta)|s_i, m_i, y_i, k_i) \\ &= \sum \frac{\partial\mu_i}{\partial\beta}\left[\frac{\partial\mu_i}{\partial\theta_i}\right]^{-1}\frac{\partial}{\partial\theta_i}\log\pi(\theta_i(\beta)|s_i, m_i, y_i, k_i) \\ &= \sum d_i^{\mathrm{T}} v_i^{-1}(m_i + k_i)\left(\frac{s_i + y_i}{m_i + k_i} - \mu_i\right) \\ &= \sum d_i^{\mathrm{T}} v_i^{-1} w_i(z_i - \mu_i) = D^{\mathrm{T}} V^{-1} W(z - \mu) = 0,\end{aligned} \tag{8}$$

where $d_i = \partial\mu_i/\partial\beta$, $v_i = V(\mu_i) = \mathrm{Var}(y_i)$, $w_i = (m_i + k_i)$ and $z_i = (s_i + y_i)/(m_i + k_i)$. When expressed in matrix form, $D = [\partial\mu/\partial\beta]$ is a matrix of derivatives of the means with respect to the model parameters, $V = \mathrm{Cov}(Y)$ is a diagonal covariance matrix and $W = \mathrm{diag}\{w_i\}$ is a diagonal weight matrix.

Please cite this article as: David H. Annis, A note on quasi-likelihood for exponential families, Statistics & Probability Letters (2006), doi:10.1016/j.spl.2006.08.011

The posterior score equations (8) resemble Wedderburn's quasi-score equations (2)—the difference being that the stochastic elements are the means μ rather than the data y. Consequently, posterior maximization requires knowledge of only the first two moments of the data and the prior. Note that when the data are, in fact, i.i.d. $\mu_i = \beta_0$ and (8) is a scalar multiple of (3).

4. Quasi-posterior analysis

We have shown that for exponential family data with QVF, the posterior distribution resulting from updating a conjugate prior on the natural parameter can be maximized with knowledge of only the first two moments of the data and of the prior. This suggests an extension to quasi-likelihood estimation when there exists vague prior information of the mean parameters (in the form of a prior mean and variance). Analogous to standard quasi-likelihood methods, mean parameters associated with non-diagonal covariance structures can still be estimated, realizing, though, that the choice of non-diagonal V matrix may not correspond to a proper likelihood-prior structure. This is to be expected as quasi-likelihood based on an arbitrary non-diagonal covariance matrix does not necessarily correspond to a proper likelihood.

Quasi-posterior maximization proceeds via Newton's method. Define $\partial \log \pi(\theta|s,m,y,k)/\partial\beta = \Xi(\beta) = 0$ as estimating equations for β. Then given a current estimate of β, say $\beta^{(r)}$, an iterative update satisfies $\beta^{(r+1)} = \beta^{(r)} - [\Xi'(\beta^{(r)})]^{-1}\Xi(\beta^{(r)})$, where $\Xi'(\beta) = -D^{\mathrm{T}}V^{-1}WD$. Note that the posterior distribution of θ (viewed as a function of β) is functionally identical to the likelihood of a WGLM with data $(s/m, y)$ and weights (m, k). Therefore, the maximum quasi-posterior estimate $\tilde{\beta}$ behaves as would a maximum likelihood estimate for the equivalent WGLM. Thus, the same asymptotic arguments apply and

$$\pi(\theta(\tilde{\beta})|s,m,y,k) \to N_p(\beta,(D^{\mathrm{T}}V^{-1}WD)^{-1}). \tag{9}$$

Unlike many situations, the effect of the prior distribution does not necessarily vanish as $n \to \infty$. This potential arises because each new observation coincides with a mean parameter on which a prior distribution may be placed. Thus, the information contained in the prior can increase with that of the sample. This can be remedied by limiting the total prior precision—for example, by constraining $\sum m_i$. For fixed $\sum m_i$ and $k_i = 1\ \forall i$, W converges to the identity matrix and the limiting variance–covariance matrix in (9) is the familiar $(D^{\mathrm{T}}V^{-1}D)^{-1}$.

5. Example

Prentice (1976) examines mortality of adult flour beetles in the five hours following exposure to varying doses of gaseous carbon disulfide (originally in Bliss, 1935). Dosages are in units of $\log_{10}\mathrm{CS}_2$, and raw data are given in Table 1. Though Prentice investigates the influence of the link function, for ease of exposition we use the complementary log–log (which accounts for the asymmetry observed by Prentice) and differs insignificantly from his best fit.

Because the binomial model depends on only one parameter (the success probability, μ) the variance is uniquely determined by this mean parameter. Often, however, observed variance differs from the prescribed model. If Y_i denotes the number of beetles killed at dosage level i, a natural choice of weights, k, is the number of insects exposed at each level. Absent overdispersion Y_i is a binomial random variable with size k_i and success probability μ_i. To allow for potential extra-binomial variation, an overdispersed binomial

Table 1
Mortality of adult flour beetles after five hours exposure to gaseous CS_2

Dosage	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
Insects	59	60	62	56	63	59	62	60
Killed	6	13	18	28	52	53	61	60

Please cite this article as: David H. Annis, A note on quasi-likelihood for exponential families, Statistics & Probability Letters (2006), doi:10.1016/j.spl.2006.08.011

quasi-likelihood is specified:

$$E(Y_i|\theta_i,k_i)=k_i\frac{\mathrm{e}^{\theta_i}}{1+\mathrm{e}^{\theta_i}}=k_i\mu_i,\quad \mathrm{Var}(Y_i|\theta_i,k_i)=\phi k_i\frac{\mathrm{e}^{\theta_i}}{(1+\mathrm{e}^{\theta_i})^2}=V(\mu_i)=\phi k_i\mu_i(1-\mu_i).$$

Interestingly, a dispersion parameter of $\phi=0.5$ (underdispersion relative to the binomial model) is consistent with these data, suggesting some degree of negative correlation between beetles subject to the same dosages.

It is reasonable to assume that the particular dosages used in the experiment were chosen to coincide with all levels of mortality. This is borne out by the data, as small numbers of beetles die (roughly 10%) at the lowest dosage while nearly all die at the high dosage.

By allowing different m_i, the prior precision can vary with the covariate. Furthermore, letting $k_i=0$ and $m_i>0$ allows specification of prior information about the response associated with covariates not present in the data. In particular, since a flour beetle may have a lifespan of two years, a beetle is unlikely to die during a five-hour experiment (in the absence of toxins). Letting $\lambda_0=s_0/m_0\approx 0$ implies the prior estimate of the mortality when the dosage is zero is virtually nil. Likewise, for a large enough dosage, say 3, virtually all beetles exposed would die. This suggests $\lambda_3=s_3/m_3\approx 1$. For moderate values, one might assume that the probability of death is increasing in dosage and that the mortality rates corresponding to these particular dosages are linear. Thus $\lambda_{1.6907}=(1-0.5)/8$; $\lambda_{1.7242}=(2-0.5)/8;\ldots;\lambda_{1.8839}=(8-0.5)/8$ are plausible. The form of (8) suggests interpreting m_i as the equivalent weight or sample size corresponding to the prior confidence in λ_i. For instance, $m_0=20$ implies the confidence in the prior belief $\lambda_0=0$ is equivalent to having observed a sample of 20 beetles of which none were killed. For the interior design points $m_{1.6907}=m_{1.7242}=\cdots=m_{1.8839}=4$ implies the confidence in the assumed linearity is equal to what it would be had we observed the same pattern for samples of four beetles at each dosage level. Choosing four hypothetical observations for each of the eight dosages ensures that the total prior confidence ($4\times 8=32$) is roughly half of the observed sample size for any particular dosage (≈ 60). Finally, since there may be more confidence in the prior mean when the dosage is 3 than when it is between 1.69 and 1.88 but less than when it is 0, $m_3=10$ indicates the belief is about as strong as if there had been a sample of 10 beetles at this dosage, all of which died. For subsequent comparison, this set of prior beliefs will be referred to as “linear.”

Another equally reasonable prior belief is approximate non-informativeness over the range of experimental dosages. Specifically, one might assume the same behavior for the extreme doses but $\lambda_{1.6907}=0.5$; $\lambda_{1.7242}=0.5;\ldots;\lambda_{1.8839}=0.5$. Here, we may specify a prior variance of $\tau^2=\frac{1}{16}$ so a Wald-type prior confidence interval for the mortality, $\lambda\pm 2\tau$ is $(0,1)$. Together $\phi=0.5$ and $\tau^2=\frac{1}{16}$ require $m_{1.6907}=m_{1.7242}=\cdots=m_{1.8839}=3.5$. For subsequent comparison, this set of prior beliefs will be referred to as “flat.”

Elicitation of prior information from experts hinges on specifying only a mean and an approximate sample size which indirectly expresses their estimate of the prior variance. Of course, in the case where the prior mean and variance can be elicited directly, m_i and s_i are immediate consequences of these specifications. Table 2 gives the two proposed choices of quasi-priors (linear and flat) and compares fits of the maximum quasi-likelihood and quasi-posterior analyses. A plot of estimated mortality curves is given in Fig. 1. Observed empirical mortalities for each dosage are marked using “+” symbols, and implied quasi-prior mortality curves are given for each set of assumptions.

Table 2
Comparison of fitted values of number of beetles killed at each dosage for maximum quasi-likelihood (QL) and quasi-posterior (QP) for two different quasi-priors

Dosage	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
Insects (n_i)	59	60	62	56	63	59	62	60
Killed ($n_i\bar{y}_{i.}$)	6	13	18	28	52	53	61	60
QL	5.59	11.28	20.95	30.37	47.78	54.14	61.11	59.95
QP (linear)	5.93	11.56	20.87	29.68	46.41	52.99	60.60	59.87
QP (flat)	6.74	12.58	21.83	30.07	46.09	52.34	60.15	59.76

Please cite this article as: David H. Annis, A note on quasi-likelihood for exponential families, Statistics & Probability Letters (2006), doi:10.1016/j.spl.2006.08.011

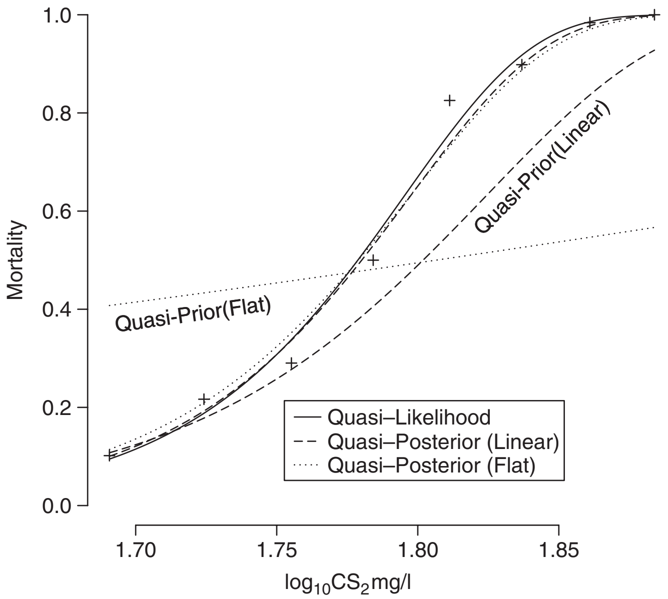

Fig. 1. Comparison of fitted values and prior beliefs for maximum quasi-likelihood (QL) and quasi-posterior (QP) methods.

The results suggest that when the m_i are small relative to the sample size, analysis is robust to choice of the prior beliefs. Even if the prior beliefs are patently false, provided $\sum m_i / \sum k_i \to 0$, the effect of the prior is abated, allowing the true relationship to emerge. In this sense, the proposed method may be preferable to explicitly constrained quasi-likelihood estimation (Heyde and Morton, 1993) in some cases.

6. Conclusion

In many areas of applied statistics it is sufficient to consider only two moments. Even in many Bayesian contexts, the posterior mean and variance (or estimates thereof, in the case of Markov chain Monte Carlo methods) are given as summary measures in lieu of a full posterior distribution. This is especially true of estimating values for model coefficients (as compared with population parameters) where the ultimate goal is the model's predictive output, not quantiles for the model's coefficients. In such cases, prior information can still be incorporated using only two moments rather than a complete Bayesian specification.

Specifically, when the data belong to an exponential family with QVF, conjugate analysis yields posterior estimates which rely on only the first two moments of the data and the prior. This leads to a Bayesian analog to quasi-likelihood estimation, appropriate when complete specification of the likelihood and prior are infeasible.

As is expected, the ability to calculate these quasi-posterior estimates easily and without a completely specified model comes at a price. Since this is not a fully Bayesian procedure, additional quantities, such as posterior probabilities and credible sets cannot be determined directly.

References

Bliss, C.I., 1935. The calculation of the dosage–mortality curve. Ann. Appl. Biol. 22, 134–167.

Chen, M.-H., Ibrahim, J.G., 2003. Conjugate priors for generalized linear models. Statist. Sinica 13, 461–476.

Consonni, G., Veronese, P., 1992. Conjugate priors for exponential families having quadratic variance functions. J. Amer. Statist. Assoc. 87, 1123–1127.

Diaconis, P., Ylvisaker, D., 1979. Conjugate priors for exponential families. Ann. Statist. 7, 269–281.

Heyde, C.C., Morton, R., 1993. On constrained quasi-likelihood estimation. Biometrika 80, 755–761.

Please cite this article as: David H. Annis, A note on quasi-likelihood for exponential families, Statistics & Probability Letters (2006), doi:10.1016/j.spl.2006.08.011

Jackson, D.A., O'Donovan, T.M., Zimmer, W.J., Deely, J.J., 1970. G2 minimax estimators in the exponential family. Biometrika 57, 439–443.

Morris, C.N., 1982. Natural exponential families with quadratic variance functions. Ann. Statist. 10, 65–80.

Morris, C.N., 1983. Natural exponential families with quadratic variance functions: statistical theory. Ann. Statist. 11, 515–529.

Prentice, R.L., 1976. A generalization of the probit and logit methods for dose response curves. Biometrics 32, 761–768.

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

Please cite this article as: David H. Annis, A note on quasi-likelihood for exponential families, Statistics & Probability Letters (2006), doi:10.1016/j.spl.2006.08.011