Statistica Sinica 4(1994), 199-218

EMPIRICAL LIKELIHOOD FOR GENERALIZED LINEAR MODELS

Eric D. Kolaczyk

Stanford University

Abstract: We show that empirical likelihood is justified as a method of inference for a class of models much larger than the class of linear models considered by Owen (1991). In particular, we show how empirical likelihood may be used with generalized linear models. Quasi-likelihood and extended quasi-likelihood are used to derive the necessary estimating functions, but the method can be applied similarly using other sources. We consider separately those models in which the dispersion parameter is fixed and known, those in which it is fixed but unknown, and those in which it is itself modeled linearly via a link function.

Key words and phrases: Dispersion modeling, empirical likelihood, estimating functions, extended quasi-likelihood, generalized linear models, quasi-likelihood.

1. Introduction

Empirical likelihood is a nonparametric method of inference with sampling properties similar to those of the bootstrap, but instead of resampling it works by profiling a multinomial likelihood supported on the sample. In settings with independent and identically distributed random variables, its properties have been examined by Owen (1988, 1990), Hall (1990), DiCiccio, Hall, and Romano (1991), and Qin and Lawless (1991). Recently, Owen (1991) has extended the applicability of empirical likelihood to the context of linear models. In this paper we show that the results of Owen (1991) imply that empirical likelihood is justified for use in a much larger class of models. We focus, in particular, on how it may by used for generalized linear models (GLMs).

Specifically, let data $(Y_1, X_1), \ldots, (Y_n, X_n)$ be observed, where $Y_i \in \mathbb{R}$ are independent random variables and $X_i \in \mathbb{R}^p$ are fixed covariates. A GLM models the Y_i in terms of the X_i by specifying that

$$E(Y_i) = \mu_i, \quad g(\mu_i) = X_i'\theta, \quad \text{var}(Y_i) = \phi V(\mu_i), \tag{1.1}$$

where $g(\cdot)$ and $V(\cdot)$ are real-valued link and variance functions respectively, and $\theta \in \mathbb{R}^p$. If it is thought appropriate to model the dispersion, ϕ, as well as the mean, a second level is specified for the model, i.e.

$$E(d_i) = \phi_i, \quad h(\phi_i) = U_i'\gamma, \quad \text{var}(d_i) = \tau V_D(\phi_i). \tag{1.2}$$

Here $d_i \equiv d(Y_i; \mu_i)$ is some statistic that measures dispersion, $h(\cdot)$ and $V_D(\cdot)$ are real-valued link and variance functions for the dispersion, and $\gamma \in \mathbb{R}^q$. The $U_i \in \mathbb{R}^q$ are covariates for the dispersion and are often a subset of the X_i.

Not surprisingly, it becomes more difficult to develop parametric methods of inference for GLMs as the complexity of the models increases and/or the number of distributional assumptions decreases. In the original form in which GLMs were introduced by Nelder and Wedderburn (1972), where the Y_i are assumed to have an exponential family distribution, approximate confidence regions can be found for θ by using a normal approximation to the distribution of the maximum likelihood estimate $\hat{\theta}$. Wedderburn's (1974) introduction of quasi-likelihood showed that the distributional assumption on the Y_i can be replaced by a weaker one on the form of their mean and variance, and still have normal theory confidence regions justified. The extended quasi-likelihood of Nelder and Pregibon (1987) is a natural generalization of quasi-likelihood which enables one to estimate or model the dispersion ϕ or nonlinear parameters in the variance $V(\cdot)$. However, first and second moment assumptions are insufficient to justify typical parametric methods of inference in this context and we are forced to make assumptions about the forms of higher moments. Davidian and Carroll (1988) examine a suggestion by Nelder and Pregibon (1987) to treat the extended quasi-likelihood as an actual likelihood and approximate its distribution by that of a chi-square. They conclude that while such an approach is valid for exponential and near-exponential family distributions, it may be misleading in many other cases. For the combined model of (1.1) and (1.2), Smyth (1989) asserts that extended quasi-likelihood estimates of γ have an asymptotic normal distribution, but there is no usable variance estimator for these estimates unless the form of $V_D(\cdot)$ is specified. Efron (1986) and Jørgensen (1987) grapple with these types of problems in methods similar to extended quasi-likelihood. Some methods of non-parametric inference have been investigated by Simonoff and Tsai (1988) for use with quasi-likelihood functions. They find that application of the bootstrap is not at all straightforward and that it seems to suffer poor performance as a result. Several jackknife-based estimators are proposed, and some are found, in simulations, to have certain robustness qualities.

Empirical likelihood has an advantage over some parametric methods of inference for GLM's in that it makes only mild assumptions on the existence of certain moments. The number of moments depends, to some degree, on the complexity of the statistics being used. Although numerous methods have been suggested for estimating θ and ϕ or γ, quasi-likelihood and extended quasi-likelihood lend themselves particularly well to empirical likelihood because of their simplicity and their semi-parametric character. For this reason, our results show how to use empirical likelihood mainly with estimating functions derived from these

two methods. However, in principal, empirical likelihood can give estimates and confidence regions with almost any method that yields a "reasonable" set of estimating functions. By "reasonable" we mean that the estimating functions have zero mean, finite variance, and either are based on independent and identically distributed sampling or have higher order moments which allow the use of a triangular array argument. In Section 2, a brief introduction is given to empirical likelihood. Section 3 discusses why empirical likelihood confidence regions are justified for a general class of models which estimate parameters via a set of estimating equations. In Section 4, it is shown, explicitly, how empirical likelihood can be used with GLMs. Three situations are examined: those models in which the dispersion parameter is fixed and known, those in which the dispersion is fixed but unknown, and those in which the dispersion itself is also modeled linearly via a link function. Some examples are given in Section 5. Section 6 contains concluding remarks.

2. Introduction to Empirical Likelihood

Let $X_1, X_2, \ldots$ be independent and identically distributed random vectors in $\mathbb{R}^p$ with distribution function F_0. Based on a sample of size n from F_0, the empirical distribution

$$F_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} I(\cdot \in \{x_i\})$$

is well known to be the nonparametric maximum likelihood estimate of F_0. The function that it maximizes is the empirical likelihood function,

$$L(F) = \prod_{i=1}^{n} F\{x_i\},$$

where $F\{x_i\}$ is the probability, under F, of the set $\{x_i\}$ and x_i is the observed value of X_i. Analogous to the parametric case, the empirical likelihood ratio function is defined by

$$R(F) = \frac{L(F)}{L(F_n)}.$$

If the x_i are all distinct we may write $R(F) = \prod np_i$, where $p_i = F\{x_i\}$. In the case where the x_i are not all distinct, Owen (1988) shows that this expression for $R(\cdot)$ is still appropriate, but with the modification that $\sum_{j:x_j=x_i} p_j = F\{x_i\}$. In other words, ties among the data do not affect this natural re-expression of the likelihood ratio.

Suppose that F_0 has mean $\mu_0 = (\mu_0^1, \ldots, \mu_0^p) \in \mathbb{R}^p$ and variance V_0 of full-rank. In order to form an empirical likelihood confidence region for μ_0, we define

the profile empirical likelihood ratio function

$$\Re(\mu) = \sup\left\{ \prod_{i=1}^{n} np_i \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i x_i = \mu \right\}.$$

Owen (1990) shows that $-2\log\Re(\mu) \rightarrow \chi^2_{(p)}$ in distribution as $n \rightarrow \infty$, which is analogous to the parametric case shown by Wilks (1938). Therefore, to construct an approximate $(1-\alpha)$-level confidence region for μ_0 one computes the set

$$C_{\mu_0} = \{\mu \in \mathbb{R}^p | -2\log\Re(\mu) \leq c_\alpha\},$$

where c_α is defined such that $P(\chi^2_{(p)} \leq c_\alpha) = 1-\alpha$.

A discussion of the computation of $\Re(\mu)$ can be found in Owen (1990, Section 3). The problem of maximizing $\prod np_i$, subject to the constraints $p_i \geq 0$, $\sum p_i = 1$, and $\sum p_i x_i = \mu$, is shown, using a Lagrange multiplier argument, to be equivalent to minimizing the expression $-\sum\log(1+\lambda'(x_i-\mu))$ over $\lambda \equiv \lambda(\mu) \in \mathbb{R}^p$, when μ is in the convex hull of the data, i.e. $ch(\{x_1,\ldots,x_n\})$. This alternative version of the problem is the convex dual of the original. Instead of attempting to solve a constrained maximization problem, the researcher is faced with the much easier task of finding the unconstrained minimum of a convex function, a problem for which many algorithms exist.

3. Empirical Likelihood and Modeling

For independent and identically distributed random variables, Owen (1990) extends the applicability of empirical likelihood beyond single functionals by justifying its confidence regions for functions of several means, Frechet differentiable functions, and M-estimates. Qin and Lawless (1991) show that empirical likelihood may be used in this context to perform inference on a parameter which is estimated by the solution of a general set of possibly nonlinear estimating equations. However, it is from the following result of Owen (1991) that justification is derived for the use of empirical likelihood with data which are independent but not necessarily identically distributed.

Theorem 1. *Empirical Likelihood for Triangular Arrays. Let $Z_{in} \in \mathbb{R}^p$, for $1 \leq i \leq n$ and $p \leq n < \infty$, be a set of random vectors such that $Z_{1n},\ldots,Z_{nn}$ are independent for each n. Suppose that $E(Z_{in}) = \mu_n$, $\mathrm{var}(Z_{in}) = V_{in}$, and define $V_n = \frac{1}{n}\sum_{i=1}^n V_{in}$, $\sigma_{1n} = \mathrm{maxeig}(V_n)$, and $\sigma_{pn} = \mathrm{mineig}(V_n)$.*

Assume that the following three conditions hold.

$$P\Big(\mu_n \in ch(\{Z_{1n},\ldots,Z_{nn}\})\Big) \rightarrow 1 \quad as \quad n \rightarrow \infty \tag{3.1a}$$

$$n^{-2}\sum_{i=1}^{n} E\Big(\|Z_{in}-\mu_n\|^4 \sigma_{1n}^{-2}\Big) \rightarrow 0 \quad as \quad n \rightarrow \infty \tag{3.1b}$$

$$For\ some\ c > 0\ and\ all\ n \geq p,\ \sigma_{pn}/\sigma_{1n} \geq c. \tag{3.1c}$$

Then $-2\log\Re(\mu_n) \to \chi^2_{(p)}$ *in distribution as* $n \to \infty$, *where*

$$\Re(\mu) = \sup\left\{\prod_{i=1}^{n} np_i \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i(Z_{in} - \mu) = 0\right\}.$$

To see how we may use this theorem for modeling, assume that we have pairs $(Y_1, X_1), \ldots, (Y_n, X_n)$ where the Y_i are independent real random variables and the $X_i \in \mathbb{R}^p$ are fixed covariates. In addition, suppose that we have a vector of real functions $f(W_i;\theta)$ such that

$$f(W_i;\theta) = \left\{f_1(W_i;\theta), \ldots, f_p(W_i;\theta)\right\}', \quad \text{and} \quad E\{f(W_i;\theta)\} = 0, \tag{3.2}$$

where $W_i = (Y_i, X_i)$ and $\theta \in \mathbb{R}^p$ is to be estimated. It follows immediately by Theorem 1, with $Z_{in} = f(W_{in};\theta)$, and the assumption that $E\{f(W_{in};\theta)\} = 0$ under the true θ, that $-2\log\Re(\theta) \to \chi^2_{(p)}$ in distribution as $n \to \infty$, where

$$\Re(\theta) = \sup\left\{\prod_{i=1}^{n} np_i \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i f(W_{in};\theta) = 0\right\}. \tag{3.3}$$

Note that if we define $\Psi_{in} = \text{var}\{f(W_{in};\theta)\}$, $\Psi_n = \frac{1}{n}\sum_{i=1}^{n}\Psi_{in}$, $\psi_{1n} = \text{maxeig}(\Psi_n)$, and $\psi_{pn} = \text{mineig}(\Psi_n)$, conditions (3.1a), (3.1b), and (3.1c) are replaced by the following.

$$P\Big(0 \in ch[\{f(W_{1n};\theta), \ldots, f(W_{nn};\theta)\}]\Big) \to 1 \quad \text{as} \quad n \to \infty \tag{3.1a$'$}$$

$$n^{-2}\sum_{i=1}^{n} E\left\{\|f(W_{in};\theta)\|^4 \psi_{1n}^{-2}\right\} \to 0 \quad \text{as} \quad n \to \infty \tag{3.1b$'$}$$

$$\text{For some } c' > 0 \text{ and all } n \geq p,\ \psi_{pn}/\psi_{1n} \geq c'. \tag{3.1c$'$}$$

Although we restrict our attention to generalized linear models in this paper, it is clear that Theorem 1 implies that empirical likelihood should be valid over a large class of linear, nonlinear, and semi-parametric models. For a given set of data, if the hypothesized model admits a set of estimating functions of the form (3.2) which satisfy (3.1a$'$), (3.1b$'$), and (3.1c$'$), then the use of empirical likelihood confidence regions is theoretically justifiable. Further research needs to be done examining the application and performance of empirical likelihood for specific classes of models.

4. Generalized Linear Models

The class of generalized linear models arises through a natural generalization of the ideas behind classical linear regression. Hence, this class of models would

seem a logical area in which to begin applying our extension of Theorem 1. In this section, we consider, in sequence, the cases where the dispersion ϕ is fixed and known, fixed but unknown, and itself modeled as in (1.2). As mentioned above, we use the methods of quasi-likelihood and extended quasi-likelihood to derive estimating functions for empirical likelihood. Note that, since maximum likelihood and quasi-likelihood coincide for all linear exponential families, the class of estimating functions derived using the former method automatically falls within that of the latter. See Morris (1982) and Nelder & Pregibon (1987).

4.1. GLMs with fixed, known dispersion

Suppose for the moment that we are working with Model (1.1) in which the dispersion parameter ϕ is fixed and known. The quasi-likelihood (or, more precisely, log quasi-likelihood) for Y_i is defined to be

$$Q(\mu_i; Y_i) = \int_{Y_i}^{\mu_i} \frac{Y_i - t}{\phi V(t)} dt,$$

when this integral exists, and hence, by independence, the quasi-likelihood for the complete set of data is

$$Q(\mu; \boldsymbol{Y}) = \sum_{i=1}^{n} Q(\mu_i; Y_i).$$

Differentiating with respect to θ and substituting from (1.1), the quasi-score function may be written as

$$\frac{\partial}{\partial \theta} Q(\mu_i; Y_i) = \frac{Y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \theta}, \tag{4.1.1}$$

where, by definition, $\mu_i = g^{-1}(X_i'\theta)$. Note that, if $V(\mu)$ is the variance function of an exponential family, then (4.1.1) will simply be the maximum likelihood score function.

Since $E\{\frac{\partial}{\partial \theta} Q(\mu_i; Y_i)\} = 0$, we take as the $f(W_i; \theta)$ of (3.2)

$$Z_i = Z\{(Y_i, X_i); \theta\} = \frac{Y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \theta} \in \mathbb{R}^p. \tag{4.1.2}$$

Table 1 gives some examples of such Z_i. Using the results following Theorem 1, the maximum empirical likelihood estimate (MELE) of θ is given by

$$\tilde{\theta} = \arg \max_{\theta \in \mathbb{R}^p} \Re(\theta),$$

and an approximate confidence region can be found for θ by using a chi-square approximation to the distribution of $-2 \log \Re(\theta)$.

The usual inference for quasi-likelihood estimation uses either a normal approximation to the distribution of the maximum quasi-likelihood estimate, $\hat{\theta}_{QL}$, or a chi-square approximation to the distribution of the quasi-likelihood ratio statistic, when it exists. Although the point estimates $\tilde{\theta}$ and $\hat{\theta}_{QL}$ will be the same, this need not be true for the corresponding likelihood ratio curves, and hence the confidence regions. At first glance it may appear that empirical likelihood and the parametric methods of inference rely on more or less the same assumptions, namely the specification of the first two moments and some weak conditions on a few of the higher moments. However, while quasi-likelihood requires that the first two moments be specified correctly, empirical likelihood only requires that the estimating functions have expectation zero. Hence empirical likelihood will be robust to misspecification of the variance function, as long as the equation for the mean is correct. In the parametric case, a normal approximation will still be valid in the presence of variance misspecification, but the wrong quantity will be used in setting the confidence limits; the chi-square approximation will not even be valid.

As an example, consider the case where the variance is a power of the mean. Suppose that while we specify $V(\mu_i) = \mu_i^{\alpha}$, the true variance is in fact $V_T(\mu_i) = \mu_i^{\alpha+\epsilon}$, for some $\alpha, \epsilon > 0$. Define D to be the $n \times p$ matrix with (i,j)th entry $\frac{\partial \mu_i}{\partial \theta_j}$, $V = \text{diag}\{\mu_1^{\alpha}, \ldots, \mu_n^{\alpha}\}$, and $V_\epsilon = \text{diag}\{\mu_1^{\epsilon}, \ldots, \mu_n^{\epsilon}\}$. Then, whereas we would use the relation

$$\sqrt{n}(\theta - \hat{\theta}_{QL}) \sim N\Big(0, n\phi(D'V^{-1}D)^{-1}\Big)$$

in setting our approximate normal confidence limits, the correct relation to use would be

$$\sqrt{n}(\theta - \hat{\theta}_{QL}) \sim N\Big(0, n\phi(D'V^{-1}D)^{-1}(D'V_\epsilon V^{-1}D)(D'V^{-1}D)^{-1}\Big).$$

Furthermore, if the quasi-likelihood ratio statistic, $QLR = -2\{Q(\mu, Y) - Q(\hat{\mu}_{QL}, Y)\}$, exists, we would approximate its distribution by a $\chi^2_{(p)}$ random variable, when in reality we have

$$QLR \sim \chi^2_{(p)} + U_\theta + O_p(n^{-1/2}),$$

where U_θ is a random variable with expectation $E[U_\theta] = \text{tr}\{(D'V^{-1}D)^{-1}(D'V_\epsilon V^{-1}D)\} - n$. As a check, note that when $\epsilon = 0$ the two normal approximations are identical and $U_\theta \equiv 0$.

In general, empirical likelihood may suffer a loss of efficiency under misspecification of the variance, but the confidence regions for θ will still be correct, provided that conditions (3.1a′)-(3.1c′) are satisfied. Condition (3.1a′) insures that the mean of the quasi-score functions (i.e. zero) is in the convex hull of the

data with high probability. Similar to Owen (1991), a sufficient condition for (3.1a′) to hold is that

$$ch(P) \cap ch(N) \neq \emptyset, \tag{4.1.3}$$

where $P = \{x_i \mid y_i - \mu_i > 0\}$, $N = \{x_i \mid y_i - \mu_i < 0\}$, and $\mu_i = g^{-1}(x_i'\theta)$. Conditions (3.1b′) and (3.1c′) are mild conditions on the moments of Y to allow a moment approximation of $-2\log\Re(\theta)$ and the use of the Lindberg-Feller Central Limit Theorem. If we define $\mu_4(x_i) = (\frac{\partial g^{-1}(x_i'\theta)}{\partial\theta})^4 \int [\frac{Y_i-\mu_i}{\phi V(\mu_i)}]^4 dF_{x_i}$, then

$$n^{-2}\sum_{i=1}^{n} \|x_i\|^4 \mu_4(x_i) \to 0 \tag{4.1.4}$$

is a sufficient condition for (3.1b′). To understand condition (3.1c′) better in the context of this section, first note that $\Psi_{in} = \mathrm{var}(Z_i) = \tau(x_i)x_ix_i'$, where

$$\tau(x_i) = \frac{\left(\dfrac{\partial g^{-1}(x_i'\theta)}{\partial\theta}\right)^2}{\phi V(\mu_i)},$$

and so $\Psi_n = \frac{1}{n}\sum \tau(x_i)x_ix_i'$. Trivial inequalities then show that it is sufficient for (3.1c′) that the $\tau(x_i)$'s and the maximum and minimum eigenvalues of $\frac{1}{n}(X'X)$ be bounded away from zero and infinity. The following corollary formalizes the above statements.

Corollary 1. *Let Z_i be defined as in* (4.1.2). *Assume* (4.1.4) *holds, and* (4.1.3) *holds with probability tending to 1 as $n \to \infty$. Suppose there exist constants $a, b > 0$ such that $a < \tau(x_i)$ for $i = 1, \ldots, n$, and for all $n \geq p$, $a < \mathrm{mineig}(\frac{1}{n}(X'X))$ and $\frac{1}{n}\sum \tau(x_i)\|x_i\|^2 \leq b < \infty$. Then $-2\log\Re(\theta) \to \chi^2_{(p)}$ in distribution as $n \to \infty$.*

The condition that $\frac{1}{n}\sum \tau(x_i)\|x_i\|^2 < \infty$ comes from using Rayleigh's principle on Ψ_n. For normal data and the model in Example 5.2 below, (a non-exponential family), $\tau(x_i)$ is a constant. For binomial data, $\tau(x_i) \leq \frac{1}{4}$. Thus for these three cases, this condition may be replaced by $\frac{1}{n}\sum \|x_i\|^2 < \infty$. For the cases where the data is modeled as Poisson, gamma, or inverse Gaussian, $\tau(x_i)$ is simply μ_i, μ_i^2, and μ_i^3, respectively. Hence, with the assumption that the μ_i are bounded above, we may again use the simpler condition that $\frac{1}{n}\sum \|x_i\|^2 < \infty$ in these three settings. Note that this condition is satisfied easily in the cases where the x_i's are bounded and where they are sampled from a normal distribution. Since $\tau(\cdot)$ is just a function of the link and variance functions, other models may be checked on a case by case basis.

4.2. GLMs with fixed, unknown dispersion

Consider Model (1.1) again, but now suppose that, although ϕ is believed to be fixed, its value is unknown. Such an assumption is common in binomial and

Poisson overdispersion models. A commonly used point estimate for ϕ is

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \tag{4.2.1}$$

a moment estimator based on the squared Pearson residuals $r_p^2 = \frac{(Y-\mu)^2}{V(\mu)}$. Note that by assuming the form of only the first two moments, quasi-likelihood does not enable one to do inference on $\hat{\phi}$. In this respect, Nelder & Pregibon's (1987) extended quasi-likelihood offers an improvement by augmenting the quasi-likelihood $Q(\mu;Y)$ so that it will behave like a likelihood with respect to both θ and ϕ. The resulting likelihood is

$$Q^+(\mu_i;Y_i) = -\frac{1}{2}\log\{2\pi\phi V(Y_i)\} - \frac{1}{2}\frac{D(Y_i;\mu_i)}{\phi},$$

where

$$D(Y_i;\mu_i) = -2\phi\{Q(\mu_i;Y_i) - Q(Y_i;Y_i)\} = -2\int_{Y_i}^{\mu_i}\frac{Y_i-t}{V(t)}dt,$$

is the deviance function.

By construction, $\frac{\partial Q^+}{\partial\theta} = \frac{\partial Q}{\partial\theta} = \frac{Y_i-\mu_i}{\phi V(\mu_i)}\frac{\partial\mu_i}{\partial\theta}$ has zero mean. With the assumption that $\kappa_r(Y_i) = O(\phi^{r-1})$ for $r \geq 2$, a saddlepoint argument using small-dispersion asymptotics shows that $E\{D(Y_i;\mu_i)\} \simeq \phi$ and hence $\frac{\partial Q^+}{\partial\phi} = \frac{D(Y_i;\mu_i)}{2\phi^2} - \frac{1}{2\phi}$ has approximately zero mean (McCullagh and Nelder (1989)). In contexts where these assumptions are tenable, empirical likelihood confidence intervals may be obtained for (θ,ϕ) using the estimating functions

$$Z_i^* = Z\{(Y_i,X_i);(\theta,\phi)\} = \left\{Z_i', \frac{D(Y_i;\mu_i)}{2\phi^2} - \frac{1}{2\phi}\right\}' \in \mathbb{R}^{p+1} \tag{4.2.2}$$

in place of the Z_i of (4.1.2). Since the assumptions above on the cumulants are properties of the exponential families, empirical likelihood confidence intervals based on Z_i^* should be especially useful in models for binomial and Poisson data with mild overdispersion.

In addition to the deviance function $D(Y_i;\mu_i)$, another common measure of dispersion is the squared Pearson residual, r_p^2, defined above. As use of the deviance function can be motivated by extended quasi-likelihood, use of the Pearson residual can be motivated by pseudo-likelihood methods. Differentiating the pseudo-likelihood

$$PL_i = \frac{r_{pi}^2}{2\phi} + \frac{1}{2}\log(2\pi\phi V(\mu_i))$$

with respect to ϕ, we get $\frac{\partial PL_i}{\partial \phi} = \frac{1}{2\phi} - \frac{r_{pi}^2}{2\phi^2}$, which has zero mean under the true (θ, ϕ). This would suggest using

$$\tilde{Z}_i^* = \tilde{Z}\{(Y_i, X_i); (\theta, \phi)\} = \left(Z_i', \frac{r_{pi}^2}{2\phi^2} - \frac{1}{2\phi} \right)' \in \mathbb{R}^{p+1} \tag{4.2.3}$$

as our estimating functions. Simply noting that r_{pi}^2 is an unbiased estimate of ϕ and inserting it in place of $D(Y_i; \mu_i)$ in (4.2.2) would lead to the same formula.

In the case of a normal model the above two measures of dispersion coincide, but they differ when one moves away from normality. Nelder & Lee (1992) cite "considerable disagreement" about which is better and address the question in the context of finite samples through an extensive series of simulations. This approach is motivated, in part, by the work of Davidian & Carroll (1988), which found the Pearson residuals to be preferable in certain cases due to the asymptotic bias of $D(Y; \mu)$. Their asymptotics were based on small ϕ or large μ. In terms of mean square error, Nelder & Lee found that the deviance did noticeably better than the Pearson residuals in examples with small values of μ, and as well or slightly better with large values. The authors reason that these results arise from the bias of $D(Y; \mu)$ being overwhelmed by the larger variance of r_{pi}^2 in finite samples, the latter being due to the fact that the deviance residuals are very close to the best normalizing transformation. Hence, in general, it appears that Equation (4.2.2) is preferable to (4.2.3). However, in practice if the μ_i are felt to be large, (4.2.3) at times may be more attractive because it will often be less complicated to code into an optimization routine. Moreover, in some situations, the $D(Y_i; \mu_i)$ may not exist, as will be seen in Example 5.2.

Regardless of whether Z_i^* or $\tilde{Z}_i^*$ is used, the MELE of ϕ is given by

$$\tilde{\phi} = \arg \max_{\phi \in (0,\infty)} \Re\{(\theta, \phi)\}.$$

An approximate confidence interval for ϕ can be found by approximating the distribution of $-2 \log \Re(\phi)$ by $\chi_{(1)}^2$, where $\Re(\phi) = \sup_{\theta \in \mathbb{R}^p} \Re\{(\theta, \phi)\}$. A set of sufficient conditions for (3.1a')-(3.1c') to hold in the present context would be similar to those in Section 4.1, but necessarily involving higher moments of the Y_i's. It should be noted that the validity of empirical likelihood confidence regions for a portion of the parameter vector obtained by profiling out the remaining portion as nuisance parameters does not follow directly from Theorem 1. However, Corollary 5 of Qin & Lawless (1991) supports this in the case of independent and identically distributed random variables, and can be adapted to the present context by appealing to the Lindberg-Feller Central Limit Theorem.

As mentioned above, there are certainly other sources for the variables Z_i^* and $\tilde{Z}_i^*$. For example, Efron (1986) and Jørgensen (1987) use slightly stronger

assumptions than those above to arrive at expressions similar to those of extended quasi-likelihood. Also, Godambe and Thompson (1989) derive optimal sets of estimating equations for either θ or (θ, ϕ), but knowledge of the skewness and kurtosis of the underlying distribution is necessary in order to use them. However, Godambe (1991) has shown, recently, that the dependence on these moments is "slight" in some situations.

4.3. Simultaneous modeling of mean and dispersion

The model (1.1) can be extended even further by attempting to model the mean and dispersion simultaneously. This type of modeling has received increased attention, recently, in areas where quality control and quality improvement are of interest. We begin by assuming that (1.1) and (1.2), together, form a reasonable model for the data. Empirical likelihood can then be applied in a straightforward manner by using the method of Section 4.1 on both the mean and dispersion submodels. The resulting estimating functions are simply

$$\begin{aligned} Z_i^{**} &= Z\{(Y_i, X_i, U_i); (\theta, \gamma)\} \in \mathbb{R}^{p+q} \\ &= \left[\frac{Y_i - \mu_i}{\phi_i V(\mu_i)} \frac{\partial \mu_i}{\partial \theta}, \frac{d_i - \phi_i}{\tau V_D(\phi_i)} \frac{\partial \phi_i}{\partial \gamma}\right]', \end{aligned} \tag{4.3.1}$$

where $\mu_i = g^{-1}(X_i'\theta)$ and $\phi_i = h^{-1}(U_i'\gamma)$.

Unfortunately, in practice (4.3.1) cannot often be used because the form of $V_D(\cdot)$ in (1.2) cannot be specified with any confidence. Typically, we may be comfortable with specifying only that $h(\phi_i) = U_i'\gamma$ for some real-valued function $h(\cdot)$, covariates U_i, and parameter $\gamma \in \mathbb{R}^q$. Using extended quasi-likelihood, this is still sufficient to obtain a set of estimating functions. By differentiating $Q^+(Y_i; \mu_i)$ with respect to γ instead of ϕ, we get

$$\frac{\partial}{\partial \gamma} Q^+(\mu_i; Y_i) = \frac{D(Y_i; \mu_i) - \phi_i}{2\phi_i^2} \frac{\partial \phi_i}{\partial \gamma}, \tag{4.3.2}$$

and can, therefore, use

$$Z_i^{***} = Z\{(Y_i, X_i, U_i); (\theta, \gamma)\} = \left[Z_{1i}^{**\prime}, \frac{d_i - \phi_i}{2\phi_i^2} \frac{\partial \phi_i}{\partial \gamma}\right]' \in \mathbb{R}^{p+q}, \tag{4.3.3}$$

where Z_{1i}^{**} is the first component of (4.3.1).

Note from the form of (4.3.2) that this approach implicitly assumes that $\text{var}(d_i) = 2\phi_i^2$. Since $\text{var}(r_p^2) = 2\phi^2(1 + \rho_4/2)$, where $\rho_4 = \kappa_4(Y)/\kappa_2^2(Y)$, and $\text{var}(D(Y; \mu)) \simeq 2\phi^2(1 + \rho_4/2)$, both $D(Y_i; \mu_i)$ and r_{pi}^2 seem to be reasonable candidates for d_i. Note that, if an estimate of ρ_4 is available from the data, using

$$\tilde{d}_i = \frac{d_i}{\sqrt{1 + \rho_4/2}} \tag{4.3.4}$$

should improve efficiency of the corresponding estimating equations. See McCullagh and Nelder (1989, Chapter 10), for a more complete discussion. Godambe & Thompson's (1989) paper also develops a set of optimal estimating functions for estimating θ and γ simultaneously, although these, too, require knowledge of the skewness and kurtosis. However, in situations where the data are substantial enough to estimate ρ_4 well in (4.3.4), estimates of skewness and kurtosis may be good enough to make the equations of Godambe & Thompson worth trying as well.

The manner in which the Z_i^{**} and Z_i^{***} are used to form empirical likelihood confidence regions for (θ, γ) is completely analogous to Sections 4.1 and 4.2, using $\chi^2_{(p+q)}$ to approximate the distribution of $-2 \log \Re\{(\theta, \gamma)\}$. Confidence regions for components of either θ or γ are obtained by profiling out the other components as nuisance parameters. Because the theory behind empirical likelihood implies that it may be used with any reasonable set of estimating functions, the primary difficulty in modeling the mean and dispersion simultaneously is in finding such functions. The performance of empirical likelihood may vary greatly among competing sets of estimating functions, depending on their relative efficiencies.

5. Some Examples

5.1. Kyphosis data

This example involves the presence or absence of *kyphosis*, a post-operative deformity, following corrective spinal surgery (Chambers & Hastie (1991)). Measurements were taken on 81 children, with 17 cases of kyphosis being observed. The response y_i was a binary variable for the presence or absence of kyphosis, while the covariates were x_1 the age of the child, x_2 the number of vertebrae involved in the operation, and x_3 the beginning of the range of the vertebrae. Physicians were interested in whether the latter three variables are related to the former in such a way that they could be used for pre-operative screening. A logit model was fit to the data, with the dispersion parameter ϕ identically equal to 1, i.e.

$$E(y_i) = \pi_i, \quad \text{var}(y_i) = \pi_i(1 - \pi_i), \quad \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3.$$

One question to ask of these data is whether the age of a child has a significant effect on the odds of developing kyphosis after surgery. For example, if younger children seem more inclined to develop kyphosis, perhaps postponing the time of operation would be beneficial in some cases. Thus we test the hypothesis $H_0 : \theta_1 = 0$.

In order to find empirical likelihood confidence intervals for θ_1, the FORTRAN package NPSOL (Gill, Murray, Saunders & Wright (1986)) was used. The empirical likelihood estimating functions for the model are

$$Z_i = X_i\Big[Y_i - \{1 + e^{(-X_i'\theta)}\}^{-1}\Big].$$

The problem of finding

$$\Re(\theta_1) = \sup_{(\theta_0,\theta_2,\theta_3)\in\mathbb{R}^3} \Re(\theta)$$

was interpreted as one of maximizing $\sum \log(np_i)$ over 84 variables, i.e. the p_i, θ_0, θ_2, and θ_3, subject to the constraints $p_i \geq 0$, $\sum p_i = 1$, and $\sum p_i Z_i = 0$. Although the dimension of this problem appears daunting, it turns out to be relatively unimportant to NPSOL. Much more important is the smoothness of the function being optimized. For this problem, NPSOL was able to compute reliable values of the profile empirical likelihood ratio far beyond any necessary levels, as determined by a chi-square approximation.

Figure 1 shows the resulting profile empirical likelihood curve, with asymptotic 95% and 99% lines drawn in by comparing the distribution of $-2\log\Re(\theta_1)$ with the $\chi^2_{(1)}$ distribution. For the hypothesis H_0, empirical likelihood gives a p-value of slightly less than 0.05, thereby suggesting that there is some evidence for rejecting at the 95%-level but not at the 99%-level. The dotted curve in Figure 1 is the profiled parametric binomial likelihood ratio. The fact that the two curves are so close suggests that the logistic model does a good job describing the relationship underlying the data. Empirical likelihood gives a slightly larger right endpoint, and a noticeably smaller left endpoint, at both the 95% and the 99% levels. Note that, based on the binomial likelihood ratio, we would not have rejected the hypothesis that $\theta_1 = 0$ at the 95% level. The actual confidence intervals were $(-0.00109, 0.02468)$ and $(-0.00475, 0.02958)$ for the binomial likelihood ratio, and $(0.00055, 0.02488)$ and $(-0.00247, 0.03028)$ for empirical likelihood, at the 95% and 99% levels respectively.

5.2. Leaf-blotch data

As an example of doing empirical likelihood inference on the dispersion parameter ϕ, we consider the leaf-blotch data of Wedderburn (1974). Ten varieties of barley were grown on nine sites and examined for *Rhynchosporium Secalisi* or leaf blotch. The response y_i was the proportion of the leaf area affected, while the covariates were indicators for variety and site. Following the example of Wedderburn (1974) and McCullagh & Nelder (1989), we treat the proportions as pseudo-binomial data. These authors found the usual binomial variance to be

inadequate for the data, and instead suggested using the model

$$E(y_i) = \pi_i, \quad \text{var}(y_i) = \phi\pi_i^2(1-\pi_i)^2, \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \theta_0 + \theta_1 x_1 + \cdots + \theta_{17}x_{17}.$$

Here the parameter ϕ cannot strictly be called an overdispersion parameter, but instead acts as a simple scale parameter. Hence, for example, if $\phi = 1$ it follows that a more suitable variance for this data than the binomial variance is just its square.

In order to construct empirical likelihood confidence intervals for ϕ we first note that the variance we have chosen corresponds to the deviance function

$$D(y_i;\mu_i) = -2\left\{(2y_i-1)\log\left(\frac{\mu_i}{1-\mu_i}\right) - \frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i}\right\},$$

which is undefined when μ_i is 0 or 1. Since some of the observed proportions were zero, this is not actually a proper deviance function in the context of our problem (see McCullagh & Nelder (1989)). Therefore, for the purpose of defining estimating functions, we prefer to use r_{pi}^2 with (4.2.3) instead of $D(y_i;\mu_i)$ with (4.2.2). It follows that the estimating functions take the form

$$\tilde{Z}_i^* = \left[s_i(\theta)\phi^{-1}q_i(\theta)X_i, \frac{\{q_i(\theta)s_i(\theta)\}^2}{2\phi^2} - \frac{1}{2\phi}\right],$$

where $s_i(\theta) = 2 + \exp(X_i'\theta) + \exp(-X_i'\theta)$ and $q_i(\theta) = y_i - \{1+\exp(-X_i'\theta)\}^{-1}$.

Again, the computations were done using NPSOL, this time to find $\Re(\phi)$. Here the problem takes the form of optimizing $\sum\log(np_i)$ over 108 variables, i.e. the p_i's and the 18 coordinates of θ. The maximum empirical likelihood estimate for ϕ was $\tilde{\phi} = 0.791$, suggesting that $\pi_i^2(1-\pi_i)^2$ alone is an inflated model for the variance by about 25%. However, the commonly used moment estimator (4.2.1) estimates ϕ by $\hat{\phi} = 0.99$, which suggests that the variance can indeed be modeled effectively as $\pi_i^2(1-\pi_i)^2$. The difference between these estimates can be explained easily by noting that

$$\tilde{\phi} = \frac{1}{n}\sum_{i=1}^{n} r_{p_i}^2 = \frac{n-p}{n}\hat{\phi}. \tag{5.1}$$

By estimating (θ,ϕ) simultaneously using likelihood-based methods, $\tilde{\phi}$ automatically incorporates weights of $\frac{1}{n}$. On the other hand, $\hat{\phi}$ uses weights of $\frac{1}{n-p}$ to adjust for the biasedness from having first estimated $\hat{\theta}_{QL}$. A similar situation arises with the use of parametric likelihood methods. For example, in a model where $Y_i \sim N(X_i'\theta,\sigma^2)$, the maximum likelihood estimate for σ^2 is $\frac{1}{n}\sum(Y_i - X_i'\hat{\theta})^2$, while the uniform minimum variance unbiased estimator is $\frac{1}{n-p}\sum(Y_i - X_i'\hat{\theta})^2$.

Cox and Reid (1987) have used conditional likelihoods in the parametric case to get "maximum likelihood" estimates which are adjusted. It remains to be determined whether conditioning can be used to get an analogous result for empirical likelihood.

Figure 2 shows the profile empirical likelihood curve for ϕ with 95% and 99% lines. Considering the fact that ϕ is a parameter in the variance of the model, and hence implicitly more difficult to estimate accurately than θ, empirical likelihood gives much narrower confidence intervals than might be expected. The ratios of the right endpoint to the left endpoint are about 1.69 and 2.15 respectively, at the 95% and 99% levels. In comparison, with the same sample size and the same dimension for θ, confidence intervals for σ^2 in the above $N(X_i'\theta, \sigma^2)$ model would have ratios 1.93 and 2.38 respectively. These ratios are the same regardless of whether the adjusted or unadjusted estimator is used to estimate σ^2. Note, however, that some allowance may need to be given for the fact that ϕ is bounded above, whereas σ^2 is unbounded.

If the adjusted estimator $\hat{\phi}$ is preferred to $\tilde{\phi}$, Equation (5.1) implies that the corresponding confidence intervals can be obtained simply by shifting the unadjusted curve by a factor of $\frac{90}{72} = 1.25$. This can be seen by noting that the confidence intervals based on $\tilde{\phi}$ are found using the equation

$$\sum_{i=1}^{n} p_i r_{pi}^2 - \phi = 0,$$

but those based on $\hat{\phi}$ would use

$$\sum_{i=1}^{n} p_i r_{pi}^2 - \frac{n-p}{n}\phi = 0.$$

The resulting curve is plotted in a dashed line in Figure 2. Because the number of parameters being fit to the mean is so large with respect to the number of observations, we feel that confidence intervals based on the adjusted estimator are more reliable here. Note that the adjusted 95% confidence interval, (0.73, 1.27), is almost symmetric about the point estimate $\hat{\phi} = 0.99$ and fairly tight. Combining this with the fact that we cannot reject the hypothesis $H_0 : \phi = 1$ at the 95%-level based on the unadjusted estimator, we suggest that ϕ may be taken to be 1 and hence the variance may be modeled effectively as $\pi_i^2(1-\pi_i)^2$.

6. Discussion

We have shown that empirical likelihood is justified as a method of inference for a large class of models extending far beyond that of linear models. Work

remains to be done examining the behavior of empirical likelihood when applied to specific sub-classes of models such as nonlinear or semi-parametric models. For models with dependent data however, such as time-series, Theorem 1 is not enough to justify empirical likelihood confidence regions.

The fact that empirical likelihood may be applied immediately to almost any reasonable set of estimating functions means that care should be given ahead of time when choosing such functions, for reasons of efficiency. In Section 4 we concentrated on using empirical likelihood mainly with quasi-likelihood and extended quasi-likelihood estimating functions, because of their flexibility and their use of only first and second moment assumptions. In specific cases, the methods of Efron (1986), Jørgensen (1987), or Godambe and Thompson (1989) might prove to be more appropriate for deriving estimating functions, in which case empirical likelihood should still be easily applicable.

Empirical likelihood shares properties of various non-parametric methods based on resampling, yet it works by optimizing a continuous function, which makes it particularly amenable to the imposition of side constraints. This is advantageous, for example, when Model (1.1) is further complicated with the addition of an unknown ϕ or Model (1.2), since the effect on empirical likelihood is equivalent to simply imposing more side constraints. We expect empirical likelihood to exhibit many of the robustness properties of Simonoff and Tsai's (1988) jackknife estimates for quasi-likelihood, but this needs to be examined more closely. Extensions of the bootstrap and jackknife to extended quasi-likelihood estimating functions have yet to be made, so no comparisons can be offered.

Acknowledgements

I am grateful to Art Owen for many helpful discussions. I would also like to thank the referee and an Associate Editor for their valuable comments and suggestions.

Table 1. Estimating functions for some common mean-variance relationships

ϕ	$V(\mu)$	$g(\mu)$	Z	Family
1	1	μ	$X(Y - X'\theta)$	$N(\mu, 1)$
1	μ	$\log(\mu)$	$X(Y - e^{X'\theta})$	$\text{Pois}(\mu)$
$\frac{1}{m}$	$\mu(1-\mu)$	$\log\left(\frac{\mu}{1-\mu}\right)$	$X\left[Y - m\{1 + e^{(-X'\theta)}\}^{-1}\right]$	$\text{Bin}(m;\mu)$

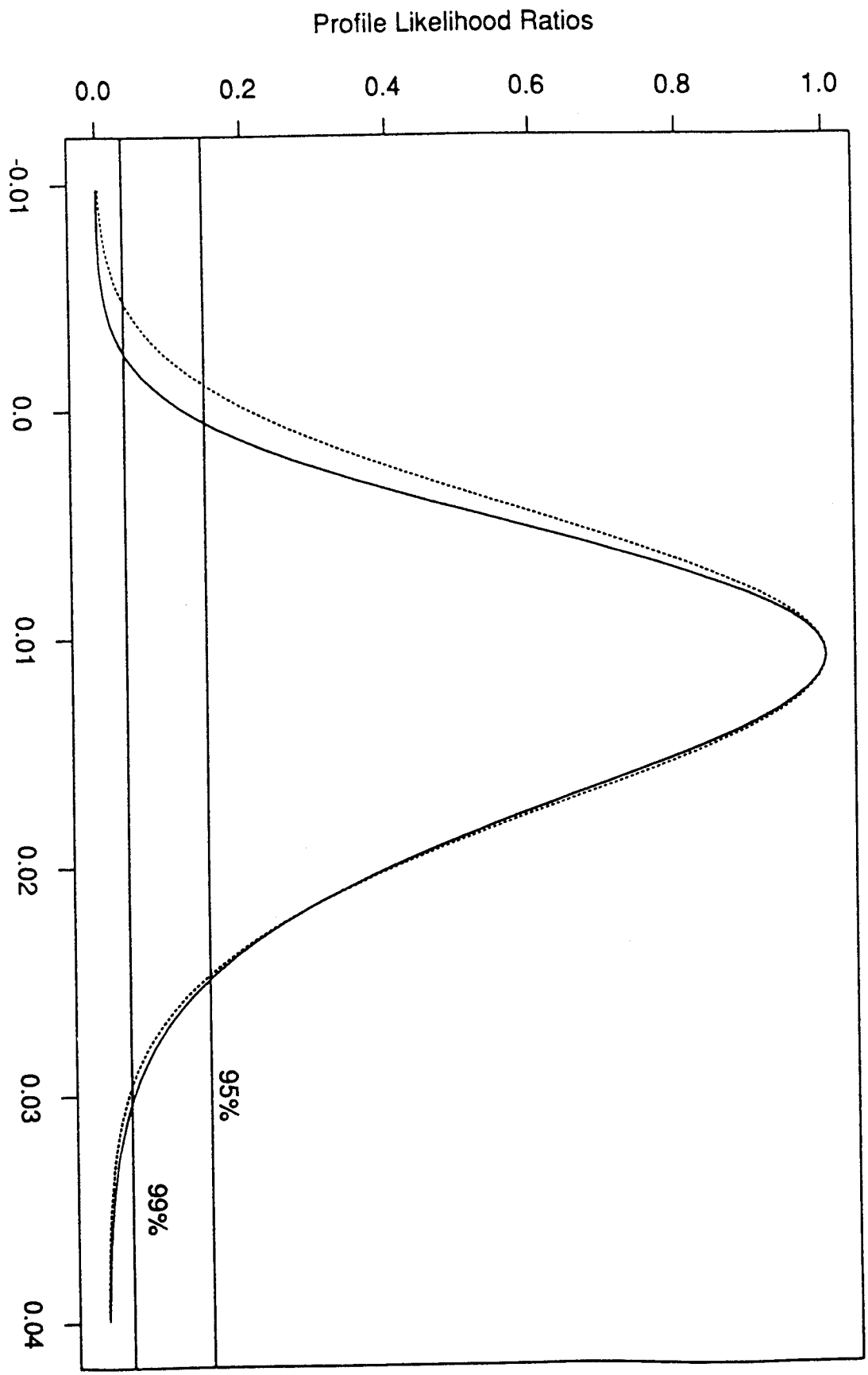

Figure 1. Kyphosis data

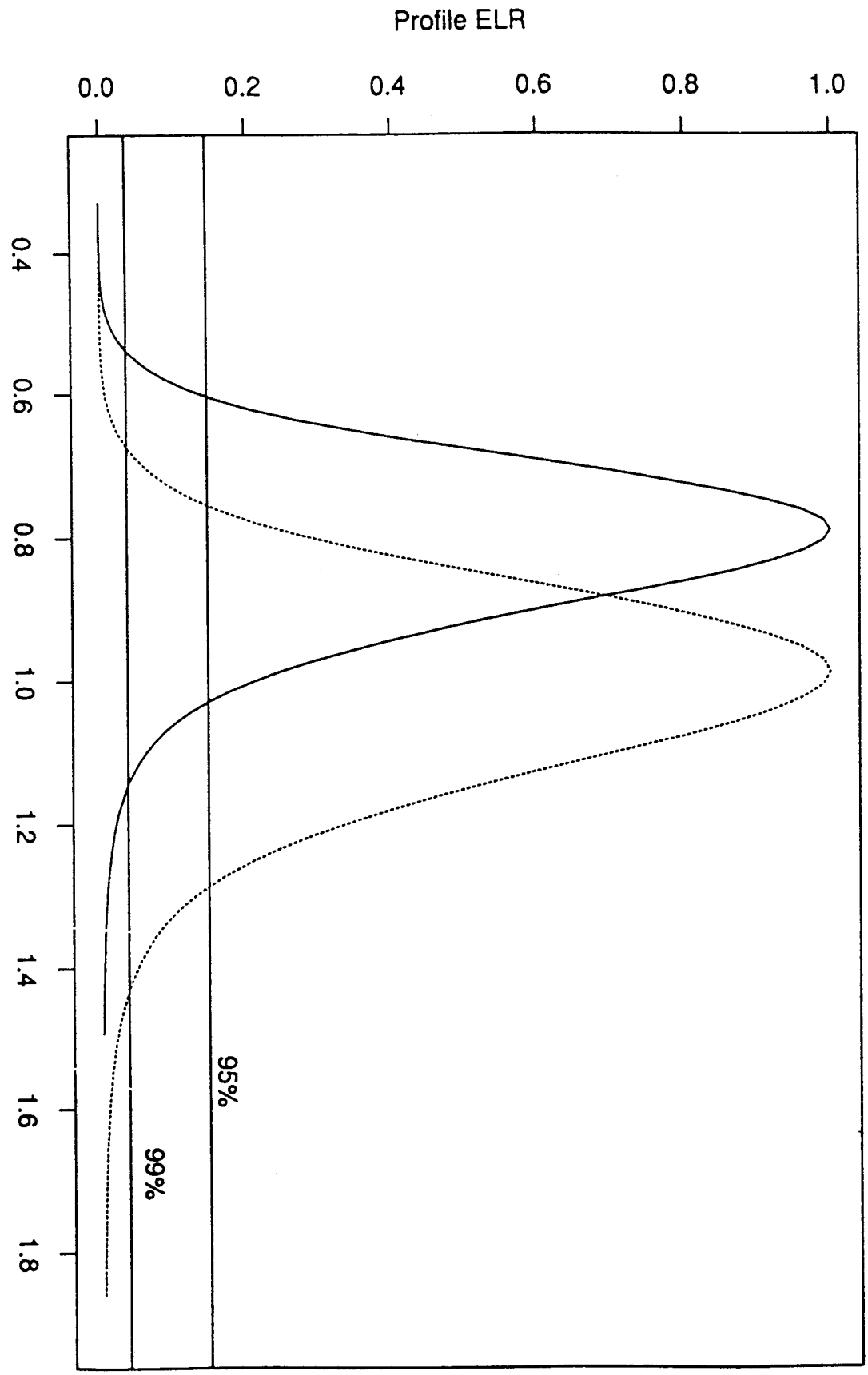

Figure 2. Leaf-blotch data

References

Chambers, J. M. and Hastie, T. J. (1991). *Statistical Models in S.* Wadsworth and Brooks / Cole Advance Books and Software, Pacific Grove, California.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser.B* **49**, 1–39.

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

DiCiccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19**, 1053–1061.

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

Gill, P. E., Murray, W., Saunders, M. A. and Wright, M. H. (1986). User's guide for NPSOL (Version 4.0): A FORTRAN package for nonlinear programming. Technical Report SOL 86-2, Department of Operations Research, Stanford University.

Godambe, V. P. (1991). Non-exponentiality and orthogonal estimating functions. *Proceedings of a Symposium in Honor of Professor V. P. Godambe.*

Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation. *J. Statist. Plann. Inference* **22**, 137–152.

Hall, P. (1990). Pseudo-likelihood theory for empirical likelihood. *Ann. Statist.* **18**, 121–140.

Jørgensen, B. (1987). Exponential dispersion models (with discussion). *J. Roy. Statist. Soc. Ser.B* **49**, 127–162.

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

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

Nelder, J. A. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood: Some comparisons. *J. Roy. Statist. Soc. Ser.B* **54**, 273–284.

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

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

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90–120.

Owen, A. (1991). Empirical likelihood for linear models. *Ann. Statist.* **19**, 1725–1747.

Qin, J. and Lawless, J. (1991). Empirical likelihood and general estimating equations I. Submitted to *Ann. Statist.*

Simonoff, J. S. and Tsai, C. L. (1988). Jackknifing and bootstrapping quasi-likelihood estimators. *J. Statist. Comput. Simul.* **30**, 213–232.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *J. Roy. Statist. Soc. Ser.B* **51**, 47–60.

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

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9**, 60–62.

Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.

(Received May 1992; accepted August 1993)