The Annals of Statistics
1991, Vol. 19, No. 4, 1725–1747

EMPIRICAL LIKELIHOOD FOR LINEAR MODELS[1]

BY ART OWEN

Stanford University

Empirical likelihood is a nonparametric method of inference. It has sampling properties similar to the bootstrap, but where the bootstrap uses resampling, it profiles a multinomial likelihood supported on the sample. Its properties in i.i.d. settings have been investigated in works by Owen, by Hall and by DiCiccio, Hall and Romano. This article extends the method to regression problems. Fixed and random regressors are considered, as are robust and heteroscedastic regressions. To make the extension, three variations on the original idea are considered. It is shown that when some functionals of the distribution of the data are known, one can get sharper inferences on other functionals by imposing the known values as constraints on the optimization. The result is first order equivalent to conditioning on a sample value of the known functional. The use of a Euclidean alternative to the likelihood function is investigated. A triangular array version of the empirical likelihood theorem is given. The one-way ANOVA and heteroscedastic regression models are considered in detail. An example is given in which inferences are drawn on the parameters of both the regression function and the conditional variance model.

1. Introduction. Empirical likelihood is a nonparametric technique for constructing confidence intervals and tests. It has sampling properties similar to the bootstrap, but achieves them through profiling a multinomial with one parameter per (distinct) data point instead of through resampling.

Properties of empirical likelihood in i.i.d. settings are described in Owen (1990), Hall (1990) and DiCiccio, Hall and Romano (1991). This article makes the extension to regression models.

Inferences based on an assumption of homoscedasticity are sometimes invalidated by heteroscedasticity, even in large samples. Empirical likelihood applied in the regression setting accounts for the heteroscedasticity, in much the same way as Wu's (1986) reweighted jackknife or the bootstrap resampling of data vectors.

Section 2 introduces empirical likelihood, gives Theorem 1 from Owen (1990) and describes related work. In Section 3 it is shown that knowledge of one statistical functional can be used to sharpen confidence regions for another, that a Euclidean distance can be used in much the same way as the empirical log-likelihood and that the assumption of identity of distribution can be relaxed. Section 4 introduces the notation and assumptions behind our linear models. Section 5 shows that empirical likelihood confidence regions for

Received April 1989; revised October 1990.
[1]Research supported by Stanford Linear Accelerator Center and NSF Grant DMS-86-00235.
AMS 1980 *subject classification*. Primary 62E20.
Key words and phrases. Bootstrap, jackknife, heteroscedasticity, nonparametric likelihood, variance modeling.

regression parameters have an asymptotic justification with either fixed or random regressors. Homoscedasticity is not required. An extension to robust regression is made. In section 6 the one-way ANOVA model is considered as a special case. In Section 7 we consider simultaneous modeling of the conditional mean and log standard deviation by linear models. The method is illustrated on a heteroscedastic regression in Section 8. Expected breast cancer mortality is modeled as a linear function of the underlying population size while the standard deviation of the mortality rate is modeled by a power law in population size. Some concluding remarks are given in Section 9, on the relative merits of empirical likelihood versus resampling methods and parametric methods. An Appendix contains the proof of Theorem 2.

2. Empirical likelihood. Let $X_1, X_2, \ldots$ be independent random vectors in $\mathbb{R}^p$ for $p \geq 1$, with common distribution function F_0. The empirical distribution

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

is well known to be the nonparametric maximum likelihood estimate of F_0 based on $X_1, \ldots, X_n$. Here δ_x denotes a point mass at x. The likelihood function that F_n maximizes is

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

where $F\{X_i\}$ is the probability of $\{X_i\}$ under F. The notion of nonparametric likelihood can be carried further. One approach is through the empirical likelihood ratio function

$$R(F) = L(F)/L(F_n).$$

$R(F)$ has some of the properties of parametric likelihood ratio functions. Owen (1990) contains the following.

THEOREM 1. *Let $X, X_1, X_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^p$, with $E(X) = \mu_0$ and $\operatorname{var}(X) = \Sigma$ of rank $q > 0$. For positive $r < 1$, let $C_{r,n} = \{\int X\,dF | R(F) \geq r, F \ll F_n\}$. Then $C_{r,n}$ is a convex set and*

$$\lim_{n\to\infty} P(\mu_0 \in C_{r,n}) = P\left(\chi^2_{(q)} \leq -2\log r\right).$$

Moreover if $E(\|X\|^4) < \infty$, then

$$\left|P(\mu_0 \in C_{r,n}) - P\left(\chi^2_{(q)} \leq -2\log r\right)\right| = O(n^{-1/2}).$$

Owen (1990) extends the result to statistics that depend smoothly on several means and to statistics with linear estimating equations. Examples of the former include the variance and the product moment correlation. The latter include quantiles and M-estimates. The restriction to distributions that reweight the sample, that is, $F \ll F_n$, is a technical one and is unnecessary if a

bounded support can be prespecified for F_0 or if the statistic under consideration has positive breakdown. The chi-square limit and even the rate given in Theorem 1 are the same ones found for the parametric case by Wilks (1938).

Define the profile empirical likelihood ratio function

$$\mathscr{R}(\mu) = \max\left\{R(F) \mid \int X\,dF = \mu, F \ll F_n\right\}$$

for all μ in the convex hull of the data $X_1, \ldots, X_n$, and let it be zero outside the convex hull. It follows from the proof of Theorem 1 that for $\mu = \mu_0 + O_p(n^{-1/2})$,

$$-2\log \mathscr{R}(\mu) = n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu) + O_p(n^{-1/2}) \tag{2.1}$$

provided fourth moments of X are finite. The lead term in (1) may be taken with a sample version of Σ in which case it becomes Hotelling's T^2. The $O_p(n^{-1/2})$ term is dominated by sample third moments and it elongates the confidence regions in directions of positive skewness.

DiCiccio, Hall and Romano (1991) show that the error in coverage probability in Theorem 1 is $O(n^{-1})$, provided certain Edgeworth expansions are justified, and this holds for smooth functions of means. This implies that central confidence intervals for a real parameter have coverage errors $O(n^{-1})$, although one-sided intervals have coverage errors $O(n^{-1/2})$. These are also the typical rates when parametric likelihood intervals are used as confidence intervals. A consequence is that, compared to Hotelling's T^2, empirical likelihood does not increase the *order* of coverage accuracy, although simulations in Owen (1988b) suggest an improvement is obtained for the mean of a skewed distribution.

Many techniques for improving the accuracy of parametric likelihood intervals apply also to empirical likelihood. DiCiccio, Hall and Romano (1991) show that a Bartlett correction reduces the central coverage errors to $O(n^{-2})$ and DiCiccio and Romano (1988b) show that a location scale modification to the signed root of the profile empirical likelihood ratio function reduces one-sided errors to $O(n^{-1})$.

Hall (1990) gives a location adjustment of order $O(n^{-1})$ to the family of empirical likelihood confidence regions that makes them second order correct. Since only a location adjustment is required, the regions are, to second order, of correct size, shape and orientation.

When there are no ties among the X_i, the empirical likelihood ratio function takes the form

$$R(F) = \prod_{i=1}^{n} nw_i, \qquad w_i = F\{X_i\}.$$

Owen (1988a) shows that this formula is still appropriate even when there are ties in the data, with the natural modification $\sum_{j:\, X_j = X_i} w_j = F\{X_i\}$. Taking the supremum of $R(F)$ subject to a constraint $T(F) = t$, forces $w_i = w_j$ whenever $X_i = X_j$.

The computation of $\mathscr{R}(\mu)$ is discussed in Owen [(1990), Section 5] and Owen (1988c). We record some of the details here. To compute $\mathscr{R}(\mu)$, one must maximize $\prod w_i$ subject to $w_i \geq 0$, $\sum w_i = 1$ and $\sum w_i X_i = \mu$. Assume μ is inside the convex hull of $X_1, \ldots, X_n$ in which case the problem can be reduced to one of minimizing $-\sum \log(1 + \lambda'(X_i - \mu))$ over $\lambda \in \mathbb{R}^p$. The minimum value is $\log \mathscr{R}(\mu)$. The new problem is the convex dual to the original constrained maximization. It is an unconstrained minimization in a smaller number of unknowns. The function to minimize is convex, so there exist algorithms to find the global minimum. Attempting to compute $\mathscr{R}(\mu)$ and then inspecting the solution (to see whether it is in the unit simplex) provides a way to determine whether μ is indeed inside the convex hull of $X_1, \ldots, X_n$.

Empirical likelihood methods were first used by Thomas and Grunkemeier (1975) to construct confidence intervals for survival times under censoring. A discussion of the connections between empirical likelihood, the nonparametric tilting bootstrap of Efron (1981) and the Bayesian bootstrap of Rubin (1981) is given in Owen (1990).

3. Extensions to empirical likelihood. In this section we present three extensions of the empirical likelihood method. First we show that if we know the value of a statistical functional T of the unknown distribution F_0, that we can sharpen our inferences by restricting consideration to those distributions F for which $T(F) = T(F_0)$. Then we consider replacing the likelihood criterion by one based on the Euclidean distance from $(w_1, \ldots, w_n)$ to $(1, \ldots, 1)/n$. The resulting method is equivalent to Hotelling's T^2. Finally we introduce a version of Theorem 1 in which the assumption of identity of distribution is relaxed.

3.1. *Constrained empirical likelihood.* If we know $T(F_0)$, then it would be natural to consider only distributions F for which $T(F) = T(F_0)$. This information should allow us to sharpen our inferences for other functionals. Corollary 1 shows that such side information can indeed be used by constraining the empirical likelihood, when both functionals are means.

COROLLARY 1. *Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. random vectors in $\mathbb{R}^{p+q}$, where $Z_i = (X_i', Y_i')'$ with $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}^q$ for $p, q > 0$. Suppose that Z_1 has finite fourth moments and that*

$$\operatorname{var}(Z_1) = \Sigma_{zz} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$$

is of full rank, where the partition above is the natural one. Write $E(Z_1) = \mu_{z0} = (\mu_{x0}', \mu_{y0}')'$, again with the natural partition.

Let

$$\mathscr{R}_{Y|X}(\mu_y|\mu_x) = \frac{\sup\{L(F)|F \ll F_n, \int X\,dF = \mu_x, \int Y\,dF = \mu_y\}}{\sup\{L(F)|F \ll F_n, \int X\,dF = \mu_x\}}, \tag{3.1}$$

where F_n is the empirical function of the Z_i. If $\mu_x = \mu_{x0} + O_p(n^{-1/2})$ and $\mu_y = \mu_{y0} + O_p(n^{-1/2})$, then

$$-2\log \mathscr{R}_{Y|X}(\mu_y|\mu_x)$$
$$= n\left(\left(\bar{Y} - \mu_y\right) - \beta_{y\cdot x}\left(\bar{X} - \mu_x\right)\right)'\Sigma_{y|x}^{-1}\left(\left(\bar{Y} - \mu_y\right) - \beta_{y\cdot x}\left(\bar{X} - \mu_x\right)\right) + O_p(n^{-1/2}),$$

where $\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ and $\beta_{y\cdot x} = \Sigma_{yx}\Sigma_{xx}^{-1}$ and so

$$-2\log \mathscr{R}_{Y|X}(\mu_{y0}|\mu_{x0}) \to \chi^2_{(q)}$$

in distribution as $n \to \infty$.

PROOF. Let

$$A = \begin{pmatrix} I_p & 0 \\ -\beta_{y\cdot x} & I_q \end{pmatrix}$$

so that

$$\operatorname{var}(AZ_1) = \begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{y|x} \end{pmatrix}.$$

We may replace $L(F)$ by $R(F)$ in the definition of $\mathscr{R}_{Y|X}$. Then using (2.1) with $\mu_z' = (\mu_x', \mu_y')$,

$$\begin{aligned} &-2\log \mathscr{R}_{Y|X}(\mu_y|\mu_x) + O_p(n^{-1/2}) \\ &\quad = n\left(\bar{Z} - \mu_z\right)'\Sigma_{zz}^{-1}\left(\bar{Z} - \mu_z\right) - n\left(\bar{X} - \mu_x\right)'\Sigma_{xx}^{-1}\left(\bar{X} - \mu_x\right) \\ &\quad = n\left(\bar{Z} - \mu_z\right)'A'\left(A\Sigma_{zz}A'\right)^{-1}A\left(\bar{Z} - \mu_z\right) - n\left(\bar{X} - \mu_x\right)'\Sigma_{xx}^{-1}\left(\bar{X} - \mu_x\right) \\ &\quad = n\left(\left(\bar{Y} - \mu_y\right) - \beta_{y\cdot x}\left(\bar{X} - \mu_x\right)\right)'\Sigma_{y|x}^{-1}\left(\left(\bar{Y} - \mu_y\right) - \beta_{y\cdot x}\left(\bar{X} - \mu_x\right)\right) \end{aligned} \tag{3.2}$$

as required. The second conclusion follows easily. □

For known μ_{x0}, we can get sharper inferences on μ_{y0} by using $\mathscr{R}_{Y|X}(\mu_y|\mu_{x0})$ instead of $\mathscr{R}_Y(\mu_y)$. The maximum empirical likelihood estimate of μ_{y0} is then approximately $\bar{Y} - \beta_{y\cdot x}(\bar{X} - \mu_x)$ and the likelihood regions are based on the conditional variance matrix $\Sigma_{y|x}$. These constrained empirical likelihood regions are asymptotically at least as small as the unconstrained ones. They can be much smaller if X and Y are highly correlated. This is essentially the same result one sees in the regression estimator from sampling theory [Cochran, (1977), Chapter 7]. Note that it is not necessary to know Σ in order to use the constrained method. The resulting inferences are, to the order given above, equivalent to using the conditional distribution of $\bar{Y} - \mu_{y0}$ given the observed value of $\bar{X}$. That is, imposing side knowledge of a population mean is like conditioning on the observed value of a sample mean.

The value $\int Y d\tilde{F}$, where $\tilde{F}$ maximizes $R(F)$ subject to $\int X dF = \mu_{x0}$ may be of some independent interest as a point estimate of μ_{y0} that makes use of the knowledge of μ_{x0}. In Section 8, we consider a regression through the origin in

which the errors about the regression line are thought to have mean zero and be uncorrelated with the regressor. The usual least squares estimate of the slope uses the second of these conditions and results in residuals that do not have mean zero. In the example of Section 8, the zero mean of the residuals gets imposed as a side constraint and this changes the estimate of the slope.

When a relevant ancillary statistic is available, it is usually considered proper to conduct inferences conditionally upon the observed value of the ancillary; see, for instance, McCullagh [(1987), Chapter 8]. Perhaps this can be done here by imposing side constraints on the ancillary statistic. In this nonparametric setting, a definition of ancillarity of $\bar{X}$ for μ_{y0} should entail zero correlation between $\bar{X}$ and $\bar{Y}$. In that case the conditioning would make no difference in the first order calculations given here, though it may make a difference in higher order considerations. Ancillarity is typically not a first order effect in parametric problems either [McCullagh (1987)].

Side information in the form of the probability of a set can be incorporated through the introduction of indicator variables. It follows that if a quantile is known, it can be used to sharpen inferences. If for events A and B, the conditional probability of A given B is known to be p, this information can be exploited by imposing the linear constraint $E(I_B(I_A - p)) = 0$, where I_A and I_B are appropriate indicator variables.

Delta method arguments extend the constraint method to sufficiently smooth nonlinear statistics. If $T_1(F_0)$ is known, inferences for T_2 can be sharpened by holding $T_1(F) = T_1(F_0)$ during the profiling.

Sheehy (1987) considers distributions that minimize a distance measure from the empirical distribution subject to linear constraints. Likelihood, Kullback–Leibler and Hellinger distances are used and the result is an empirical measure with a reduction in variability, similar to that given in Corollary 1. Lippman (1986) considers minimization of Kullback–Leibler distance from the empirical subject to linear constraints, including specification of conditional probabilities, in the context of pattern recognition.

3.2. *Euclidean likelihood.* Likelihood is not the only distance in the simplex that can be used to generate confidence sets with a chi-square calibration. It is an easy exercise to show that using Euclidean distance in the set where $\sum w_i = 1$ gives rise to Hotelling's T^2 apart from the denominator in the estimate of var(X). One defines the log-likelihood $l_E = -\frac{1}{2}\sum(nw_i - 1)^2$ and maximizes l_E subject to $\sum w_i X_i = \mu$ and $\sum w_i = 0$. A little calculus shows that minus twice this constrained maximum is $n(\bar{X} - \mu)'S^{-1}(\bar{X} - \mu)$, where $S = n^{-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'$. That is, $-2\max\{l_E | \sum w_i X_i = \mu, \sum w_i = 1\} = (1 - n^{-1})T^2$, where T^2 is Hotelling's T^2. The statistic $-2l_E$ can also be thought of as Neyman's chi-squared $\sum_i (O_i - E_i)^2/O_i$, where there are n cells with observed counts $O_i = 1$ and expected counts $E_i = nw_i$.

The Euclidean profile likelihood was constructed without imposing the constraint $w_i \geq 0$. When μ is far enough from $\bar{X}$, the maximum of l_E involves some negative w_i. For instance, if μ lies outside the convex hull of the data, some of the maximizing w_i will be negative. This is advantageous when n is

small, since it allows confidence regions to extend outside the convex hull of the data. There are also computational advantages since algorithms can exploit the quadratic nature of l_E. It has the consequence that confidence intervals for correlations may extend outside $[-1, 1]$ and intervals for a variance may include negative values, when l_E is used. Since ordinary empirical likelihood considers only proper distributions on the data, it automatically obeys range restrictions like those for the correlation and variance.

Other distance measures in the simplex may be used. Efron (1981) and DiCiccio and Romano (1988a) consider the one-dimensional subfamily of multinomials generated by minimizing the Kullback–Leibler distance $D(F, F_n) = \sum w_i \log(nw_i)$ subject to $\sum w_i X_i = \mu \in \mathbb{R}$ and $\sum w_i = 1$. Both sources also consider the likelihood family. DiCiccio and Romano (1988a) describe one-sided confidence intervals based on inverse testing that have coverage errors of order $O(n^{-1})$ for the Kullback–Leibler family, the likelihood family and another family obtained by using the likelihood family on a linearization of the statistic of interest. They also show that likelihood intervals have the usual chi-square coverage, to order $O(n^{-1})$ for central intervals and $O(n^{-1/2})$ for one-sided intervals, in the Kullback–Leibler and in the linearized family.

The likelihood and Euclidean distances have the advantage that the Lagrange multiplier corresponding to $\sum w_i = 1$ can be solved for in terms of the multiplier for $\sum w_i X_i = \mu$ and this simplifies many expressions. The Euclidean distance has the further advantage that the multiplier for the constraint $\sum w_i X_i = \mu$ can also be solved for leading to the closed form $-2 \max l_E = (1 - n^{-1})T^2$. Apart from factors like $1 - n^{-1}$, the Euclidean method reproduces some other well known statistics in the anova setting of Section 6 and in linear regression setting of Section 5.

3.3. *Triangular array empirical likelihood.* In this section, we relax the assumption of identical distribution made in Theorem 1. Such a relaxation is essential to handle regressions with nonrandom regressors. The i.i.d. central limit theorem used in the proof of Theorem 1 gets replaced by an appeal to the Lindeberg–Feller central limit theorem. Since the latter is stated for triangular arrays, it is natural to define a triangular array empirical likelihood theorem.

Some notation is needed. Let maxeig(V) and mineig(V) denote the maximum and minimum eigenvalues of the symmetric matrix V. Let ch(A) denote the convex hull of the set $A \subseteq \mathbb{R}^p$.

THEOREM 2 (Empirical likelihood for triangular arrays). *Let $Z_{in} \in \mathbb{R}^p$ for $1 \le i \le n$ and $p \le n < \infty$, be a collection of random vectors, with $Z_{1n}, \ldots, Z_{nn}$ independent for each n. Suppose that $E(Z_{in}) = m_n$, $\operatorname{var}(Z_{in}) = V_{in}$ and let $V_n = (1/n)\sum_{i=1}^n V_{in}$, $\sigma_{1n} = \operatorname{maxeig}(V_n)$ and $\sigma_{pn} = \operatorname{mineig}(V_n)$.*

Assume that as $n \to \infty$,

$$P\big(m_n \in \operatorname{ch}(\{Z_{1n}, \ldots, Z_{nn}\})\big) \to 1 \tag{3.3a}$$

and

$$n^{-2}\sum_{i=1}^{n} E\left(\|Z_{in}-m_n\|^4\sigma_{1n}^{-2}\right)\to 0 \tag{3.3b}$$

and that for some $c>0$ *and all* $n\ge p$,

$$\sigma_{pn}/\sigma_{1n}\ge c. \tag{3.3c}$$

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

$$\mathscr{R}(m)=\sup\Big\{\prod nw_i\Big|w_i\ge 0,\ \sum w_i=1,\ \sum w_iZ_{in}=m\Big\}.$$

The proof of Theorem 2 appears in the Appendix. Note that the random vectors in any row of the triangular array have a common mean, but may have different variances. In most applications, the mean would be common to all rows as well. The largest eigenvalue of V_n is used to scale the problem in the same way that row sums of variances are commonly used to scale triangular arrays in the central limit theorem.

4. Linear regression models and notation. This section introduces the two versions of the linear model corresponding to fixed (Section 4.1) and random (Section 4.2) designs along with the notation to be used. We adopt the terms regression model and correlation model from Freedman (1981). We include a brief discussion of the standard inference methods for these problems. Section 4.3 considers resampling methods. Our emphasis throughout is on point and set estimates for the coefficients in a linear model.

4.1. *Regression model.* In the regression model, the data are of the form (x_i,y_i) for $1\le i\le n$. Here x_i is a row vector of dimension $p\ge 1$ and $y_i\in\mathbb{R}$. The vector x_i contains explanatory variables on the ith case for which y_i is the response. The response y_i is the observed value of

$$Y_i=x_i\beta_0+\varepsilon_i,$$

where $\beta_0\in\mathbb{R}^p$ is a column vector of coefficients and ε_i is a random variable with mean 0 and variance $\sigma^2(x_i)<\infty$. The n random variables ε_i are independent. The distribution of ε_i may depend on x_i but it does not depend on x_j for any $j\ne i$. We write F_x for the distribution of Y_i when $x_i=x$. We use the notation $\mu(x)=\int Y\,dF_x=x\beta_0$ below.

Our regression model differs from Freedman's in that he assumes homoscedasticity, $\sigma^2(x_i)=\sigma^2$.

It is convenient to let X denote the matrix (assumed here to be of full rank p) whose n rows are the x_i and Y denote the column vector whose n elements are the y_i.

The regression model with homoscedasticity and a multivariate normal distribution for the ε_i is the one usually considered in introductory texts. In this context the maximum likelihood estimate of β_0 is $\beta_{LS}=(X'X)^{-1}X'Y$, which is unbiased and has variance $(X'X)^{-1}\sigma^2$. The usual t, F and χ^2 tests

based on the estimate $\hat{\sigma}^2 = (n - p)^{-1}\|Y - X\beta_{LS}\|^2$ are exact for this model. If the assumption of normality is dropped, but the ε_i are still i.i.d., the normal theory inferences are approximately correct by the central limit theorem. See Anderson (1971). Other versions of the central limit theorem would not require identically distributed ε_i, though stronger moment conditions would have to hold.

When the homoscedasticity condition is dropped, β_{LS} is still unbiased for β_0, but by the Gauss–Markov theorem, it is inefficient. This inefficiency is usually rather mild in applications [Bloch and Moses (1988)]. More seriously the variance of β_{LS} is no longer consistently estimated by $(X'X)^{-1}\hat{\sigma}^2$ and confidence regions and tests for β_0 will not have even asymptotic justification.

4.2. *Correlation model.* In the correlation model, (x_i, y_i) are realizations of n i.i.d. random vectors (X_i, Y_i). We assume that $\mu(x) = E(Y_i|X_i = x)$ and $\sigma^2(x) = \text{var}(Y_i|X_i = x)$ exist and that $E(X_i'X_i)$ is positive definite. Then we may write

$$Y_i = X_i\beta_0 + \eta(X_i) + \varepsilon_i,$$

where $\varepsilon_i = Y_i - \mu(X_i)$ and

$$\beta_0 = E(X_i'X_i)^{-1}E(X_i'Y_i).$$

The random variable ε_i may be interpreted as a measurement error and $\eta(X_i)$ is a misspecification error. It follows from the definition of β_0 that $E(X_i\eta_i) = 0$.

In this model, F_x denotes the conditional distribution of Y_i given that $X_i = x$.

When each F_{x_i} is normal and there is no misspecification ($\eta = 0$ a.s.) and assuming homoscedasticity, the usual normal theory inferences are exact conditionally on the observed values of $X_1, \ldots, X_n$. The central limit theorem can be appealed to if normality fails, but homoscedasticity holds and there is no misspecification. Heteroscedasticity or model misspecification invalidate the normal theory confidence methods.

4.3. *Resampling inference.* The jackknife may be used to estimate the variance of β_{LS} in a way that accounts for heteroscedasticity and nonnormality. For a survey, see Wu (1986) and the discussion that follows. The jackknife was applied to the linear model by Miller (1974). Hinkley (1977) modified the jackknife to account for the unbalanced nature of regression data points. Wu (1986) proposed a method weighted by the information for β_0 in subsamples. Wu's method extends naturally to jackknife schemes leaving out more than one observation.

There are two main approaches to bootstrap resampling for regression. One approach [Efron (1979)] is to fit a linear model, estimate an error distribution from the residuals and generate resampled data from the fitted linear model plus independent errors from the estimated distribution. Another approach is to resample the (x_i, y_i) pairs. Freedman (1981) shows that the former provides asymptotically valid inferences on β_0 in the homoscedastic regression model

and the latter does so in the correlation model, both requiring some moment conditions. Freedman allows for misspecification in his correlation model.

5. Empirical likelihood inference. In this section we consider inferences based on empirical likelihood in both the regression and correlation models.

5.1. *Correlation model.* In the correlation model, the parameter β_0 is defined as a smooth function of population moments. It follows from a delta method argument in Owen [(1990), Theorem 2] that inferences based on empirical likelihood have a large sample justification. All that is required is that the second moments of $X_i'X_i$ and $X_i'Y_i$ exist.

An approach based on using the normal equations as estimating equations for β_0 requires somewhat weaker moment conditions and leads to simpler computations. If for some distribution F on pairs (x, y) with x a row vector of length p and y a scalar we have $\beta = (\int x'x\,dF)^{-1}\int x'y\,dF$, then we also have $\int x'(y - x\beta)\,dF = 0$ and vice versa if $\int x'x\,dF$ is of full rank. So we introduce the random variable $Z_i = Z_i(\beta) = X_i'(Y_i - X_i\beta)$, a column vector of p components. The Z_i are i.i.d. by construction and $\beta_0 = \beta$ if and only if $E(Z_i) = 0$. To test $\beta_0 = \beta$, we test whether the Z_i have mean 0. This may be done using empirical likelihood, according to Theorem 1, assuming only that the Z_i have a finite variance, that is $E\|X_i'(Y_i - X_i\beta_0)\|^2 < \infty$.

The computational problem is to calculate

$$\mathscr{R}(\beta) = \max\Big\{\sum \log nw_i \Big| w_i \geq 0,\ \sum w_i = 1,\ \sum w_i x_i'(y_i - x_i\beta) = 0\Big\},$$

the profile empirical likelihood for β. Minus twice the log empirical likelihood may be referred to a $\chi^2_{(p)}$ distribution, rejecting $\beta_0 = \beta$ for large values of $-2\log\mathscr{R}(\beta)$. If interest centers on a subset of r components of β, one can profile out the other components and refer to a $\chi^2_{(r)}$ distribution. Similar remarks hold if one is interested only in r contrasts. Owen (1990) indicates that a better reference might be based on an appropriate F distribution. If there are no nuisance parameters, then $\mathscr{R}(\beta)$ can be computed using the methods sketched in Section 2. If some components β are to be profiled out, then the nested algorithm of Owen [(1990), Section 6.3] can be used. In Section 8, we use sequential quadratic programming to solve a similar problem.

5.2. *Regression model.* We show here that the empirical likelihood function $\mathscr{R}(\beta)$ used in the correlation model leads to confidence regions with an asymptotic chi-square calibration under mild conditions.

As above, it is most convenient to work with the normal equations. We construct the auxiliary variables $Z_i = Z_i(\beta) = x_i'(Y_i - x_i\beta)$, where now the x_i are not random. If $\beta = \beta_0$, the Z_i all have mean 0. They are not identically distributed. For instance, $\mathrm{var}(Z_i) = x_i'x_i\sigma^2(x_i)$. Homoscedasticity would not suffice to make the Z_i i.i.d.

To handle this case, we appeal to Theorem 2, for triangular arrays. We let $Z_{in} = Z_i(\beta)$. Under the hypothesis that $\beta_0 = \beta$, we have $m_n = 0$ for all n and $V_{in} = x_i' x_i \sigma^2(x_i)$. Empirical likelihood confidence regions for β_0 in this model are asymptotically correctly calibrated by the chi-square, provided that conditions (3.3a, b, c) apply to the Z_{in}.

Condition (3.3a) of Theorem 2 is a mild one for regression problems. We need the convex hull of the Z_{in} to contain 0, when β_0 is used in the construction of Z_{in}. The condition is violated, for example, if $x_i = (1, t_i)$, $\beta_0 = (\beta_1, \beta_t)'$ with $\beta_t < 0$ and the sample values y_i are increasing in t_i, for then there is no way to reweight the data to get a negative least squares slope for the y_i versus the t_i. Under i.i.d. sampling, (3.3a) is satisfied exponentially fast as may be determined by the Vapnik–Cervonenkis theory. The event in (3.3a) has a simple sufficient condition in the regression model. Let $P = \{x_i | Y_i - x_i\beta_0 > 0\}$ and $N = \{x_i | Y_i - x_i\beta_0 < 0\}$. If

$$\text{ch}(N) \cap \text{ch}(P) \neq \varnothing, \tag{5.1}$$

then 0 is in the convex hull of the Z_{in}. In the case of simple linear regression, with $x_i = (1, t_i)$, it suffices to have at least one sequence $t_i < t_j < t_k$, where the jth error $y_j - x_j\beta_0$ differs in sign from the ith and kth.

Conditions (3.3b) and (3.3c) are mild regularity conditions, used to justify a moment approximation to $\log \mathscr{R}(\beta_0)$. They are also strong enough to allow the application of the central limit theorem to the moment approximation and to establish the asymptotic negligibility of certain remainder terms. The matrix V_n from Theorem 2 is, in the regression model, $(1/n)\Sigma x_i' x_i \sigma^2(x_i)$. Trivial inequalities show that $\text{maxeig}(V_n) \leq \max_{1 \leq i \leq n} \sigma^2(x_i) \text{maxeig}((1/n)X'X)$ and $\text{mineig}(V_n) \geq \min_{1 \leq i \leq n} \sigma^2(x_i) \text{mineig}((1/n)X'X)$, so (3.3c) holds if the eigenvalues of $(1/n)X'X$ are bounded away from zero and infinity and the $\sigma^2(x_i)$ are also bounded away from zero and infinity. If for $\alpha \geq 0$, we have $\sigma^2(x_i) \leq A\|x_i\|^\alpha$, then (3.3c) follows if $(1/n)\Sigma \|x_i\|^{2+\alpha} < \infty$ and both $\text{mineig}((1/n)X'X)$ and $\sigma^2(x_i)$ are bounded away from zero.

As for condition (3.3b), introduce $\mu_4(x) = \int (Y - \mu(x))^4 \, dF_x$. If $\text{mineig}(V_n) > a > 0$ for all sufficiently large n, a sufficient condition for (3.3b) is

$$n^{-2} \sum_{i=1}^{n} \|x_i\|^4 \mu_4(x_i) \to 0. \tag{5.2}$$

To summarize:

COROLLARY 2. *Let $n_0 \geq p$, $\alpha \geq 0$ and $a, b > 0$. Assume that* (5.2) *holds and as $n \to \infty$,* (5.1) *holds with probability tending to* 1. *Suppose $a < \sigma^2(x_i) < b\|x_i\|^\alpha$ for all i and that for all $n \geq n_0$, $a < \text{mineig}(X'X)/n$ and $(1/n)\Sigma \|x_i\|^{2+\alpha} < b$. Then $-2 \log \mathscr{R}(\beta_0) \to \chi^2_{(p)}$ in distribution as $n \to \infty$.*

5.3. *Comparison to resampling methods.* We now compare empirical likelihood inferences with those of the resampling methods. For β near β_0, the lead

term in $-2\log \mathcal{R}(\beta)$ is

$$\begin{aligned}(5.3)\quad &\Big(\sum x_i(Y_i - x_i\beta)\Big)\Big(\sum (Y_i - x_i\beta)^2 x_i' x_i\Big)^{-1}\Big(\sum x_i'(Y_i - x_i\beta)\Big)\\ &= (\beta_{LS} - \beta)'\Big((X'X)^{-1}\sum (Y_i - x_i\beta)^2 x_i' x_i (X'X)^{-1}\Big)^{-1}(\beta_{LS} - \beta).\end{aligned}$$

In the correlation model, the error is $O_p(n^{-1/2})$ and to this order of accuracy one can replace $x_i\beta$ by $x_i\beta_0$ or $x_i\beta_{LS}$ in (5.3). In this sense the empirical likelihood method uses

$$(X'X)^{-1}\sum \hat{\sigma}^2(x_i)x_i' x_i (X'X)^{-1}$$

with $\hat{\sigma}^2(x_i) = (Y_i - x_i\beta)^2$ as an estimated variance for β_{LS}. By way of comparison, Hinkley's (1977) weighted pseudovalue version of the jackknife uses $\hat{\sigma}^2(x_i) = (Y_i - x_i\beta_{LS})^2/(1 - p/n)$, Wu's (1986) information weighted delete one jackknife uses $\hat{\sigma}^2(x_i) = (Y_i - x_i\beta_{LS})^2/(1\text{-}x_i(X'X)^{-1}x_i')$, bootstrap resampling of (x_i, Y_i) pairs uses $\hat{\sigma}^2(x_i) \doteq (Y_i - x_i\beta_{LS})^2$ and the usual normal theory uses $\hat{\sigma}^2(x_i) = \|Y - X\beta_{LS}\|^2/(n - p)$ as does bootstrap resampling from residuals. These results may be found in Wu (1986). In the Euclidean version of empirical likelihood, minus twice the log relative likelihood is

$$(5.4)\quad (\beta_{LS} - \beta)'\Big((X'X)^{-1}\sum (Y_i - x_i\beta_{LS})^2 x_i' x_i (X'X)^{-1}\Big)^{-1}(\beta_{LS} - \beta)$$

so that, in comparison with the above it uses $\hat{\sigma}^2(x_i) = (Y_i - x_i\beta_{LS})^2$, providing a close match to the bootstrap with resampled (x_i, Y_i) pairs. The covariance estimate implicit in (5.4), namely $(X'X)^{-1}\Sigma(Y_i - x_i\beta_{LS})^2 x_i' x_i (X'X)^{-1}$, is the one produced by the ACOV option of the REG procedure in SAS (1985); see also White (1980).

5.4. *Misspecification in the regression model.* Empirical likelihood results are available assuming that there is either no model misspecification in the regression model or that the predictors are i.i.d. Can one use empirical likelihood to form confidence regions assuming fixed regressors and some model misspecification? For example, one might seek asymptotic confidence regions for $\beta_0 = (\Sigma x_i' x_i)^{-1}\Sigma x_i\mu(x_i)$, where $\mu(x)$ is not necessarily of the form $x\beta$.

The situation is analogous to one in which we observe $Z_1, \ldots, Z_n \in \mathbb{R}^p$, where $E(Z_i) = \mu_i$ and $\operatorname{var}(Z_i) = V_i$ has full rank and we wish to test whether $\bar{\mu} = (1/n)\Sigma\mu_i$ takes a given value μ_0. For simplicity only, we consider $p = 1$, letting $V_i = \sigma_i^2$. Then the empirical likelihood test refers

$$\frac{n(\bar{Z} - \mu_0)^2}{(1/n)\Sigma(Z_i - \mu_0)^2}$$

to a chi-square distribution. If $\mu_0 = \bar{\mu}$, then $n^{1/2}(\bar{Z} - \mu_0)$ has mean 0 and variance $(1/n)\Sigma\sigma_i^2$. The denominator $(1/n)\Sigma(Z_i - \mu_0)^2$ should estimate this variance, but has expectation $(1/n)\Sigma\sigma_i^2 + (1/n)\Sigma(\mu_i - \bar{\mu})^2$. The variance estimate is thus biased upward by a term due to the model misspecification,

and this term would usually be of the same order of magnitude as the variance itself.

Consequently, confidence sets for β_0 would not have consistent coverage levels in the regression model with misspecification. They would be conservative to the extent that the model misspecification inflates the estimate of variance.

5.5. *Robust regression.* Robust regressions can be obtained by substituting a location M-estimate in the definition of β_0. Introduce the robust correlation model in which β_0 is the solution of

$$0 = E(X_i'\psi(Y_i, X_i\beta_0)),$$

where the (X_i, Y_i) pairs are i.i.d. and ψ is an appropriate function [see Huber (1981)] Typically, ψ depends on Y_i and $X_i\beta_0$ only through $Y_i - X_i\beta_0$. Empirical likelihood confidence regions for this problem will have asymptotically correct coverage under conditions on ψ that make sure β_0 is well defined and provided that the variance of $Z_i = X_i'\psi(Y_i, X_i\beta_0)$ is finite and nonzero. See Owen [(1990), Theorem 3] for M-estimates and empirical likelihood.

For the robust regression model, we assume that

$$E(x_i'\psi(Y_i - x_i\beta_0)) = 0$$

for $i = 1, \ldots, n$. That is, we do not allow misspecification of the model. The Z_i (with x_i in place of X_i) must now satisfy the conditions of Theorem 2, and of course, identifiability conditions on ψ are still needed.

The resulting family of confidence regions are nested around a point estimate of β_0 that is (for the common choices of ψ) robust against occasional extreme values of $y_i - x_i\beta_0$. The regions are also robust in that their coverage properties do not depend on homoscedasticity or normality.

6. ANOVA. The one factor analysis of variance is a commonly used linear model and studying it gives some insight into empirical likelihood. Let the observations be Y_{ij}, where $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$. Let $N = \sum_{i=1}^{k} n_i$. Suppose that $Y_{ij} \sim F_{i0}$ are independent samples from the k different distributions. Use F_i to denote a candidate for F_{i0}. We present two approaches to forming an empirical likelihood ratio function.

It is natural to take the product of k empirical likelihoods, from the k independent samples. This leads to the likelihood

$$L(F_1, \ldots, F_k) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} v_{ij},$$

where $v_{ij} = F_i\{Y_{ij}\}$. The empirical likelihood ratio function is

$$R(F_1, \ldots, F_k) = \prod_i \prod_j n_i v_{ij}.$$

A second formulation is to consider N random pairs (I, Y), where $I \in \{1, \ldots, k\}$ and the Y's for $I = i$ are denoted Y_{ij}. Let F be a distribution on

pairs (I, Y). The data are not i.i.d. from this F because the index I in each pair is nonrandom. Consider

$$L(F) = \prod_i \prod_j w_{ij},$$

where $w_{ij} = F\{(i, Y_{ij})\}$. Let $w_{i\cdot} = \sum_j w_{ij}$ and $w_{j|i} = w_{ij}/w_{i\cdot}$. The empirical likelihood ratio function is

$$\begin{aligned} R(F) &= \prod_i \prod_j N w_{ij} \\ &= \prod_i \prod_j N w_{i\cdot} w_{j|i} \\ &= \Big(\prod_i (N w_{i\cdot}/n_i)^{n_i}\Big)\Big(\prod_i \prod_j n_i w_{j|i}\Big). \end{aligned}$$

If we are interested in statistics that depend only on the $w_{j|i}$'s, we may maximize $R(F)$ by taking $w_{i\cdot} = n_i/N$. Both formulations lead to the same profile empirical likelihood ratio functions for statistics that depend only on the F_i. So empirical likelihood applied to the (I, Y) data set automatically keeps the relative weights of the k groups fixed at the sample values. The second formulation is directly covered by Theorem 2, so both approaches have an asymptotic justification as $n_0 = \min_{1 \le i \le k} n_i \to \infty$.

This equivalence does not hold for the Euclidean distance. That is, profiles of $-\frac{1}{2}\sum_{ij}(N w_{ij} - 1)^2$ can differ from those of $-\frac{1}{2}\sum_{ij}(n_i v_{ij} - 1)^2$, even when the statistics depend on w_{ij} only through $w_{j|i}$. (The two are asymptotically equivalent as $n_0 \to \infty$ since they are each asymptotic to the corresponding empirical likelihood method.)

Suppose that $E(Y_{ij}) = \mu_{i0}$. Let $\mathscr{R}(\mu_1, \ldots, \mu_k)$ be the maximum value of $R(F)$ subject to $\sum_j w_{i|j} Y_{ij} = \mu_i$, $i = 1, \ldots, k$. If each $\mu_i = \mu_{i0} + O(n_o^{-1/2})$, then

$$-2 \log \mathscr{R}(\mu_1, \ldots, \mu_k) = \sum_{i=1}^{k} n_i \frac{\big(\bar{Y}_{i\cdot} - \mu_i\big)^2}{s_i^2} + O_p\big(n_0^{-1/2}\big) \to \chi^2_{(k)}$$

in distribution as $n_0 \to \infty$, where $\bar{Y}_{i\cdot} = n_i^{-1} \sum_j Y_{ij}$ and

$$s_i^2 = n_i^{-1} \sum \big(Y_{ij} - \bar{Y}_{i\cdot}\big)^2.$$

It is necessary that all of the group variances be finite and nonzero. They do not have to be equal. We can change $\bar{Y}_{i\cdot}$ to μ_i or to μ_{i0} in the definition of s_i^2 without changing the order of the approximation.

The most common test in a one-way analysis of variance is for H_0: $\mu_1 = \cdots = \mu_k$. Let μ_0 denote the common mean. Since H_0 forces $k - 1$ contrast means to be 0, the limit law for the empirical likelihood test is $\chi^2_{(k-1)}$. To first

order, under H_0,

$$(6.1) \qquad -2\log\max_{\mu}\mathscr{R}(\mu,\ldots,\mu) = \sum_{i=1}^{k} n_i \frac{\left(\overline{Y}_{i\cdot}-\hat{\mu}\right)^2}{s_i^2} + O_p\left(n_0^{-1/2}\right),$$

where

$$(6.2) \qquad \hat{\mu} = \frac{\Sigma_i n_i \overline{Y}_{i\cdot}/s_i^2}{\Sigma_i n_i/s_i^2}.$$

The estimate of the common mean in (6.2) is not the mean of all the Y_{ij}, which is the usual estimate in the analysis of variance. Instead it weights the group means in inverse proportion to the group variances. If the variances were known a priori to be equal, one could impose such homoscedasticity as a side constraint in the profiling of the likelihood. This would still not lead to the standard anova, even asymptotically. The reason is that the sample variance and sample mean are correlated in general. So information about the group variances will change the estimates for the group means.

In the case of anova, the convex hull criterion (3.3a) becomes $\min_j Y_{ij} \le \mu_i \le \max_j Y_{ij}$ for $i = 1,\ldots,k$. It follows that the method would not work well if k were large and n_0 were small. The Euclidean method might be preferable in this case.

Using the Euclidean distance $-\frac{1}{2}\Sigma_{ij}(n_i v_{ij} - 1)^2$, we can solve for minus twice the log-likelihood ratio under H_0 in closed form. The result is the lead term in (6.1). Replacing n_i by $n_i - 1$ in the denominator of s_i^2, we get a statistic used by James (1951) to test differences among group means when population variance ratios are unknown. James does not use a chi-square criterion. Instead he uses a criterion that depends on the values of s_i^2.

7. Variance modeling. Suppose that the variance $\sigma^2(x)$ is not constant. We saw in Section 5 that empirical likelihood confidence regions still have the correct asymptotic level under both the correlation model and the regression model. Although β_{LS} is not an efficient estimate of β_0, the uncertainty in β_{LS} is adequately assessed.

If we knew $\sigma(x)$, we would use a weighted least squares estimate of β_0, equivalent to the solution β of

$$(7.1) \qquad 0 = \sum_{i=1}^{n} \frac{x_i'(y_i - x_i\beta)}{\sigma^2(x_i)}.$$

We can still use (7.1) if we only know $\sigma(x)$ up to a constant multiplicative factor. If we use (7.1) with an incorrect function σ, our point estimate of β_0 will be inefficient, though it may represent an improvement over β_{LS}. Theorem 2 applies to the weighted least squares estimate too, with somewhat different moment requirements, so the confidence regions for β_0 will attain their nominal levels, asymptotically.

In many applications we will want to use the data to guide us in a choice of weights for weighted least squares. One way to do this is by modeling σ^2. Suppose that

$$\log \sigma = u\theta \tag{7.2}$$

for some parameter vector θ and row vector of explanatory variables u. We change our two models to the heteroscedastic regression model and the heteroscedastic correlation model. We will assume in the heteroscedastic correlation model that u_i is the observed value of U_i and that the triples (X_i, U_i, Y_i) are i.i.d. In the heteroscedastic regression model, u_i and x_i are fixed, while $E(Y_i) = \mu(x_i) = x_i\beta_0$ and $\operatorname{var}(Y_i) = \sigma^2(u_i) = \exp(2u_i\theta)$. Some of the components of u may match components of x or be deterministic functions of components of x.

The normal theory maximum likelihood estimates for β and θ solve

$$0 = \sum x_i(y_i - x_i\beta)\exp(-2u_i\theta) \tag{7.3a}$$

and

$$0 = \sum u_i\bigl(1 - (y_i - x_i\beta)^2\exp(-2u_i\theta)\bigr) \tag{7.3b}$$

jointly. Even if normality does not hold, we can still interpret $x\beta$ as a conditional mean and $\exp(u\theta)$ as a conditional standard deviation. We could replace (7.3a, b) by estimating equations based on a location-scale family other than the normal one, or more generally use M-estimates of location and scale. For a discussion of variance modeling in regression see Davidian and Carroll (1987).

Under i.i.d. sampling, we can test specified values of β and θ by constructing the auxiliary variables

$$Z_i = \Bigl(X_i(Y_i - X_i\beta)\exp(-2U_i\theta), U_i\bigl(1 - (Y_i - X_i\beta)^2\exp(-2U_i\theta)\bigr)\Bigr)' \tag{7.4}$$

and testing whether their common mean is 0. To test only θ or β the empirical likelihood for testing (β, θ) is maximized over the other vector.

In the heteroscedastic correlation model, only moment conditions are needed for the resulting inferences to be valid for the corresponding population values. Misspecification does however complicate our interpretations of the coefficients.

The heteroscedastic regression model requires that the conditional mean and variance functions be specified correctly. Otherwise at least one of the components of the vectors Z_i in (7.4) will have a mean that depends on i. At first this seems to contradict the remarks in the second paragraph of this section. But there, the variance model does not enter into the estimating equations, so its incorrectness does not affect the applicability of Theorem 2.

8. Example. Rice [(1988), page 221] gives a table of breast cancer data. Each data point is from a county in North Carolina, South Carolina or Georgia. For each county the number of adult white females living there in 1960 is given, as is the number of deaths due to breast cancer among adult

white females from 1950 through 1969 inclusive. The population sizes range from about 500 to about 100,000. Rice [(1988), Chapter 14] gives some plots and an analysis of this data. This data is also discussed in Royall and Cumberland (1981). It is well described by a regression through the origin and it is evident that the variance of the mortality count increases as the population increases. Royall and Cumberland consider weighted regressions based on a model in which the variance is proportional to the square of the mean. Rice (1988) fits a regression of the square root of the mortality on the square root of the population. We will only use the data points on page 221 of Rice (1988) which represent $n = 151$ of the 301 counties.

Let X_i denote the population in the ith county (divided by 10,000) and let Y_i denote the mortality due to breast cancer (divided by 10). Scaling the data this way improves numerical stability for the computations that follow. We model the mean of Y_i given $X_i = x$ by $\beta_1 + \beta_2 x$ and the log standard deviation of Y_i given $X_i = x$ by $\theta_1 + \theta_2 \log x$. That is, $\sigma(x) = e^{\theta_1} x^{\theta_2}$. The most interesting parameters are β_2, which may be interpreted as a cancer rate and θ_2 through which we can compare the conditional variance to the Poisson model ($\theta_2 = 0.5$) and to the Gamma model ($\theta_2 = 1.0$).

The normal theory maximum likelihood estimates are $\beta_1 = 0.0073$, $\beta_2 = 3.57$, $\theta_1 = -0.0764$ and $\theta_2 = 0.0749$. That is to say these values satisfy equations (7.3a, b). It might be more natural to use a negative binomial model since the response is discrete, but the normal likelihood equations are simpler and the resulting estimates still have intuitive meaning since they specify conditional moments. Davidian and Carroll (1987) advocate the use of robust scale M-estimates in place of (7.3b). This was not done here because it is easy to inspect the data and plots showed no outliers. A normal QQ plot of $(y_i - \hat{\mu}(x_i))/\hat{\sigma}(x_i)$ is very straight.

The cancer mortality rate is approximately 3.6 per 1000 of population, which translates into 1.8 per 10,000 of population per year, since the figures are over 20 years. This translation is rough since the populations would not be constant over the 20 years. The intercept β_1 is quite close to the origin and the estimate of θ_2 is intermediate between the 0.5 expected under a Poisson model and the 1.0 expected under a gamma model.

Profile empirical likelihood curves were computed for each of the parameters. For example, for a specified value of θ_2, it is necessary to maximize $\prod n w_i$ over $w_i \geq 0$, β_1, β_2 and θ_1 subject to the constraints $\sum w_i = 1$ and

$$\begin{aligned} \sum w_i r_i/\sigma_i^2 = \sum w_i x_i r_i/\sigma_i^2 &= \sum w_i\left(1 - r_i^2/\sigma_i^2\right) \\ &= \sum w_i \log(x_i)\left(1 - r_i^2/\sigma_i^2\right) = 0, \end{aligned} \tag{8.1}$$

where $r_i = Y_i - \beta_1 - \beta_2 x_i$ and $\sigma_i = \exp(\theta_1 + \theta_2 \log x_i)$. The optimization thus has 155 variables, namely the w_i, the β_i and the θ_i, has one linear constraint on the w_i, an equality constraint on θ_2 and four nonlinear constraints. This was done for two sequences of values of θ_2 starting at the maximum likelihood estimate; one increasing until the empirical likelihood was small and the other decreasing. The FORTRAN function NPSOL of Gill, Murray, Saunders and

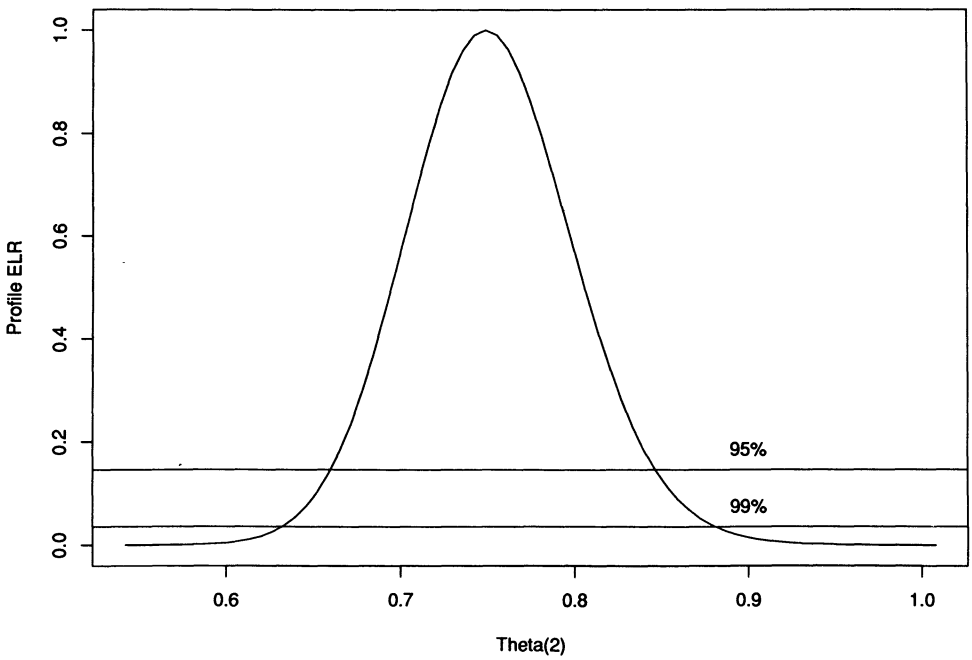

FIG. 1. *Exponent for S.D.*

Wright (1986) was used to do the optimization. NPSOL implements a sequential quadratic programming algorithm. The objective function was taken to be $\sum \log nw_i$ for numerical stability. Each optimization in a sequence provided starting values for its successor.

The profile empirical likelihood curve for θ_2 appears in Figure 1 with asymptotic 95% and 99% lines indicated. There is sufficient evidence in this data to reject the mean-variance relationships of both Poisson and gamma distributions. A similar curve for β_1 has a peak centered near 0 as would be expected. The 99% confidence interval for β_1 has endpoints that lie very close to values translating into -1 and 1 cancer mortalities in the twenty year period.

Because regression through the origin makes sense and fits the data well, it was decided to constrain $\beta_1 = 0$ in the optimization. Holding $\beta_1 = 0$ is not the same as simply dropping the intercept from the regression. In terms of equation (8.1), dropping the intercept would amount to taking $\beta_1 = 0$ in the last three weighted sums and ignoring the first weighted sum. Instead, we are imposing the constraint in all four estimating equations. This asserts that the errors are uncorrelated with the population sizes and have mean zero. Simply dropping the intercept and fitting a regression through the origin would impose only the first of these conditions.

The maximum empirical likelihood possible under the constraint $\beta_1 = 0$ is 0.984. Figure 2 shows a plot of the profile empirical likelihood ratio function of

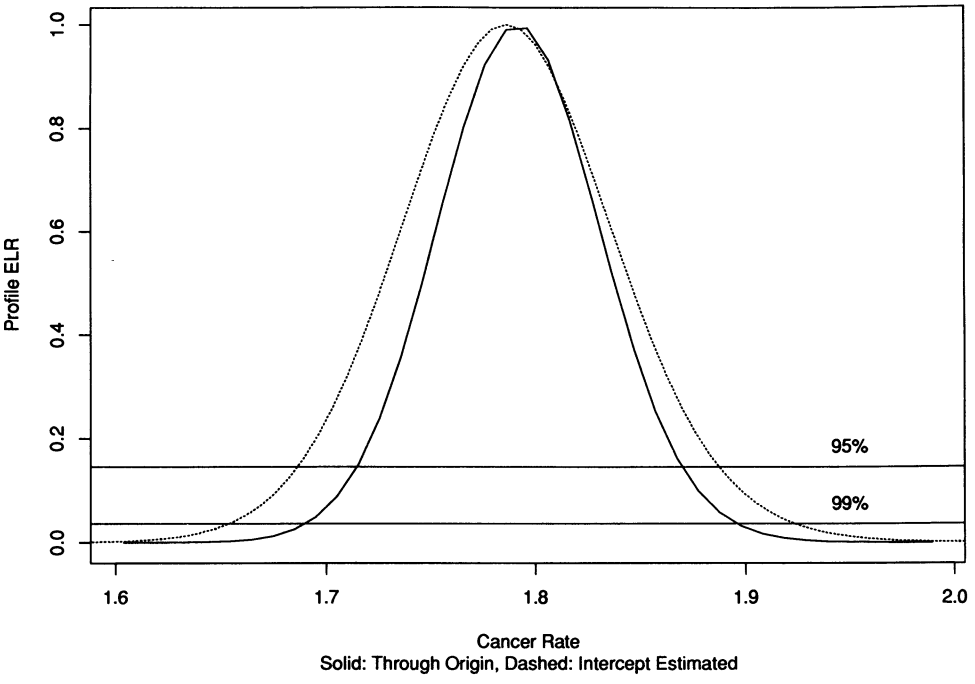

FIG. 2. *Breast cancer deaths per 10,000 population per year.*

β_2 subject to the constraint that $\beta_1 = 0$. The units have been converted to deaths per 10,000 of population per year. The likelihoods achieved for various values of β_2 have been divided by 0.984. Also plotted in a dashed line is the profile empirical likelihood function of β_2 without imposing the constraint that $\beta_1 = 0$. The constraint results in a narrowing of the confidence intervals and a shift in location, as predicted by Corollary 1.

Constrained ($\beta_1 = 0$) and unconstrained plots for θ_2 are almost indistinguishable visually. The same is true for θ_1. These results would be expected if it were known that the residual distribution had zero skewness.

9. Conclusions. The method of empirical likelihood applied to regression problems, has been seen to have wider validity than the usual parametric methods. The asymptotic behavior of the method is much the same as that of various resampling methods, and in fact there is a lot of similarity in the test statistics used by the nonparametric methods.

Empirical likelihood is qualitatively different from the resampling methods in that it uses optimization of a continuous function instead of discrete simulation. This is not an advantage per se, but there may be situations in which it is convenient. For example, empirical likelihood lends itself naturally to the imposition of side constraints. In the example in Section 8, this allowed us to estimate regression coefficients where the normal equations impose more

constraints than the number of regression parameters that are free to vary. If the design matrix X is of full rank, then any of the reweighted versions of it used by empirical likelihood will also be of full rank. This stands in contrast to the resampling methods in which resampled design matrices might be rank deficient. When confidence regions for a pair of parameters are sought, the profiling of an empirical likelihood ratio function provides a natural choice. With resampling methods, the problem of finding the central $1 - \alpha$ fraction of a point cloud arises; see Owen (1990). Empirical likelihood chooses the shape as well as the size of the confidence region. On the other hand, the resampling methods may be applied to enormously complicated statistics much more easily than can empirical likelihood.

To use empirical likelihood, one must specify the estimating equations for the parameters of interest, but need not specify explicitly how to construct standard errors for them. The latter can be quite difficult, and it is often tempting to make assumptions based on tractability instead of wideness of applicability. For instance, one might specify a parametric family of distributions in which it is easy to get interval estimates. A case in point is inference on variances, in which it is convenient to consider normal populations. But normality carries with it the assumption that the kurtosis is zero. If the kurtosis is not zero, normal theory methods do not correctly assess the sampling variability of a quantity like s^2. The well-known test of Bartlett for comparing two or more variances is, for this reason, very sensitive to the normality assumption. Similarly, assumptions of homoscedasticity or symmetry are often made for convenience.

APPENDIX

This appendix contains the proof of Theorem 2. Recall that maxeig(V) and mineig(V) denote the maximum and minimum eigenvalues of the symmetric matrix V and ch(A) denotes the convex hull of the set $A \subseteq \mathbb{R}^p$.

PROOF OF THEOREM 2. Without loss of generality, we may assume that $m_n = 0$ and $\sigma_{1n} = 1$. This follows from considering standardized vectors: $(Z_{in} - m_n)\sigma_{1n}^{-1/2}$. For simplicity of notation, we drop the second subscript n. Let $\hat{V}_n = (1/n)\sum Z_i Z_i'$.

By (3.3a), we may assume that the convex hull of $Z_1, \ldots, Z_n$ contains the origin. Then a straightforward argument based on Lagrange multipliers shows that

$$\mathscr{R}(0) = \prod_{i=1}^{n} n\tilde{w}_i, \tag{A.1}$$

where

$$\tilde{w}_i = \frac{1}{n}\frac{1}{1 + \lambda Z_i} \tag{A.2}$$

are the (strictly positive) coordinates in the simplex of the maximizing $F \ll F_n$

and the multiplier $\lambda \in \mathbb{R}^p$ is uniquely determined by

$$0 = \sum \tilde{w}_i Z_i. \tag{A.3}$$

First we show that

$$\lambda = O_p(n^{-1/2}). \tag{A.4}$$

Introduce $\bar{Z} = (1/n)\sum Z_i$ and $Z^* = \max_{1 \le i \le n} \|Z_i\|$. Let $\lambda = \rho\theta$, where $\rho \ge 0$ and $\|\theta\| = 1$. From (A.2) and θ' times (A.3)

$$\begin{aligned}
0 &= \frac{1}{n}\theta' \sum \frac{Z_i}{1 + \rho\theta' Z_i} \\
&= \frac{1}{n}\theta' \sum Z_i - \rho \frac{1}{n} \sum \frac{(\theta' Z_i)^2}{1 + \rho\theta' Z_i} \\
&\le \theta'\bar{Z} - \frac{\rho}{1 + \rho Z^*} \frac{1}{n} \sum (\theta' Z_i)^2 \\
&= \theta'\bar{Z} - \frac{\rho}{1 + \rho Z^*} \theta' \hat{V}_n \theta \\
&\le \theta'\bar{Z} - \frac{\rho}{1 + \rho Z^*} \operatorname{mineig}\left(\hat{V}_n\right),
\end{aligned}$$

where we have used $0 < 1 + \rho\theta' Z_i \le 1 + \rho Z^*$ which follows from the positivity of $\tilde{w}_i$ and the definition of Z^*. Rearranging terms we find that

$$\rho\left(\operatorname{mineig}\left(\hat{V}_n\right) - \theta'\bar{Z}Z^*\right) \le \theta'\bar{Z}.$$

From (3.3b) and Chebychev's inequality, it follows that $\hat{V}_n - V_n = o_p(1)$ and so by (3.3c),

$$c + o_p(1) \le \operatorname{mineig}\left(\hat{V}_n\right) \le 1 + o_p(1).$$

It also follows from (3.3b) that

$$Z^* = o_p(n^{-1/2}) \tag{A.5}$$

and by the central limit theorem [(A.5) implies Lindberg's condition]

$$\theta'\bar{Z} = O_p(n^{-1/2})$$

so that $\rho = O_p(n^{-1/2})$ establishing (A.4).

Put $\gamma_i = \lambda' Z_i$. It follows from (A.4) and (A.5) that

$$\max_i |\gamma_i| = o_p(1). \tag{A.6}$$

Now from (A.2) and (A.3),

$$0 = \frac{1}{n} \sum Z_i - \frac{1}{n} \sum \lambda' Z_i Z_i + \frac{1}{n} \sum \frac{Z_i \gamma_i^2}{1 + \gamma_i}$$

from which

$$\lambda = \hat{V}_n^{-1}\overline{Z} + \delta,$$

where $\delta = o_p(n^{-1/2})$.

Because the γ_i are uniformly small, we may expand

$$\log(1 + \gamma_i) = \gamma_i - \gamma_i^2 + \eta_i,$$

where for some finite $B > 0$,

$$P\big(|\eta_i| \le B|\gamma_i|^3, 1 \le i \le n\big) \to 1$$

as $n \to \infty$.

Now $-2\log \mathscr{R}(0) = 2\sum \log(1 + \gamma_i)$ and after some algebra this reduces to

$$-2\log \mathscr{R}(0) = n\overline{Z}'\hat{V}_n^{-1}\overline{Z} - n\delta'\hat{V}_n^{-1}\delta + 2\sum \eta_i.$$

The lead term tends to $\chi^2_{(p)}$ by the central limit theorem, the second term is $o_p(1)$ by the bound on δ and because $\operatorname{maxeig}(\hat{V}_n)^{-1} \le c^{-1} + O_p(n^{-1/2})$ and finally

$$\Big|\sum \eta_i\Big| \le B\sum (\lambda' Z_i)^3 \le nB\|\lambda\| Z^* \sum (\lambda' Z_i)^2$$

$$\le nBZ^*\|\lambda\|^3 \operatorname{maxeig}\big(\hat{V}_n\big) = o_p(1). \qquad \square$$

There is no guarantee of a rate in the above theorem. By strengthening the moment conditions, the rate $O(n^{-1/2})$ should be attainable.

Acknowledgment. The author gratefully acknowledges helpful comments of an Associate Editor.

REFERENCES

ANDERSON, T. W. (1971). *The Statistical Analysis of Time Series Data*. Wiley, New York.

BERAN, R. (1986). Discussion of "Jackknife, bootstrap and other resampling methods in regression analysis" by C. F. J. Wu. *Ann. Statist.* **14** 1295–1298.

BLOCH, D. A. and MOSES, L. E. (1988). Nonoptimally weighted least squares. *Amer. Statist.* **42** 50–53.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

DAVIDIAN, M. and CARROLL, R. J. (1987). Variance function estimation. *J. Amer. Statist. Assoc.* **82** 1079–1091.

DICICCIO, T. J. and ROMANO, J. (1988a). Nonparametric confidence limits by resampling methods and least favorable families. Technical Report 295, Dept. Statistics, Stanford Univ.

DICICCIO, T. J. and ROMANO, J. (1988b). On adjustments to the signed root of the empirical likelihood ratio statistic. Technical Report 303, Dept. Statistics, Stanford Univ.

DICICCIO, T., HALL, P. and ROMANO, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19** 1053–1061.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canad. J. Statist.* **9** 139–172.

FREEDMAN, D. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218–1228.

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, Dept. Operations Research, Stanford Univ.

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

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

HINKLEY, D. (1986). Discussion of "Jackknife, bootstrap and other resampling methods in regression analysis" by C. F. J. Wu. *Ann. Statist.* **14** 1312–1316.

HINKLEY, D. and SCHECHTMAN, E. (1987). Conditional bootstrap methods in the mean-shift model. *Biometrika* **74** 85–94.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

JAMES, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika* **38** 324–329.

LIPPMAN, A. (1986). A maximum entropy method for expert system construction. Ph.D. dissertation, Div. Applied Mathematics, Brown Univ.

MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall, London.

MILLER, R. G., JR. (1974). An unbalanced jackknife. *Ann. Statist.* **2** 880–891.

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

OWEN, A. B. (1988b). Small sample central confidence intervals for the mean. Technical Report 302, Dept. Statistics, Stanford Univ.

OWEN, A. B. (1988c). Computing empirical likelihoods. In *Computing Science and Statistics, Proceedings of the 20th Symposium on the Interface* (E. J. Wegman, P. T. Gantz and J. J. Miller, eds.) 442–447. Amer. Statist. Assoc., Alexandria, Va.

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

RICE, J. A. (1988). *Mathematical Statistics and Data Analysis*. Wadsworth, Pacific Grove, Calif.

ROYALL, R. and CUMBERLAND, W. (1981). An empirical study of the ratio estimator and estimators of its variance (with discussion). *J. Amer. Statist. Assoc.* **76** 66–88.

RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134.

SHEEHY, A. (1987). Kullback–Leibler estimation of probability measures with an application to clustering. Ph.D. dissertation, Dept. Statistics, Univ. Washington, Seattle.

SAS INSTITUTE INC. (1985). SAS user's guide: Statistics. SAS Inst. Inc., Cary, N.C.

THOMAS, D. R. and GRUNKEMEIER, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70** 865–871.

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrika* **48** 817–838.

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

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

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305