Ann. Inst. Statist. Math.
39 (1987), Part A, 247-261

A NOTION OF AN OBSTRUCTIVE RESIDUAL LIKELIHOOD

TAKEMI YANAGIMOTO

(Received Mar. 11, 1986; revised Nov. 14, 1986)

Summary

A new notion of an obstructive residual likelihood is proposed and explored. Examples where the conditional maximum likelihood estimator is preferable to the unconditional maximum likelihood estimator are discussed. In these examples the residual likelihood can be obstructive in deriving a preferable estimator, when the maximum likelihood criterion is applied. This notion is different from a similar notion ancillarity, which simply emphasizes that a residual likelihood is uninformative.

1. Introduction

Consider an estimation problem of a scalar parameter of interest θ based on $p(\boldsymbol{x}; \theta, \mu)$, under the existence of a nuisance (vector) parameter μ. We will assume that the dominating measure and the support are common with parameters. Suppose that there exists a statistic t such that the density is factored into $p(\boldsymbol{x}; \theta, \mu)=pc(\boldsymbol{x}; \theta|t)\cdot pr(t; \theta, \mu)$ where $pr(t; \theta, \mu)$ is the marginal density of t, that is, $pr(t; \theta,\mu)=\int_{\Omega} p(\boldsymbol{x}; \theta, \mu)d\boldsymbol{x}$ with $\Omega=\{\boldsymbol{x}|t(\boldsymbol{x})=t\}$. The former term is referred to as a conditional density, and the latter as a residual density. When we regard them as likelihoods, we will write L, LC and LR in place of p, pc and pr, respectively. In many situations an estimator of θ based on the conditional likelihood is recommended rather than that based on the unconditional likelihood. Neyman and Scott [17] presented explicit examples showing that the unconditional maximum likelihood estimator (m.l.e.) can be inconsistent, and also that it can be inefficient even if it is consistent. Andersen [2] showed that under mild regularity conditions the conditional m.l.e. is consistent and efficient.

Much effort has been devoted to study the reasons why the con-

Key words and phrases: Ancillary statistic, conditional likelihood, bias, consistency, maximum likelihood estimator, residual likelihood.

ditional m.l.e. can be superior to the unconditional m.l.e. Most interest has been paid to the notion of ancillarity introduced by Fisher [5]. A statistic t is called to be ancillary with respect to a parameter of interest θ, if the conditional distribution of $\boldsymbol{x}$ given t does not depend on μ and t is noninformative with respect to θ. When t is ancillary, it is believed that we can find an appropriate estimator based on the conditional likelihood given t, which is not inferior to all possible estimators based on the unconditional likelihood. Several definitions of ancillarity such as B, S, G, M, affine and weak ancillarity have been introduced (Barndorff-Nielsen [3] and [4]). Godambe [8] and [9] introduced and discussed other notions of ancillarity.

Our attention will be limited to the conditional and the unconditional m.l.e.'s. The limitation of our interest to the maximum likelihood criterion does not seem too restrictive in practice. When a model contains many nuisance parameters, an estimator other than that based on the maximum likelihood criterion is likely to be difficult to obtain. We know that some estimators of common odds ratio in multiple 2×2 tables are undesirable (see Fleiss [6]). It seems that the conditional m.l.e. is believed to be strictly superior to the unconditional m.l.e. Since the difference between the two estimators comes from the residual likelihood, the residual likelihood may be obstructive in deriving a preferable estimator of θ. On the other hand, it is obvious that the conditional m.l.e. is not always superior to the unconditional m.l.e. The notion of an obstructive residual likelihood is roughly defined as follows: If a residual likelihood satisfies a condition for obstructiveness, then the conditional m.l.e. is superior to the unconditional m.l.e.

The aim of this paper is to propose a definition of a notion of an obstructive residual likelihood. The notion is intuitively appealing and relates directly with possible biasedness and inconsistency of the unconditional m.l.e. A notion of an obstructive residual likelihood is introduced in Section 2. In the following section some properties are presented. Some examples of an obstructive residual likelihood with explicit form are discussed in Section 4. Section 5 provides an example in contrast, where a residual likelihood is not obstructive. Finally, two notions of ancillarity are discussed in Section 6.

2. Obstructive residual likelihood

Suppose that a scalar parameter θ in a likelihood $L(\boldsymbol{x}; \theta, \mu)$ is in an open interval I in R^1, and that the i-th component of a nuisance parameter $\mu=(\mu_1, \cdots, \mu_k)'$ is in an open interval J_i in R^1. The closure of I and J_i in the extended real value are written as $\bar{I}$ and $\bar{J}_i$. As in the Introduction we assume that there exists a statistic t such that

$$L(\boldsymbol{x};\theta,\mu)=LC(\boldsymbol{x};\theta|t)\cdot LR(t;\theta,\mu)\,. \tag{2.1}$$

We assume that each likelihood is first differentiable, and that each estimation equation has a unique solution. For simplicity of description, the ordinary derivative is denoted by a prime and the partial derivative is denoted by a subscript of a corresponding variable. It is convenient to rewrite (2.1) as $LL(\boldsymbol{x};\theta,\mu)=LLC(\boldsymbol{x};\theta|t)+LLR(t;\theta,\mu)$ with $LL(\cdot)=\log L(\cdot)$ and vice versa. The conditional m.l.e. is given by a unique solution of the estimation equation,

$$LLC'(\boldsymbol{x};\theta|t)=0\,. \tag{2.2}$$

On the other hand, the unconditional m.l.e. is given under regularity conditions by a unique solution of the estimation equation

$$LLC'(\boldsymbol{x};\theta|t)+LLR_{\theta}(t;\theta,\mu(\theta))=0 \tag{2.3}$$

with a unique solution $\mu(\theta)$ of the equation

$$LLR_{\mu_i}(t;\theta,\mu)=0 \qquad i=1,\cdots,k\,, \tag{2.4}$$

where we suppress t in $\mu(\theta,t)$.

Before introducing an obstructive residual likelihood, we proceed to recall some favorable properties of a likelihood $L(\boldsymbol{x};\theta)$ when no nuisance parameter exists. The log-likelihood $LL(\boldsymbol{x};\theta)$ is expected to be strictly increasing in θ up to $\hat{\theta}$ and strictly decreasing after $\hat{\theta}$; in other words, $LL(\boldsymbol{x};\theta)$ is unimodal. In addition $\hat{\theta}$ is required not to be a constant but to distribute around a true value θ_0. A regularity condition for this requirement is unbiasedness, $\mathrm{E}(LL'(\boldsymbol{x};\theta)|\theta)=0$. Regularity conditions for the asymptotic normality can be consulted in a standard textbook (Rao [18], for example). Our notion of obstructiveness corresponds to adverse properties in contrast with the above favorable ones. Write $R(t;\theta)=\sup_{\mu} LR(t;\theta,\mu)$, $(=LR(t;\theta,\mu(\theta)))$. Aitkin [1] called this type of an induced likelihood the profile likelihood.

DEFINITION 1. A residual likelihood $LR(t;\theta,\mu)$ is said to be obstructive in estimating θ, if there exists a point $\theta_m\in\bar{I}$, not depending on the data, and a subset T of the range of t, such that

(i) For a fixed $t\notin T$, $R(t;\theta)$ is strictly decreasing for $\theta\leqq\theta_m$, and it is strictly increasing for $\theta>\theta_m$.

(ii) For a fixed $t\in T$, $R(t;\theta)$ is constant, and $\Pr(T|\theta,\mu)<1$ for any θ and μ.

The point θ_m is said to be the stable minimum point.

The condition (i) is essential in the above definition. The definition means that a global behavior of $R(t;\theta)$ is reversely unimodal

when the stable minimum point is in I, and that it is strictly monotone when the point is on a boundary of $\bar{I}$. Therefore, $R(t;\theta)$ always attains the maximum at a boundary of $\bar{I}$.

We introduce another notion of a residual likelihood, which means that the residual likelihood is useless in distinguishing the difference between the conditional and the unconditional m.l.e.'s.

DEFINITION 2. A residual likelihood $LR(t;\theta,\mu)$ is said to be plain in estimating θ, if $R(t;\theta)$ is constant for any t.

The above definition is interpreted as a limiting case of the condition (ii) in Definition 1. It is obvious that both the m.l.e.'s are identical, when the residual likelihood is plain.

3. Some properties

First, we note that the notion of an obstructive residual likelihood is invariant under a strictly monotone transformation. The invariance property is presumed in likelihood estimation.

PROPOSITION 1. (i) *Suppose* $f(x)$ *in a first differentiable and strictly monotone function on the common support of* $p(\boldsymbol{x};\theta,\mu)$. *Let* $\boldsymbol{y}=(y_1,\cdots,y_n)'=(f(x_1),\cdots,f(x_n))$ *and* $s=s(\boldsymbol{y})=t(f^{-1}(y_1),\cdots,f^{-1}(y_n))$. *If* $LR(t;\theta,\mu)$ *is obstructive, then the residual likelihood of the induced likelihood* $LR(s;\theta,\mu)$ *is also obstructive.*

(ii) *Suppose* g *is strictly monotone in* θ *with the domain* I, *and let* $h=(h_1,\cdots,h_k)$ *be strictly monotone in* μ, *componentwise, with the domain* $J_1\times\cdots\times J_k$. *If* $LR(t;\theta,\mu)$ *is obstructive, then* $LR(t;g(\theta),(h_1(\mu_1),\cdots,h_k(\mu_k))')$ *is also obstructive.*

The following two properties show that the formal application of the maximum likelihood criterion to $pr(t;\theta,\mu)$ yields an undesirable estimator. In Propositions 3 and 4 we presume that the conditional m.l.e. is unbiased and consistent in contrast with the unconditional m.l.e. In fact, we know that the assumption is often true.

PROPOSITION 2. *Suppose that crude values of a sample* $\boldsymbol{x}$ *are unavailable and that only a summary statistic* $t(\boldsymbol{x})$ *is available. Then the m.l.e. of* θ *is on the boundaries,* $\bar{I}-I$.

PROPOSITION 3. *Suppose that the residual likelihood is obstructive. Let* $\hat{\theta}_c$ *and* $\hat{\theta}_u$ *be the conditional and the unconditional m.l.e.'s, respectively. Then it holds that for* $t\in T$: (i) *If* θ_m *is on the lower boundary of* I, *then* $\hat{\theta}_u>\hat{\theta}_c$, (ii) *if* $\theta_m\in I$, *then* $\hat{\theta}_u>\hat{\theta}_c>\theta_m$, $\theta_m>\hat{\theta}_c>\hat{\theta}_u$ *or* $\hat{\theta}_u=\hat{\theta}_c=$

θ_m and (iii) *if θ_m is on the upper bound of I, then $\hat{\theta}_u < \hat{\theta}_c$.*

COROLLARY. *If θ_m is on a boundary of $\bar{I}$ and $\hat{\theta}_c$ is unbiased, then $\hat{\theta}_u$ is not unbiased.*

Next, we consider possible inconsistency of the unconditional m.l.e. Let $\{p_s(\boldsymbol{x}_s; \theta, \mu_s)\}$ be a sequence of density functions, where n_s, the dimension of $\boldsymbol{x}_s$, is strictly increasing in s. Suppose that $p_s(\boldsymbol{x}_s; \theta, \mu_s)$ is factored into $pc_s(\boldsymbol{x}_s; \theta | t_s) \cdot pr_s(t_s; \theta, \mu_s)$. We say that a sequence of residual likelihoods has the common stable minimum point θ_m, if every residual likelihood has a stable minimum point common with s. Here we regard any point as the stable minimum point when the residual likelihood is plain.

The unconditional m.l.e. is likely to be inconsistent when the number of strata tends to infinity as $s \to \infty$. We outline the reason as follows: Let

$$\boldsymbol{x}'_{s+1} = (\boldsymbol{x}'_s, \boldsymbol{x}^{s+1\prime}) , \qquad \mu'_{s+1} = (\mu'_s, \mu^{s+1\prime})$$

and

$$p(\boldsymbol{x}_s; \theta, \mu_s) = \prod_{j=1}^{s} p_j(\boldsymbol{x}^j; \theta, \mu^j) = \prod_{j=1}^{s} \{pc_j(\boldsymbol{x}^j; \theta | t_j) \cdot pr_j(t_j; \theta, \mu^j)\} .$$

When $LR_j(t_j; \theta, \mu^j)$ has a common stable minimum point, all the first derivatives of $R_j(t_j; \theta)$ have a common sign or they are 0. Let n^i be the dimension of $\boldsymbol{x}^i$, and k^i be the dimension of μ^i. Suppose that n^i is bounded and the norm of μ_0^i is bounded. Then we can expect that $LLC'_s(\boldsymbol{x}_s; \theta_0 | t_s)/\sqrt{n_s - k_s}$ has a normal asymptotic distribution while $LLR'_s(\boldsymbol{x}_s; \theta_0, \mu_s(\theta_0))/(n_s - k_s)$ converges in probability to a nonzero constant.

The following proposition is useful in checking the possible inconsistency of the unconditional m.l.e. in practical examples.

PROPOSITION 4. *Suppose that standard regularity conditions on $LL_s(\boldsymbol{x}_s; \theta, \mu_s)$ and $LLC_s(\boldsymbol{x}_s; \theta | t_s)$ concerning the first and the second derivatives for asymptotic normality are satisfied. Assume that $LLR'_s(t_s; \theta_0, \mu_s(\theta_0))/(n_s - k_s)$ does not converge to 0 in probability when the true values of the parameters are θ_0 and μ_{s0}. Then the unconditional m.l.e. is not consistent.*

The S-ancillarity is widely used. A statistic t is said to be S-ancillary if the family of marginal densities $\{pr(t; \theta, \mu)\}$ given θ does not depend on θ. This restriction seems to be too strong. In fact, as we will discuss in Section 4, the conditional m.l.e. is widely accepted, even when t in (2.1) is not S-ancillary. A weaker and more practical notion M-ancillarity will be discussed in Section 6.

PROPOSITION 5. *If a statistic t is S-ancillary, then the residual likelihood is plain.*

4. Examples

In this section we give examples with explicit forms of likelihoods of which residual likelihoods are obstructive. It may appear that the notion of an obstructive likelihood is restrictive. We find, however, that some likelihoods appearing in familiar models satisfy the conditions of the notion. The estimation of the variance in a normal population and that of the common odds ratio especially are of practical importance. They will be discussed in Examples 4.1 and 4.3.

Example 4.1. (Multiple regression) Let $\boldsymbol{x}=(x_1, \cdots, x_n)'$ be a sample from a multiple regression model,

$$\boldsymbol{x}=Z\mu+\theta\varepsilon\,,$$

where Z is an $n\times k$ design matrix of rank $k<n$, ε is an n-dimensional normal error with mean 0 and variance-covariance matrix I. The parameter spaces of θ and μ are $(0, \infty)$ and R^k. Let t be the least square estimator of μ, $t=(Z'Z)^{-1}Z'\boldsymbol{x}$. The likelihood is factored into

$$\begin{aligned}(4.1)\qquad L(\boldsymbol{x};\theta,\mu)&=\frac{1}{(\sqrt{2\pi}\,\theta)^n}\exp-\frac{1}{2\theta^2}\|\boldsymbol{x}-Z\mu\|^2\\&=\frac{1}{(\sqrt{2\pi}\,\theta)^{n-k}}\exp-\frac{1}{2\theta^2}\boldsymbol{x}'(I-Z(Z'Z)^{-1}Z')\boldsymbol{x}\\&\quad\cdot\frac{1}{(\sqrt{2\pi}\,\theta)^k}\exp-\frac{1}{2\theta^2}\|Z(Z'Z)^{-1}Z'\boldsymbol{x}-Z\mu\|^2\\&=LC(\boldsymbol{x};\theta|t)\cdot LR(t;\theta,\mu)\,.\end{aligned}$$

It follows that the residual likelihood is obstructive with the stable minimum point $\theta_m=\infty$. The conditional likelihood is regular, unless $\boldsymbol{x}=Z(Z'Z)^{-1}Z'\boldsymbol{x}$, which holds with probability 0 for any θ_0 and μ_0. From Proposition 3 it follows that $\hat{\theta}_c>\hat{\theta}_u$ with probability 1. The conditional and the unconditional m.l.e.'s are expressed explicitly by $\hat{\theta}_c^2=n\hat{\theta}_u^2/(n-k)$ and $\hat{\theta}_u^2=\boldsymbol{x}'(I-Z(Z'Z)^{-1}Z')\boldsymbol{x}/n$. The conditional m.l.e. is unbiased.

Before comparing the risks between these estimators we consider a sequence of models,

$$(4.2)\qquad \boldsymbol{x}_s=Z_s\mu+\theta\varepsilon\,,$$

where Z_s is a design matrix of rank k_s $(<n_s)$ and $\varepsilon\sim N(0, I_{n_s})$. The number n_s is assumed to be strictly increasing. When we consider the sequence of estimators, we will write $\hat{\theta}_{cs}$ and $\hat{\theta}_{us}$. The following two

facts are well known: (i) The conditional m.l.e. is unbiased while the unconditional m.l.e. is strictly downward biased. (ii) $\hat{\theta}_{cs}$ is consistent. Unless the limit k_s/n_s is 0 as s tends to infinity, $\hat{\theta}_{us}$ is not consistent. Consider the special model that Neyman and Scott [17] discussed; the model is given by setting

$$Z_{s+1}=\begin{bmatrix} Z_s, & 0 \\ 0, & e_{s+1} \end{bmatrix}$$

in (4.2), where e_{s+1} is an m_{s+1} (>2)-dimensional vector with each component 1. They showed that the unconditional m.l.e. is not always consistent. It is easily shown that $\hat{\theta}_{us}$ is not efficient, unless the limit of $k_s/\sqrt{n_s}$ is 0. This result presents a simpler example of a consistent but inefficient estimator than that given by Neyman and Scott [17].

Next we compare the risks of the two estimators under a reasonable loss. Modifying the Kullback-Leibler separator, we employ a loss function of an estimator $(\hat{\theta}, \hat{\mu})$, which is expressed by

$$\text{Loss}\,(\hat{\theta}, \hat{\mu}\,|\,\theta_0, \mu_0)=\mathrm{E}\left\{-\log \frac{p(\boldsymbol{w}; \theta_0, \mu_0)}{p(\boldsymbol{w}; \hat{\theta}, \hat{\mu})}\,\middle|\, p(\boldsymbol{w}; \hat{\theta}, \hat{\mu})\right\}, \tag{4.3}$$

where (θ_0, μ_0) is a true value. This becomes, in our case,

$$n \log \theta_0+\frac{n\hat{\theta}^2+\|Z\mu_0-Z\hat{\mu}\|^2}{2\theta_0^2}-n \log \hat{\theta}+\frac{n}{2}.$$

The risk of an estimator $(\hat{\theta}, \hat{\mu})$ is defined by $\text{Risk}\,(\hat{\theta}, \hat{\mu}; \theta_0, \mu_0)=\mathrm{E}(\text{Loss}\,(\hat{\theta}, \hat{\mu}\,|\,\theta_0, \mu_0)\,|\,p(\boldsymbol{x}; \theta_0, \mu_0))$. Note that the likelihood (4.1) is maximized at $\hat{\mu}=(Z'Z)^{-1}Z'\boldsymbol{x}$ and that $\mathrm{E}(-\log \hat{\theta}_c\,|\,p(\boldsymbol{x}; \theta_0, \mu_0))$ is bounded. It follows that

$$\begin{aligned}
&\text{Risk}\,(\hat{\theta}_c, \hat{\mu}; \theta_0, \mu_0)-\text{Risk}\,(\hat{\theta}_u, \hat{\mu}; \theta_0, \mu_0) \\
&=\mathrm{E}\left\{n\hat{\theta}_c^2(1-(n-k)/n)/2\theta_0^2+\frac{n}{2}\log (n-k)-\frac{n}{2}\log n\,\middle|\, p(\boldsymbol{x}; \theta_0, \mu_0)\right\} \\
&=\frac{n}{2}\{k/n+\log (1-k/n)\}<0.
\end{aligned}$$

The result is summarized as follows: Under the loss in (4.3) the risk of $(\hat{\theta}_c, \hat{\mu})$ is strictly less than that of $(\hat{\theta}_u, \hat{\mu})$ for any (θ_0, μ_0). In other words the latter is inadmissible for every k and n.

Example 4.2. (Single measurement) Assume that a known density $f(x)$ is strongly unimodal, symmetric at 0 and positive on R^1. Let

$$\boldsymbol{x}=Z\eta+\theta\varepsilon, \tag{4.4}$$

where ε is an n-dimensional random error with mutually independent components having the common density $f(x)$, and Z is an $n\times(k+1)$ design matrix for a positive integer k less than n such that

$$Z=\begin{bmatrix} I_k, & 0 \\ 0, & e_{n-k} \end{bmatrix},$$

where e_{n-k} is the $(n-k)$-dimensional vector with each component 1. Let $\eta'=(\mu', c)$ with a known constant c and set $t=(x_1, \cdots, x_k)'$. The likelihood is factored into

$$(4.5) \qquad L(\boldsymbol{x}; \theta, \mu)=\left\{\prod_{i=k+1}^{n} \frac{1}{\theta} f((x_i-c)/\theta)\right\}\left\{\prod_{i=1}^{k} \frac{1}{\theta} f((x_i-\mu_i)/\theta)\right\}$$
$$=LC(\boldsymbol{x}; \theta|t)\cdot LR(t; \theta, \mu)\,.$$

The model may be interpreted as follows. Suppose that n observations are obtained to estimate measurement error using a standard instrument with a known weight c. But k measurements among them are found to be obtained using different instruments with unknown weights. Note that the conditional likelihood disregards measurements obtained using false instruments. Note also that the formal equivalence of (4.1) and (4.5) when $f(x)$ is the standard normal density and $c=0$.

The residual likelihood is obstructive with the stable minimum point $\theta_m=\infty$ and the conditional likelihood satisfies regularity conditions. Proposition 3 implies that $\hat{\theta}_c>\hat{\theta}_u$. The likelihood equations based on the conditional and the unconditional likelihoods are written as

$$-\sum_{i=k+1}^{n} \frac{f'((x_i-c)/\theta)}{f((x_i-c)/\theta)}\cdot\frac{x_i-c}{\theta^2}-\frac{g}{\theta}=0$$

with $g=n-k$ and n, respectively. Roughly speaking, the difference between the two estimators increases in k/n is large.

Example 4.3. (Logit model) Let $\boldsymbol{x}=(x_1, \cdots, x_n)'$, $n\geqq 2$ and x_i be a sample from binomial distribution with the incidence probability $\mathrm{p}(x_i; \theta, \mu)=\exp(z_i\theta+\mu)/(1+\exp(z_i\theta+\mu))$ where $\sum(z_i-\bar{z})^2\neq 0$. Kalbfleish and Sprott [12] studied this case in their Example 2.2 with restricted values of z under the different notation. The likelihood is factored into

$$(4.6) \qquad L(\boldsymbol{x}; \theta, \mu)=\prod_{i=1}^{n}\{\exp x_i(z_i\theta+\mu)/(1+\exp(z_i\theta+\mu))\}$$
$$=\left\{\prod_{i=1}^{n}\exp x_i(z_i\theta)\Big/\sum_{\phi\in\Phi}\prod_{j\in\phi}\exp(z_j\theta)\right\}$$
$$\times\left\{\sum_{\phi\in\Phi}\prod_{j\in\phi}\exp(z_j\theta)\frac{\exp t\mu}{\prod_i(1+\exp(z_i\theta+\mu))}\right\}$$
$$=LC(\boldsymbol{x}; \theta|t)\cdot LR(t; \theta, \mu)\,,$$

where $t=\sum x_i$ and Φ denotes the set of all subsets $\phi=\{\phi_1, \cdots, \phi_t\}$ of size t from $\{1, 2, \cdots, n\}$. A theorem in Yanagimoto and Kamakura

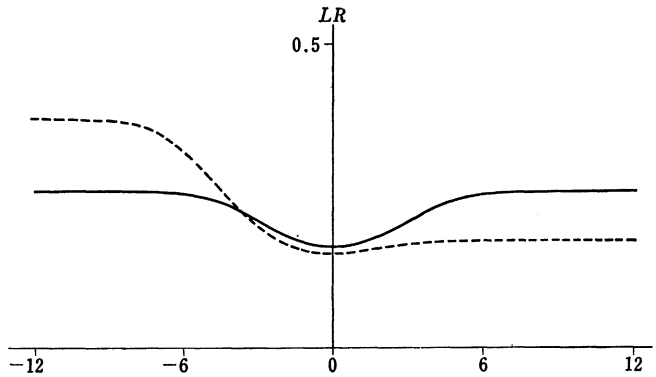

Fig. 1. Behavior of $R(t;\theta)$ induced from the logit model where the covariates consist of N1's and M0's: The real line shows the case $(N, M, t)=(12, 12, 9)$ and the dashed line $(N, M, t)=(20, 9, 10)$.

[19] implies that the residual likelihood is obstructive with the stable minimum point $\theta_m=0$. The exceptional sample region T in Definition 1 is $\{0, n\}$. An illustrative figure is given in Figure 1. When t is 0 or n, both the conditional and residual likelihoods are constant. When $\theta_0=0$, Proposition 3 yields $|\hat{\theta}_u|>|\hat{\theta}_c|$ if either of the two exists and is not equal to 0. Note that $\theta_0=0$ means all the binomial distributions have a common incidence probability.

The difference between the conditional and the unconditional m.l.e.'s is likely to become large, when many strata are taken into account as in Example 4.1. Many strata often appear, for instance, in case-control studies. Let $\boldsymbol{x}_s'=(\boldsymbol{x}_{s-1}', \boldsymbol{x}^{s'})$ with $\boldsymbol{x}^i=(x_{i1}, \cdots, x_{ini})'$, and let each $\boldsymbol{x}^i$ have the likelihood given by replacing $\boldsymbol{x}$, μ, z_i in (4.6) with $\boldsymbol{x}^i$, μ_i, z_{ij}. Setting $t_i=\sum x_{ij}$ $(=n_i\bar{x}_i)$, we find that the residual likelihood is obstructive with the common stable minimum point $\theta_m=0$. The conditional likelihood satisfies standard regularity conditions.

When each covariate z_{ij} takes only dichotomous values 0 and 1, the problem for estimating θ is known as that for estimating the common log-odds ratio. Much attention has been paid to this specific problem, which can be referred to in Fleiss [6] and Gart [7]. Hauck et al. [10] and Lubin [16] conducted simulation studies, and concluded that $\hat{\theta}_{cs}$ is superior to $\hat{\theta}_{us}$ under various conditions. Since $\hat{\theta}_{cs}$ is shown to be consistent under very mild conditions, $\hat{\theta}_{us}$ is likely to be inconsistent as shown in Proposition 4. The bias can be large when all the n_i are small. Especially when $n_i=2$ for any i and $\hat{\theta}_{cs}$ exists, it is known that $\hat{\theta}_{cs}=(1/2)\hat{\theta}_{us}$. Yanagimoto and Yamamoto [20] suggested that $LLC'(\boldsymbol{x}_s;\theta|t_1, \cdots, t_s)=\sum LLC_i'(\boldsymbol{x}^i;\theta|t_i)$ is well approximated by $\sum(\partial/\partial\theta)LL_i(\boldsymbol{x}^i;(n_i-1)\theta/n_i, \mu_i((n_i-1)\theta/n_i))$, where $\mu_i(\theta)$ is the unique solution of $(\partial/\partial\mu_i)LLR_i(t_i;\theta, \mu_i(\theta))=0$. This means that $\hat{\theta}_{us}$ is approxi-

mated by $m\hat{\theta}_{cs}/(m-1)$ when $n_i=m$ for every i.

Example 4.4. (Negative binomial distribution) Let $\boldsymbol{x}=(x_1,\cdots,x_n)'$ be a sample from a negative binomial distribution with probability function

$$\left\{\prod_{i=1}^{n}\binom{\theta+x_i-1}{x_i}\right\}\mu^{\Sigma x_i}(1-\mu)^{n\theta} \qquad (=L(\boldsymbol{x};\theta,\mu))\,.$$

The parameter spaces of θ and μ are $(0,\infty)$ and $(0,1)$, respectively. Let $t=\sum x_i$. The likelihood can be factored into

$$\left\{\prod_{i=1}^{n}\binom{\theta+x_i-1}{x_i}\right\}\Big/\binom{n\theta+t-1}{t}\cdot\binom{n\theta+t-1}{t}\mu^t(1-\mu)^{n\theta}$$
$$=LC(\boldsymbol{x};\theta|t)\cdot LR(t;\theta,\mu)\,.$$

Since the maximum of $LR(t;\theta,\mu)$ for a fixed θ is attained uniquely at $\mu(\theta)=t/(n\theta+t)$, it follows that

$$R(t;\theta)=\frac{(n\theta+t-1)\cdots(n\theta)}{t!}\,\frac{t^t(n\theta)^{n\theta}}{(n\theta+t)^{n\theta+t}}\,.$$

From this we get the residual likelihood is obstructive with the stable minimum point $\theta_m=0$, if $t>0$. Since $\log LC(\boldsymbol{x};\theta|t)$ is concave in θ when $t>0$, the conditional likelihood is unimodal when $t>0$. Proposition 3 implies that $\hat{\theta}_c<\hat{\theta}_u$ when either of the two exists.

Kalbfleish and Sprott [12] briefly discussed this model with strata. The likelihood of $\boldsymbol{x}_s'=(\boldsymbol{x}^{1\prime},\cdots,\boldsymbol{x}^{ns\prime})$ is

$$L(\boldsymbol{x}_s;\theta,\mu_s)=\prod_{i=1}^{s}L_i(\boldsymbol{x}^i;\theta,\mu^i)\,, \tag{4.7}$$

where $\mu_s=(\mu^1,\cdots,\mu^s)'$. They gave a figure for $LC(\cdot)$ and $LR(\cdot)$, and concluded that the behavior of $LC(\cdot)$ is regular and $LR(\cdot)$ does not vary rapidly when θ is not too small. Our approach provides clearer insight. Under the model (4.7) and mild regularity conditions $\hat{\theta}_{cs}$ is consistent, while $\hat{\theta}_{us}$ is not necessarily consistent. The unconditional m.l.e. is likely to be upward biased.

Other examples of an obstructive residual likelihood appear in the gamma distribution with its density $(x^{\theta-1}/\mu^\theta)(\exp -x/\mu)/\Gamma(\theta)$ and the exponential distribution with unknown support with its density $(1/\theta)\cdot\exp -(x-\mu)/\theta$. In fact, the conditional distribution given a sample mean from the gamma distribution is a function only of the shape parameter, and the residual likelihood is obstructive. This suggests the superiority of the conditional m.l.e. to the unconditional m.l.e. To make sure, further discussion is necessary. This will be done else-

where. In the latter example Lindsay [15] shows the factorization of the likelihood and possible inconsistency of the unconditional m.l.e. by setting $t=x_{(1)}$, the smallest order statistic.

5. Further example

In the preceding section we gave examples of obstructive residual likelihoods. It is obvious that a residual likelihood does not necessarily satisfy the condition, even though the full likelihood is formally factored into two terms as in (2.1). In this section we discuss an example where the residual likelihood is not obstructive. We will find that the conditional m.l.e. is not necessarily recommended, even when it possesses an optimum property. This example shows that we should look carefully at the behavior of the residual likelihood before we apply the conditional m.l.e.

Example 5.1. Let the first m components of $\boldsymbol{x}=(x_1, \cdots, x_n)'$, $m<n$, be

$$x_i=c+\theta\varepsilon_i \tag{5.1}$$

and the remaining components for $i>m$ be

$$x_i=c+\theta\varepsilon_i+\mu\eta_i\,, \tag{5.2}$$

where c is a known constant, $\varepsilon=(\varepsilon_1, \cdots, \varepsilon_n)'$ and $\eta=(\eta_{m+1}, \cdots, \eta_n)'$ are independent normal errors, $N(0, I_n)$ and $N(0, I_{n-m})$. The common parameter space of θ and μ is $(0, \infty)$. This setup is simple and realistic, since the latter $n-m$ values can be regarded as measurements with additional errors, contaminated in a standard measurement process, by which the first m measurements are obtained.

Set $t=(x_{m+1}, \cdots, x_n)'$. The likelihood is factored into

$$\begin{aligned} L(\boldsymbol{x}; \theta, \mu) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\,\theta} \exp\left(-\frac{1}{2\theta^2}(x_i-c)^2\right) \\ &\quad \times \prod_{i=m+1}^{n} \frac{1}{\sqrt{2\pi(\theta^2+\mu^2)}} \exp\left(-\frac{1}{2(\theta^2+\mu^2)}(x_i-c)^2\right) \\ &= LC(\boldsymbol{x}; \theta|t)\cdot LR(t; \theta, \mu)\,. \end{aligned} \tag{5.3}$$

Write $s_1=\sum_{i=1}^{m}(x_i-c)^2$ and $s_2=\sum_{i=m+1}^{n}(x_i-c)^2$. It follows that

$$LR(t; \theta, \mu(\theta))=\begin{cases} \dfrac{1}{(\sqrt{2\pi}\,\theta)^{n-m}} \exp\left(-\dfrac{s_2}{2\theta^2}\right) & \theta^2>s_2/(n-m) \\ \dfrac{1}{\{2\pi s_2/(n-m)\}^{(n-m)/2}} \exp\left(-\dfrac{(n-m)}{2}\right) & \theta^2\leqq s_2/(n-m)\,. \end{cases}$$

The behavior of $LR(t; \theta, \mu(\theta))$ is close to that in the previous example. It is constant up to $s_2/(n-m)$ and decreases strictly from that point. The difference between the two pertains to the upper or the lower side of a plateau part. The formal application of the maximum likelihood criterion to $LR(t; \theta, \mu(\theta))$ implies that $\hat{\theta}^2 \leqq s_2/(n-m)$. The interval estimator seems to be intuitively reasonable, since the observed variance is expected to be stochastically larger than θ^2. The behavior of $LR(t; \theta, \mu(\theta))$ leads the inequality $\hat{\theta}_u \leqq \hat{\theta}_c$. In this case both the estimators can be explicitly expressed as $\hat{\theta}_c^2 = s_1/m$ and $\hat{\theta}_u^2 = \mathrm{Min}\,(s_1/m, (s_1+s_2)/n)$. The conditional m.l.e., $\hat{\theta}_c^2$, is an unbiased estimator of θ^2. Note that s_1 and s_2 are sufficient to θ and μ, and also that the family of distributions of (5.3) is complete; these imply that the conditional m.l.e. is the uniformly minimum variance unbiased estimator.

However, the behavior of $LR(t; \theta, \mu(\theta))$ suggests that the unconditional m.l.e. may still be superior to the conditional m.l.e. In fact, consider a sequence of the two estimators under a situation that $m_s = m$ for any s and n_s tends to infinity. The limit of the estimators are $\hat{\theta}_{c\infty}^2 = s_1/m$ and $\hat{\theta}_{u\infty}^2 = \mathrm{Min}\,(s_1/m, \theta_0^2+\mu_0^2)$. The inequality $|\hat{\theta}_{c\infty}^2 - \theta_0^2| \geqq |\hat{\theta}_{u\infty}^2 - \theta_0^2|$ holds for each sample, where the strict inequality holds with positive probability for any θ_0 and μ_0. This means that the residual likelihood has certainly a little but positive information. When both m_s and $n_s - m_s$ tend to infinity, the probability $\Pr(\hat{\theta}_{cs}^2 = \hat{\theta}_{us}^2)$ tends to unity. There is no need to distinguish between the two estimators. The residual likelihood is asymptotically plain.

Finally, we compare the two estimators in a finite sample case. Consider a simple combination of $m = n-m = 2$ and $\mu_0 = 0$. Recall that the chi-square distribution with 2 degrees of freedom is the exponential distribution with mean 2. The density functions of $\hat{\theta}_c^2$ and $\hat{\theta}_u^2$ are expressed by

$$g_c(\theta) = \frac{1}{\theta_0^2} \exp -\theta/\theta_0^2$$

and

$$g_u(\theta) = \frac{1}{2\theta_0^2} \exp -2\theta/\theta_0^2 + \frac{2\theta}{\theta_0^4} \exp -2\theta/\theta_0^2 .$$

The loss given in (4.3) in Example 4.1 is written as $-\log(\hat{\theta}^2/\theta_0^2) + \hat{\theta}^2/\theta_0^2 - \log\{(\hat{\theta}^2+\hat{\mu}^2)/(\theta_0^2+\mu_0^2)\} + (\hat{\theta}^2+\hat{\mu}^2)/(\theta_0^2+\mu_0^2) - 2$. Therefore, the difference of the risk of $(\hat{\theta}_u, \hat{\mu}_u)$ to that of $(\hat{\theta}_c, \hat{\mu}_c)$ reduces to $\int \{-\log(\theta/\theta_0^2) + \theta/\theta_0^2\}(g_u(\theta) - g_c(\theta))d\theta = (1/2)\{2\log 2 - \psi(1) - \psi(2) + 3/2\} - (-\psi(1)+1) = \log 2 - 3/4 < 0$.

6. Comparison with notions of ancillarity

The notion of M-ancillarity was introduced by Barndorff-Nielsen [3], and is defined as; there exists $\mu(t;\theta)$ for any fixed θ such that $pr(t;\theta,\mu(t;\theta))\geqq pr(t';\theta,\mu(t;\theta))$ for any t'. Since the notions of B- and S-ancillarity and weak ancillarity are very strong, this notion is of more practical interest.

It is known (Barndorff-Nielsen [3]) that the statistic t in Examples 4.1, 4.3 and 4.4 is M-ancillary with respect to θ. It is easy to show that the statistic t in Example 4.2 and the sample mean from the gamma distribution with shape parameter θ under the restriction $n\theta>1$, are M-ancillary with respect to θ. The statistic $t=x(1)$ from the exponential distribution with scale parameter θ and unknown support is also M-ancillary with respect to θ. On the other hand, the statistic t in Example 5.1 is not M-ancillary with respect to θ. We find that both the notions of obstructiveness and M-ancillarity are incidentally quite close to each other in practical examples. The question arises of which notion is more reasonable. We agree that the definition of M-ancillarity is intuitively appealing. The definition, however, is inconvenient for analytical calculations. In fact, there seems to be no analytical result for the conditional m.l.e. such as Propositions 3 and 4, even when a statistic t is M-ancillary. A serious defect of M-ancillarity is that neither B- nor S-ancillarity necessarily implies M-ancillarity. This seems to come from the lack of invariance property of M-ancillarity corresponding to Proposition 1. These results lead us to prefer the notion of obstructiveness compared with that of M-ancillarity.

Another notion to be compared with ours is one of the definitions of ancillarity under the presence of a nuisance parameter by Godambe [8] and [9]. Under certain regularity conditions a statistic is ancillary in this sense, if the factorization (2.1) holds and the class of distributions of t to μ is complete for each fixed θ. This notion of ancillarity seems considerably different from other notions (Kuboki [13]). Godambe [9] showed that the statistic t in Examples 4.3 and 4.4 and the sample mean from the gamma distribution are ancillary in this sense with respect to θ. It is easy to check that the statistic t in Examples 4.1 and 4.2 and $t=x_{(1)}$ from the exponential distribution with scale parameter θ and unknown support are ancillary in this sense with respect to θ. In addition, we can show the statistic t in Example 5.1 is ancillary in this sense with respect to θ. We find that both the notions of obstructiveness and ancillarity in this sense are fairly close to each other in practical examples.

An advantage of this notion is that it possesses an optimum property of the conditional m.l.e.; $LC'(\boldsymbol{x},\theta)$ attains the inverse of the lower

bound for the variance of an unbiased estimating function. However, the restriction on an unbiased estimating function is too strong for our purpose. In fact, any unconditional likelihood function for θ in (2.3) appeared in the above examples is not an unbiased estimating function. Consequently, the ancillarity of a statistic in this sense does not necessarily suggest that the conditional m.l.e. is superior to the unconditional m.l.e. It appears that the theory of ancillarity presumes the presence of a nuisance parameter in a strict sense, though a so-called "nuisance" parameter is not necessarily nuisance in practice.

Acknowledgements

The author wishes to thank the two anonymous referees for their constructive suggestions and pointing out errors in the earlier version of the present paper.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Aitkin, M. (1982). Direct likelihood inference, *GLIM 82* (ed. R. Gilchrist), Springer, New York, 76-86.

[2] Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators, *J. R. Statist. Soc.*, **B32**, 283-301.

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

[4] Barndorff-Nielsen, O. (1980). Conditionality revaluations, *Biometrika*, **67**, 293-310.

[5] Fisher, R. A. (1935). The logic of inductive inference, *J. R. Statist. Soc.*, **98**, 39-54.

[6] Fleiss, J. L. (1981). *Statistical Method for Rates and Proportions*, 2nd ed., Wiley, New York.

[7] Gart, J. J. (1971). The comparison of proportions: A view of significant tests, confidence intervals and adjustments for stratification, *Rev. Int. Statist. Inst.*, **39**, 148-169.

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

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

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

[11] Kalbfleish, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters, *J. R. Statist. Soc.*, **B32**, 175-208.

[12] Kalbfleish, J. D. and Sprott, D. A. (1973). Marginal and conditional likelihoods, *Sankhyā*, **A35**, 311-328.

[13] Kuboki, H. (1987). Analysis of marginal and conditional density functions for separate inference, *Ann. Inst. Statist. Math.*, **39**, 1-23.

[14] Liang, K. Y. (1983). On information and ancillarity in the presence of a nuisance parameter, *Biometrika*, **70**, 607-612.

[15] Lindsay, B. G. (1980). Nuisance parameters, mixtures models, and the efficiency of partial likelihood estimators, *Phil. Trans. Royal Soc.*, London, **296**, 639-662.

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

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

[18] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Wiley, New York.

[19] Yanagimoto, T. and Kamakura, T. (1984). The maximum full and partial likelihood estimators in the proportional hazard model, *Ann. Inst. Statist. Math.*, **36**, 363-373.

[20] Yanagimoto, T. and Yamamoto, E. (1985). Simple linear approximations to the likelihood equation for combining evidence in multiple 2×2 tables: A critique of conventional procedures, *Ann. Inst. Statist. Math.*, **37**, 77-89.