Ann. Inst. Statist. Math.
Vol. 41, No. 2, 269–278 (1989)

POSSIBLE SUPERIORITY OF THE CONDITIONAL MLE OVER THE UNCONDITIONAL MLE

TAKEMI YANAGIMOTO AND KAZUO ANRAKU

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan

(Received April 8, 1988; revised November 2, 1988)

Abstract. The possibility that the conditional maximum likelihood estimator (CMLE) is superior to the unconditional maximum likelihood estimator (UMLE) is discussed in examples where the residual likelihood is obstructive. We observe relatively smaller risks of the CMLE for a finite sample size. The models in the study include the normal, inverse Gauss, gamma, two-parameter exponential, logit, negative binomial and two-parameter geometric ones.

Key words and phrases: Bias reduction, comparison of estimators, Kullback-Leibler loss, odds ratio, Pareto distribution, steep exponential distribution.

1. Introduction

Let $x_1,\dots,x_n$ be a sample of size n from a population with the density function $p(x;\theta,\mu)$, $\theta\in\Theta\subset R^1$ and $\mu\in M\subset R^1$. Suppose that there exists a statistic t such that the entire likelihood for $\boldsymbol{x}=(x_1,\dots,x_n)$ is factored into

$$p(\boldsymbol{x};\theta,\mu)=pc(\boldsymbol{x};\theta\,|\,t)\cdot pr(t;\theta,\mu)\,, \tag{1.1}$$

where pc stands for the conditional likelihood given t being free from μ and pr the residual likelihood. Traditionally, the parameter θ is called the structural parameter and μ is called the nuisance or incidental parameter, though our interest is often placed on both the parameters. Under this setup, conditional inference for θ has attracted researchers' attention. Under weak regularity conditions, the likelihood equation for the CMLE of θ is written as

$$\frac{\partial}{\partial\theta}\log pc(\boldsymbol{x};\theta\,|\,t)=0\,, \tag{1.2}$$

and the profile likelihood equation for the UMLE is written as

$$\frac{\partial}{\partial\theta}\log pc(\boldsymbol{x};\theta|t)+\frac{\partial}{\partial\theta}\log pr(t;\theta,\mu(\theta))=0\,, \tag{1.3}$$

with $\mu(\theta)$ being the unique solution of $\partial \log p(t;\theta,\mu)/\partial\mu = 0$.

We find two approaches justifying the use of conditional inference in the literature. One is to show that inference for θ based on the conditional likelihood is not inferior to that based on the unconditional likelihood, even though the residual likelihood contains some information on θ. The other is to show that conditional inference has optimality or advantages. The notion of ancillarity has been widely discussed for the former purpose (Fisher (1935), Dawid (1975) and Barndorff-Nielsen (1978, 1980)). The ancillarity of a statistic t means that t contains little or no information on θ. Kalbfleisch and Sprott (1970, 1973) studied the examples where t contains little information. The main results on optimality and the favorable properties of the CMLE are in Neyman and Scott (1948), Andersen (1970), Godambe (1976) and Lindsay (1982). However, these authors do not explicitly claim superiority of the CMLE over other estimators such as the UMLE, when the sample size is finite. We know the UMLE also has optimality in different contexts.

Yanagimoto (1987) asserted the possibility that the CMLE is properly superior to the UMLE under certain conditions, and introduced the notion of an obstructive residual likelihood. This notion means that the behavior of the profile likelihood for θ induced from the obstructive residual likelihood is undesirable. Therefore, such a residual likelihood is to be disregarded for obtaining θ by the maximum likelihood method. The disregard of the residual likelihood results in the CMLE. Although his assertion is intuitively appealing, further studies concerning the performance of the CMLE are necessary to confirm its relative usefulness.

The aim of the paper is to explore to what extent the CMLE is superior to the UMLE for a finite sample size in examples where the residual likelihood is obstructive. In Sections 2 and 3, general aspects of the two MLE's are discussed. Section 4 is devoted to a comparison study of the two MLE's in cases of continuous models using various risk functions. Comparison studies for discrete models are conducted in Section 5. Our conclusion is that the CMLE is superior to the UMLE, when a model is specified, and especially when a model has multiple strata. On the other hand the UMLE is sometimes more convenient, when multiple candidate models exist. In the final section, the other two familiar estimators, the jackknifed estimator and the uniformly minimum variance unbiased estimator (UMVUE), are briefly critiqued.

2. Advantages of the CMLE

In cases of a finite sample size we have two results by Godambe (1976) and Yanagimoto (1987). The latter author pointed out the undesirable behavior of the profile likelihood for θ induced from the residual likelihood. Let $pp(t;\theta) = pr(t;\theta,\mu(\theta))$ be the profile likelihood for θ from the residual likelihood; therefore, the derivative is equivalent to the second term in the left-hand side in (1.3). The residual likelihood is called obstructive, if there exists a subset of the support of t and a value of $\theta_m \in \bar{\Theta}$ with the closure of Θ such that $pp(t;\theta)$ is strictly decreasing in θ for $\theta \leq \theta_m$ and increasing for $\theta \geq \theta_m$ when t is in the subset, and it is constant when t is not in the subset. Here we assume that the subset has a positive probability for every θ and μ. Examples where the residual likelihood is obstructive for a suitable t are given in Table 1. Our primary interest will be placed on these models. Note that the examples also cover other ones given by the monotone parameter transformation because of the invariance property; such examples include the lognormal and Pareto distributions. The first three distributions, the normal, inverse Gauss and gamma, are members of the family of the steep exponential distributions (Bar-Lev (1984)) and also those of the reproductive exponential distributions (Blaesild and Jensen (1985)). Theorem 2.1 in Bar-Lev (1984) essentially shows the obstructiveness of the residual likelihood of the steep exponential distribution.

Under weak regularity conditions, the conditional likelihood equation (1.2) is unbiased. The definition of an obstructive residual likelihood

Table 1. Models with a distribution whose residual likelihood given a statistic t is obstructive. The sample mean and the minimum order statistic are denoted by $\bar{x}$ and $x_{(1)}$.

Model	Density function	θ_m	t
(1) Normal	$\frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2\theta}$	∞	$\bar{x}$
(2) Inverse Gauss	$\sqrt{\theta/2\pi x^3}\, e^{-\theta(x-\mu)^2/2\mu^2 x}$, $x > 0$	∞	$\bar{x}$
(3) Gamma	$\frac{1}{\Gamma(\theta)} \frac{x^{\theta-1}}{\mu^{\theta}} \exp - x/\mu$, $x > 0$	0	$\bar{x}$
(4) Two-Parameter Exponential	$\frac{1}{\theta} \exp - (x-\mu)/\theta$, $x > \mu$	∞	$x_{(1)}$
(5) Logit	$\binom{n}{x} \frac{\exp - x(\theta z + \mu)}{\{1 + \exp - (\theta z + \mu)\}^n}$, $x = 0,\ldots,n$	0	$\bar{x}$
(6) Negative Binomial	$\binom{x + 1/\theta - 1}{x} \mu^x (1-\mu)^{1/\theta}$, $x = 0, 1,\ldots$	∞	$\bar{x}$
(7) Two-Parameter Geometric	$\theta^{x-\mu}(1-\theta)$, $x = \mu, \mu+1,\ldots$	∞	$x_{(1)}$

suggests that the profile likelihood equation (1.3) is not unbiased in most cases. In fact, strict biasedness holds for every θ and μ and every model under study, except in the case of the logit model. It is difficult to prove biasedness of the equation (1.3) in the logit model, but numerical experiments support biasedness except in the case of $\theta = 0$. Except for the negative binomial model, the conditional likelihood equation can be rewritten as

$$s - \psi(\theta, n) = 0 ,$$

for a suitable statistic s. If the function ψ is free from n, the statistic $s = \psi(\hat{\theta}_c)$ with the CMLE $\hat{\theta}_c$ is an unbiased estimator of $\psi(\theta)$. The normal, inverse Gauss and two-parameter exponential models satisfy this condition. When (s, t) is sufficient for (μ, θ) and s is unbiased, it is also the UMVUE.

Godambe (1976) gave an optimality property of the conditional likelihood equation (1.3) among unbiased estimating equations. He defines a measure of an unbiased estimating equation, $g(\boldsymbol{x}; \theta)$ as

$$M(g) = \frac{V(g)}{\{E(\partial g/\partial\theta)\}^2} ,$$

which is regarded as that of the reciprocal of standardized sensitivity. Under certain regularity conditions, the estimating function appearing in the conditional likelihood equation attains the minimum for every θ. Note that his result holds for a finite sample size. The restriction on estimators induced from the solution of an unbiased estimating equation is intuitively appealing, but it is inconvenient for comparison of the CMLE with the UMLE.

To study asymptotic behaviors of the estimators, we note that there are two types of asymptotic cases. The usual one occurs where a sample size n tends to infinity. The other, which we call a sparse case, occurs where a model has multiple strata having bounded sample sizes, with the number of strata tending to infinity. The density function is expressed as

$$p(\boldsymbol{x}; \theta, \boldsymbol{\mu}) = \prod_{k=1}^{K} p_k(\boldsymbol{x}_k; \theta, \mu_k) .$$

We know the two MLE's are equally efficient in the usual asymptotic case. Neyman and Scott (1948) gave examples where the UMLE is inconsistent and also where it is consistent but inefficient in sparse asymptotic cases. Sufficient conditions that the CMLE be efficient and asymptotically normal in both the asymptotic cases were discussed in Andersen (1970). The measure $M(g)$ provides the asymptotic variance of the CMLE in the models in the study. We observe that inconsistency of the UMLE in sparse asymptotic cases is due to biasedness of the profile likelihood equation (1.3).

The above finite sample size results support the superiority of the CMLE. Because of the inconsistency of the UMLE in sparse asymptotic cases, our interest will be placed on the possible superiority of the CMLE when K is small, typically 1. Therefore, the comparison study in the case of a single stratum with a small or a moderate sample size will attract our primary attention. Although these general reasonings strongly support the superiority of the CMLE, a case study is necessary to confirm our assertion. Prior to the case study we also refer to advantages of the UMLE from other standpoints in order to make a fair comparison.

3. Advantages of the UMLE

Advantages of the UMLE with emphasis on its comparison to the CMLE have rarely been discussed. The following two facts may be known, but they have not been stressed explicitly in the literature. In addition, the importance of the two facts is not recognized in a correct way. One of the two is that the amount of computation for the CMLE is often greater than that for the UMLE. The other is that the UMLE provides all the estimates of all the parameters contained in a model without any additional principles.

Superficially, the numerical computation for the CMLE looks simpler, since the conditional likelihood in (1.1) is free from the parameter μ. However, solving the likelihood equation for the CMLE is regarded as a constrained maximization, while that for the UMLE is regarded as an unconstrained one. Actual examples show that the former is likely to be more elaborate than the latter, which is summarized in Table 2. It is

Table 2. Comparisons of rule of thumb estimates of amounts of computations for obtaining the CMLE and the UMLE. The "amount of computation" means the amount of computation for the CMLE relative to that for the UMLE.

Model	Solution		Amount of computation	Comment
	CMLE	UMLE		
(1) Normal	E	E	$\approx$	
(2) Inv. Gauss	E	E	$\approx$	
(3) Gamma	I	I	$>$	digamma, trigamma, Difficulty in initial value
(4) Exponential	E	E	$\approx$	
(5) 2×2 table	I	E	$>$	
(5') Logit	I	I	$\gg$	combinatorial
(6) Neg. Binomial	I	I	$\approx$	
(7) Geometric	I	E	$>$	

Symbols: E: explicit form is possible, I: iterative procedure is required, $\approx$: even, $>$: larger, $\gg$: much larger.

interesting that the CMLE is most widely accepted in the logit model, whereas the computation problem can be most serious.

Another advantage of the UMLE is that it presents estimates of both θ and μ by a common principle. The CMLE presents an estimate of θ only. Although the other parameter, μ, is assumed to be a nuisance in conditional inference, both the parameters can be of interest in practice. The maximized likelihood, which is given by $p(\boldsymbol{x}; \hat{\theta}_u, \hat{\mu}_u)$ with $\hat{\theta}_u$ and $\hat{\mu}_u$ being the UMLE, is useful in diagnosing the goodness of fit of the model compared with other candidate models in terms of deviance (Nelder and Wedderburn (1972), for example).

Although these two practical disadvantages of the CMLE may be dissolved by future research, the use of the CMLE is not recommended at present unless it behaves more favorably as an estimator than the UMLE. Fortunately, we can expect the superiority of the CMLE as discussed in the previous section. Therefore, our purpose in the following will be to show to what extent the CMLE is actually superior, when the residual likelihood is obstructive.

4. Case study—the continuous models

The result of a comparison study of estimators usually depends on the criteria employed. In general, it is unlikely to hold that one estimator is preferable to another estimator with respect to every criterion for every θ and μ. Therefore, we select several criteria for our comparison study.

The risks of an estimator $\hat{\theta}$ or $(\hat{\theta}, \hat{\mu})$ in cases of continuous models are:

(i) Bias of $\hat{\theta}$; $E(\hat{\theta} - \theta_0)$, MSE of $\hat{\theta}$; $E\{(\hat{\theta} - \theta_0)^2\}$,

(ii) Bias of $1/\hat{\theta}$, MSE of $1/\hat{\theta}$,

(iii) Bias of $\log \hat{\theta}$, MSE of $\log \hat{\theta}$,

(iv) K-L risk; $E\{KL(\hat{\theta}, \hat{\mu}; \theta_0, \mu_0)\} = E\Big[\int - \log\{p(\boldsymbol{z}; \theta_0, \mu_0)/p(\boldsymbol{z}; \hat{\theta}, \hat{\mu})\} \cdot p(\boldsymbol{z}; \hat{\theta}, \hat{\mu})\, d\boldsymbol{z}\Big]$,

where the expectation is taken with respect to $p(\boldsymbol{x}; \theta_0, \mu_0)$, and K-L risk denotes the Kullback-Leibler risk. To aid our understanding, explicit forms of the Kullback-Leibler loss to the normal, inverse Gauss and gamma models are exemplified as:

$$\frac{n}{2}\left\{\frac{\hat{\theta} + (\hat{\mu} - \mu_0)^2}{\theta_0} - \log\frac{\hat{\theta}}{\theta_0} - 1\right\},$$

$$\frac{n}{2}\left\{-\log\frac{\theta_0}{\hat{\theta}} + \frac{\theta_0}{\mu_0}\left(\frac{\hat{\mu}}{\mu_0} + \frac{\mu_0}{\hat{\mu}} - 2\right) + \frac{\theta_0}{\hat{\theta}} - 1\right\},$$

$$n\left\{-\log\frac{\Gamma(\hat{\theta})}{\Gamma(\theta_0)} - \theta_0 \log\frac{\hat{\mu}}{\mu_0} + (\hat{\theta} - \theta_0)\psi(\hat{\theta}) - \hat{\theta}\left(1 - \frac{\hat{\mu}}{\mu_0}\right)\right\},$$

with $\psi(\theta)$ being the digamma function. This seems to require some comments on the use of the bias and the MSE. We employ bias and the MSE for $1/\theta$ and also $\log\theta$, since the parameter of θ in a standard expression is not always of interest. In addition the criteria, the bias and the MSE are reliable, especially when an estimator $\hat{\theta}$ has support $(-\infty,\infty)$ and distributes symmetrically at a point. Since the range of θ in our examples is $(0,\infty)$, the logarithmic transformation changes it into $(-\infty,\infty)$. Recall that the parameter θ in the two-parameter exponential distribution corresponds to $1/\theta$ in the Pareto distribution by the exponential transformation. Furthermore, other transformations of θ may attract our interest. The Kullback-Leibler risk is employed as the risk of a simultaneous estimate of (θ,μ). Here the CMLE of μ is given by maximizing the entire likelihood for a given $\hat{\theta}_c$, $p(\boldsymbol{x};\hat{\theta}_c,\mu)$. We are often interested in an estimated model as well as estimates of parameters. In fact, the UMLE is regarded as an estimate of a model; this property is desirable for an estimate.

The risks of the two MLE's are compared. Since in the case of gamma distribution, analytical comparison appears impossible for some risks, simulation studies are applied. The results are summarized for comparison in Table 3. The results concerning the normal and gamma models are presented in Yanagimoto (1987) and Yanagimoto (1988*a*), respectively. Those concerning the remaining two models are obtained after straightforward calculations.

Table 3. Risk comparison between the CMLE and the UMLE. The symbol $<$ denotes smaller risk of the CMLE, and $*$ means that the evaluation is based on the simulation study.

Model	θ		$1/\theta$		$\log\theta$		K-L Loss
	Bias	MSE	Bias	MSE	Bias	MSE	
Normal	$<$	$>$	$<$	$<$	$<$	$<$	$<$
Inv. Gauss	$<$	$<$	$<$	$>$	$<$	$<$	$<$
Gamma	$<$	$<*$	$<$	$>*$	$<*$	$<*$	$<*$
Exponential	$<$	$>$	$<$	$<$	$<$	$<$	$<$

As expected, we observe that in most cases the CMLE presents a smaller risk than the UMLE. The exceptional cases are the MSE for θ in the normal and two-parameter exponential models, and that of $1/\theta$ in the inverse Gauss and gamma models. Yanagimoto (1988*a*) pointed out that the relative bias of $1/\hat{\theta}_c$ is small compared with that of $\hat{\theta}_c$. In this sense, the role of $1/\theta$ in the gamma model corresponds with θ in the normal model. We think that this result is associated with a possible defect of the MSE as a criterion. Note first that the UMLE in the above cases is downward biased; in fact, it holds $E(\hat{\theta}_u) < E(\hat{\theta}_c) < \theta_0$ or $E(1/\hat{\theta}_u) < E(1/\hat{\theta}_c) < 1/\theta_0$. We suspect that the smaller MSE of the UMLE is associated with its greater bias. The fact that the squared error takes the same value, θ_0^2, for $\hat{\theta} = 0$ and $2\theta_0$ is obviously undesirable in our examples. The MSE of an estimator of $\log \theta$ is free from this controversy, since the range of $\log \theta$ is $(-\infty, \infty)$.

The consistently smaller risk of the CMLE is also observed in the Kullback-Leibler risk. This fact looks surprising; it means that conditional inference may be useful in estimating both the parameters in a model as well as estimating simply a structural parameter. We conclude that the CMLE is a better estimator than the UMLE in continuous models where the residual likelihood is obstructive.

5. Case study—the discrete models

The comparison study for estimators in discrete models is more complicated. Although we obtained the inequalities of risks between the two MLE's in the continuous models, we can not find the corresponding inequalities in the discrete models. The analytical comparison study looks impossible, and even the numerical one is much more elaborate than that in the continuous models.

The comparison study of the two MLE's of the common log odds ratio in multiple 2×2 tables has been extensively performed (Lubin (1981) and Hauck *et al.* (1982)). The comparison study is extended to those of the logit model with multiple strata. Breslow and Cologne (1986) regard the CMLE as "a golden standard". Note that existing evidence of its superiority is concerned only with simulation studies of comparing bias. We conducted further comparison studies for a single stratum using various risks, and our results support the superiority of the CMLE, although we could not succeed in obtaining clear ones. Fortunately, there is no controversy over the conclusion of the CMLE's superiority in the literature, though further studies are necessary.

The use of the CMLE in the negative binomial models was recommended in Kalbfleisch and Sprott (1970), Godambe (1980) and Yanagimoto (1987), but no comparison study was conducted. As in the logit model, the negative binomial model also has multiple strata in practical applications (Bliss and Owen (1958), for example). Therefore, the possibility of the use of the CMLE attracts our attention. The negative binomial distribution is approximated by the gamma distribution, when μ is close to 1. We can expect from the results in the previous section the superiority of the CMLE in such a case.

Anraku and Yanagimoto (1988) conducted simulation studies using bias and the MSE for θ, $1/\theta$ and $\theta/(1+\theta)$, which are summarized as follows. The CMLE has the smaller bias for $1/\theta$ and $\theta/(1+\theta)$, for any θ_0 and μ_0 they employed. The CMLE has the smaller MSE for $1/\theta$ in all cases, and for $\theta/(1+\theta)$ in most cases. They conclude that results in terms of bias and MSE for $1/\theta$ and $\theta/(1+\theta)$ support the superiority of the

CMLE, but that results in terms of θ are confusing. When there are multiple strata, say 5 strata, simulation studies show the clear superiority of the CMLE.

Yanagimoto (1988*b*) introduced the CMLE in the two-parameter geometric model, and compared it to other estimators, including the UMLE. The limiting distribution of the two-parameter geometric distribution as θ tends to 1 is a two-parameter exponential one. Though his conclusion is not as definite as our conclusion in the continuous models, his results support the superiority of the CMLE.

6. Other estimators

Although our interest was restricted to the two MLE's, we know some estimators based on other principles could be useful. We discuss briefly the two familiar estimators among them: the UMVUE and the jackknifed estimator.

The UMVUE is appealing, if our interest is actually in the parameter. However, this assumption looks restrictive in practice. Recall that the likelihood equation for the CMLE is unbiased, which yields that the CMLE of a transformed parameter, $g(\theta)$, is the UMVUE. Therefore, the CMLE and the UMVUE coincide with each other, when the former is unbiased. The strict restriction on unbiasedness can result in an undesirable estimate. For example, the UMVUE can take the value 1 in the case of the two-parameter geometric model, though the estimated likelihood is zero.

The jackknifed estimator is introduced to eliminate first order bias in a general way. The estimator requires a large amount of computation, when the sample size is large. It coincides with the CMLE in cases of the normal and inverse Gauss models. However, the jackknifed estimator can not be recommended in our other examples. Consider the two-parameter exponential model. It is expressed as $\hat{\theta}_J = \bar{x} - x_{(1)} + (n-1)(x_{(2)} - x_{(1)})/n$. It is obviously less satisfactory than the CMLE, $n(\bar{x} - x_{(1)})/(n-1)$. As Yanagimoto (1988*a*) showed, the jackknifed estimator of the shape parameter in the gamma model takes a negative value with a positive probability.

Acknowledgements

The authors express their sincere thanks to the two referees for their kind, constructive comments, including editorial ones.

REFERENCES

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators, *J. Roy. Statist. Soc. Ser. B*, **32**, 283–301.

Anraku, K. and Yanagimoto, T. (1988). Estimation for the negative binomial distribution based on the conditional likelihood, Research Memorandum, No. 349, The Institute of Statistical Mathematics.

Bar-Lev, S. K. (1984). Asymptotic behaviour of conditional maximum likelihood estimators in a certain exponential model, *J. Roy. Statist. Soc. Ser. B*, **46**, 425–430.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*, Wiley, New York.

Barndorff-Nielsen, O. (1980). Conditionality resolutions, *Biometrika*, **67**, 293–310.

Blaesild, P. and Jensen, J. L. (1985). Saddlepoint formulas for reproductive exponential models, *Scand. J. Statist.*, **12**, 193–202.

Bliss, C. I. and Owen, A. R. G. (1958). Negative binomial distributions with a common k, *Biometrika*, **45**, 37–58.

Breslow, N. E. and Cologne, J. (1986). Methods of estimation in log odds ratio regression models, *Biometrics*, **42**, 949–954.

Dawid, A. P. (1975). On the concepts of sufficiency and ancillarity in the presence of nuisance parameters, *J. Roy. Statist. Soc. Ser. B*, **37**, 248–258.

Fisher, R. A. (1935). The logic of inductive inference, *J. Roy. Statist. Soc.*, **98**, 39–54.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations, *Biometrika*, **63**, 277–284.

Godambe, V. P. (1980). On sufficiency and ancillarity in the presence of a nuisance parameter, *Biometrika*, **67**, 155–162.

Hauck, W. W., Anderson, S. and Leahy, F. J. (1982). Finite-sample properties of some old and some new estimators of a common odds ratio from multiple 2×2 tables, *J. Amer. Statist. Assoc.*, **77**, 145–152.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussions), *J. Roy. Statist. Soc. Ser. B*, **32**, 175–208.

Kalbfleisch, J. D. and Sprott, D. A. (1973). Marginal and conditional likelihoods, *Sankhyā Ser. A*, **35**, 311–328.

Lindsay, B. (1982). Conditional score functions: Some optimality results, *Biometrika*, **69**, 503–512.

Lubin, J. H. (1981). An empirical evaluation of the use of conditional and unconditional likelihoods for case-control data, *Biometrika*, **68**, 567–571.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear model (with discussion), *J. Roy. Statist. Soc. Ser. A*, **34**, 370–384.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations, *Econometrica*, **16**, 1–32.

Yanagimoto, T. (1987). A notion of an obstructive residual likelihood, *Ann. Inst. Statist. Math.*, **39**, 247–261.

Yanagimoto, T. (1988*a*). The conditional maximum likelihood estimator of the shape parameter in the gamma distribution, *Metrika*, **35**, 161–175.

Yanagimoto, T. (1988*b*). The conditional MLE in the two-parameter geometric distribution and its competitors, *Comm. Statist. A—Theory Methods*, **17**, 2779–2787.