Ann. Inst. Statist. Math.
Vol. 45, No. 1, 55–67 (1993)

ESTIMATING FUNCTIONS FOR CONDITIONAL INFERENCE: MANY NUISANCE PARAMETER CASE

H. J. MANTEL[1] AND V. P. GODAMBE[2]

[1]*Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6*
[2]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1*

(Received January 5, 1990; revised June 10, 1991)

Abstract. When there is a complete sufficient statistic for the nuisance parameter which depends on the parameter of interest then there are locally optimal unbiased estimating functions, but generally there is no globally optimal estimating function. We consider conditioning on the minimal sufficient statistic for the nuisance parameter and find the conditional linear optimal unbiased estimating function. Since the nuisance parameter is totally eliminated in the conditional model there is no intrinsic problem in setting up conditional tests of significance and confidence intervals. A compromise between conditional and unconditional optimum estimating functions is suggested. The techniques are illustrated on three examples including the well known common means problem. The proposed hypothesis testing and confidence interval procedures work reasonably well for the examples considered.

Key words and phrases: Common means problem, conditional inference, confidence interval, estimating function, hypothesis testing, Fisher information, stratified model.

1. Introduction

A long standing and important problem in statistics is the estimation of an interesting parameter when there are nuisance parameters present, especially when the number of nuisance parameters is large. A classic reference is Neyman and Scott (1948).

The general approach taken in this paper is that of estimating functions. Others approaching the problem with these tools are Godambe (1976, 1980, 1984), Lindsay (1982) and Kumon and Amari (1984).

Godambe (1984) introduced an extended concept of Fisher information in the presence of nuisance parameters. This definition was motivated as an upper bound on the information (1.2) of any regular unbiased estimating function $g \in \boldsymbol{G}$ (1.1). In Section 2 we prove for a wide class of problems the local optimality of an

estimating function for which the upper bound on the information is attained at a fixed value of the nuisance parameter.

In Section 3 we introduce the stratified model which is the primary focus of this paper. In Section 4 conditioning is used to eliminate the nuisance parameter from the model. Two conditionally unbiased estimating functions are proposed, both motivated to some extent by the results of Section 2. The principal advantage of conditioning here is that, since the conditional model does not involve a nuisance parameter, exact significance tests for the parameter of interest are available.

We now introduce some basic definitions and concepts.

Let $\boldsymbol{X}$ be a sample space with a fixed measure μ. We consider a model consisting of a family of probability densities $p(x;\theta)$ on $(\boldsymbol{X},\mu)$ where $\theta=(\theta_1,\theta_2)$, $\theta_1\in\Omega_1$, $\theta_2\in\Omega_2$ and $\Omega=\Omega_1\times\Omega_2$ is the parameter space. θ_1 is the parameter of interest and Ω_1 is supposed to be a real interval.

A real function g on $\boldsymbol{X}\times\Omega_1$ is a regular unbiased estimating function if $E_\theta(g)=0$ and $E_\theta(g^2)<\infty$, $\theta\in\Omega$, and g satisfies appropriate regularity conditions (Godambe and Thompson (1974)). Let

$$\boldsymbol{G}=\{g: g \text{ is a regular unbiased estimating function}\}. \tag{1.1}$$

The information of a $g\in\boldsymbol{G}$, given by

$$I(g;\theta)=E_\theta^2(\partial g/\partial\theta_1)/E_\theta(g^2),\qquad \theta\in\Omega, \tag{1.2}$$

is a measure of how well g may be used to estimate θ_1. Note that it is a function of θ.

DEFINITION 1.1. An estimating function $g^*\in\boldsymbol{G}$ is said to be locally optimal at θ_{20} if

$$I(g^*;\theta)\geq I(g;\theta),\qquad g\in\boldsymbol{G},\quad \theta_1\in\Omega_1,\quad \theta_2=\theta_{20}. \tag{1.3}$$

DEFINITION 1.2. An estimating function $g^*\in\boldsymbol{G}$ is said to be optimal in $\boldsymbol{G}$ if the inequality (1.3) holds for all $\theta_{20}\in\Omega_2$.

This optimality criterion is often interpreted as minimizing an asymptotic variance, but the criterion can also be applied to finite samples. We emphasize that asymptotics do not play a central role in our development here.

2. Fisher information and optimal estimation

The concept of Fisher information $I(p;\theta_1)$ in the distribution $p(x;\theta)$ about θ_1 ignoring θ_2, referred to in Section 1 is as follows. Let

$$\boldsymbol{U}=\{u(x;\theta): E_\theta(ug)=0,\ E_\theta(u^2)<\infty,\ \theta\in\Omega,\ g\in\boldsymbol{G}\}. \tag{2.1}$$

Then we define

$$I(p;\theta_1)=\inf_{u\in\boldsymbol{U}} E_\theta\{(\partial\log p/\partial\theta_1)-u\}^2. \tag{2.2}$$

Godambe (1984) showed that $I(p;\theta_1)$ is an upper bound on the information (1.2) of any $g \in \boldsymbol{G}$.

For fixed θ the space of functions of x with finite second moment is an inner product space with the inner product of two functions being the expectation of their product. Let $\boldsymbol{V}$ be the orthogonal complement of $\boldsymbol{U}$ in that space.

DEFINITION 2.1. Let u^* be the projection of $(\partial \log p/\partial\theta_1)$ into $\boldsymbol{U}$ and $v^* = (\partial \log p/\partial\theta_1) - u^*$ be the projection of $(\partial \log p/\partial\theta_1)$ into $\boldsymbol{V}$. We emphasize that u^* and v^* may depend on θ_2, as well as on θ_1 and x.

Now for any $u \in \boldsymbol{U}$ we have

$$\begin{aligned} E_\theta\{(\partial \log p/\partial\theta_1) - u\}^2 &= E_\theta\{v^* + u^* - u\}^2 \\ &= E_\theta(v^*)^2 + E_\theta(u^* - u)^2 \geq E_\theta(v^*)^2 \end{aligned}$$

so that the infimum in (2.2) is given by

$$E_\theta(v^*)^2 = E_\theta\{(\partial \log p/\partial\theta_1) - u^*\}^2 = I(p;\theta_1). \tag{2.3}$$

2.1 *Properties of u^* and v^**

Since $u \equiv 1 \in \boldsymbol{U}$ and $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal we have

$$E_\theta(v^*) = 0. \tag{2.4}$$

Differentiating both sides of (2.4) with respect to θ_1, and assuming that differentiation and integration can be interchanged, we obtain

$$E_\theta\{\partial v^*/\partial\theta_1\} = -E_\theta\{v^*(\partial \log p/\partial\theta_1)\} = -E_\theta(v^*)^2$$

because of the orthogonality of $\boldsymbol{U}$ and $\boldsymbol{V}$.

For any $g \in \boldsymbol{G}$, differentiating $E_\theta(g) = 0$ with respect to θ_1 we get

$$E_\theta\{\partial g/\partial\theta_1\} = -E_\theta\{g(\partial \log p/\partial\theta_1)\} = -E_\theta\{gv^*\}$$

whence (1.2) may be rewritten as

$$I(g;\theta) = E_\theta^2\{gv^*\}/E_\theta\{g^2\}, \qquad \theta \in \Omega.$$

That is, $I(g;\theta)$ is proportional to the square of the correlation of g with v^*. Now if $g_0 \in \boldsymbol{G}$ attains the upper bound $I(p;\theta_1)$ on $I(g;\theta)$ then from (2.3) it follows that

$$E_\theta^2\{g_0v^*\}/E_\theta\{g_0^2\} = E_\theta(v^*)^2$$

whence $g_0 = k(\theta)v^*$ for some real function $k(\theta)$. If there is no function $k(\theta)$ such that $k(\theta)v^* \in \boldsymbol{G}$ then no $g \in \boldsymbol{G}$ attains the upper bound $I(p;\theta_1)$ on the information; however, this does not preclude the existence of an optimal unbiased

estimating function. Note that if $v^*(\theta_{20}) \in \boldsymbol{G}$ then $v^*(\theta_{20})$ is locally optimal (Definition 1.1) at $\theta_2 = \theta_{20}$.

Suppose we were interested not in θ_1, but in a smooth one-to-one transformation, $\tau = \tau(\theta_1)$. Since $(\partial \log p/\partial\tau) = (\partial \log p/\partial\theta_1)(\partial\theta_1/\partial\tau)$ it is easily seen from (2.2) that

$$I(p;\tau) = I(p;\theta_1)(\partial\theta_1/\partial\tau)^2$$

and in Definition 2.1, u^* and v^* would both be multiplied by $(\partial\theta_1/\partial\tau)$.

Suppose that the parameterization for θ_2 were changed to $\xi = \xi(\theta_1, \theta_2)$. Let $p_2(x;\theta_1;\xi) = p(x;\theta_1,\theta_2(\theta_1,\xi))$. Letting $g \in \boldsymbol{G}$ be arbitrary and differentiating $\int g(p-p_2)d\mu$ with respect to θ_1 we find that $\{(\partial \log p/\partial\theta_1) - (\partial \log p_2/\partial\theta_1)\} \in \boldsymbol{U}$. Now since

$$(\partial \log p_2/\partial\theta_1) = (\partial \log p/\partial\theta_1) - \{(\partial \log p/\partial\theta_1) - (\partial \log p_2/\partial\theta_1)\}$$

it is clear from (2.2) and Definition 2.1 that the information (2.2) and v^* do not depend on the parameterization of the nuisance parameter, a result to be expected.

2.2 *A special case*

The following theorem is important since it may be used to find v^* in a wide class of problems.

Let $S(\theta_1)$ be the minimal sufficient statistic for θ_2 when θ_1 is given and let

$$w(x;\theta_1,\theta_2) = (\partial \log p/\partial\theta_1) - E_\theta\{(\partial \log p/\partial\theta_1) \mid S(\theta_1)\} \tag{2.5}$$

where $\theta = (\theta_1, \theta_2)$.

THEOREM 2.1. *If* $E_\theta\{(\partial \log p/\partial\theta_1) \mid S(\theta_1)\} \in \boldsymbol{U}$ *in* (2.1) *then in Definition* 2.1, $v^* = w$ *from* (2.5).

PROOF. Since $S(\theta_1)$ is sufficient for θ_2 we have that $w(x;\theta_1,\theta_{20}) \in \boldsymbol{G}$ and since $\boldsymbol{G}$ is orthogonal to $\boldsymbol{U}$ it follows that $w(x;\theta_1,\theta_2) \in \boldsymbol{V}$. Now if $E_\theta\{(\partial \log p/\partial\theta_1) \mid S(\theta_1)\} \in \boldsymbol{U}$ then w in (2.5) is the projection of $(\partial \log p/\partial\theta_1)$ into $\boldsymbol{V}$ and the result follows immediately from Definition 2.1.

When Theorem 2.1 applies, $w(x;\theta_1,\theta_{20})$ is the locally optimum estimating function at $\theta_2 = \theta_{20}$ according to Definition 1.1. For the case when $S(\theta_1)$ is complete and independent of θ_1, the optimality (Definition 1.2) of w was established by Godambe (1976).

An important case where the conditions of the theorem are met is when the distributions of $S(\theta_1)$ for fixed θ_1, $\theta_2 \in \Omega_2$, are complete. Then for all $g \in \boldsymbol{G}$, $E_\theta\{g \mid S(\theta_1)\} = 0$ whence $E_\theta\{gE_\theta\{(\partial \log p/\partial\theta_1) \mid S(\theta_1)\}\} = 0$. Lindsay (1982) established the local optimality (Definition 1.1) of w in (2.5) for this special case.

3. The stratified model

We now restrict to a stratified model. The term stratified here means simply that the nuisance parameter may vary from stratum to stratum. Let $x = (x_1, x_2, \ldots, x_m)$ where the x_i are independent, x_i having density $p_i(x_i; \theta_1, \theta_{2i})$, $\theta_{2i} \in \Omega_{2i}$ and $\Omega_2 = \Omega_{21} \times \Omega_{22} \times \cdots \times \Omega_{2m}$.

For fixed θ_1 let $S_i(\theta_1)$ be the minimal sufficient statistic for θ_{2i} in the model $p_i(x_i; \theta_1, \theta_{2i})$. Following (2.5) we define

$$w_i = (\partial \log p_i/\partial\theta_1) - E_\theta\{(\partial \log p_i/\partial\theta_1) \mid S_i(\theta_1)\} \tag{3.1}$$

and let

$$w = \sum_i w_i. \tag{3.2}$$

Note that the sufficient statistic

$$S(\theta_1) = (S_1(\theta_1), \ldots, S_m(\theta_1)) \tag{3.3}$$

is complete if and only if $S_i(\theta_1)$ is complete for each $i = 1, \ldots, m$ (Lehmann and Sheffé (1955)). In the case that $S(\theta_1)$ is complete w given in (2.5) is equal to w given in (3.2) and both are equal to v^* of Definition 2.1. We assume now that $S(\theta_1)$ is complete. In many practical examples this would be the case.

Now in (3.2) $w(\theta_{20})$ is locally optimal (Definition 1.1) at $\theta_2 = \theta_{20}$. If w_i depends on θ_{2i} then no optimal estimating function (Definition 1.2) exists. One possibility is to try to estimate w at the true value of θ_2 by estimating θ_2. Let $\tilde{\theta}_2 = \tilde{\theta}_2(\theta_1, S(\theta_1))$. As noted by Lindsay (1982), by the sufficiency of $S(\theta_1)$, $E_\theta\{(\partial \log p(x; \theta_1, \tilde{\theta}_2)/\partial\theta_1) \mid S(\theta_1)\}$ does not depend on the true value of θ_2 and hence $\tilde{w} = w(\tilde{\theta}_2)$ is unbiased. For example, $\tilde{\theta}_2$ could be taken to be the maximum likelihood estimate $\hat{\theta}_2(\theta_1)$. Our approach in this paper is different. Rather than replacing θ_2 in w by an estimate we define conditional optimality and look for a conditionally optimal estimating function in a "linear" class of functions. The details are given in Section 4.

Other people have also considered this stratified model or variations of it in the literature. An important early reference is Neyman and Scott (1948). Lindsay (1982) considers restricting to a class of functions which he calls information unbiased and then finding an optimum estimating function in that restricted class. Kumon and Amari (1984) consider restricting to a class of functions which they call information uniform. In the examples of Section 5 we will compare the results of our approach, to be developed in Section 4, to those obtained in these papers.

4. Conditional inference in the stratified model

4.1 *The conditional model*

We want to consider estimation for θ_1 based only on the conditional distribution of x given the minimal sufficient statistic $S(\theta_1)$ in (3.3). To emphasize that

the inferences are based only on the conditional distribution we imagine that if θ_1 is the true value and $S(\theta_1) \neq S(\theta_{10})$ then we do not need to know the distribution of the data conditional on $S(\theta_{10})$. Because we make no assumptions about the distribution of $S(\theta_1)$ we might expect the procedures we obtain to enjoy a type of robustness; that is, the validity of the inferences will depend only on the validity of the conditional part of the model. In this conditional model there is no nuisance parameter, but the sample space may depend on θ_1.

Note that for the construction of the locally optimum estimating function w in (2.5) we need a more detailed model than just the conditional model formulated above.

4.2 *Conditional estimating functions*

We now restrict to estimating functions that are conditionally unbiased; that is, if θ_1 is the true value then $E\{g(x;\theta_1) \mid S(\theta_1)\} = 0$. The conditional information of a function in this class is defined, analogous to (1.2), as

$$I(g,\theta_1 \mid S(\theta_1)) = E^2\{(\partial g/\partial\theta_1) \mid S(\theta_1)\}/E\{g^2 \mid S(\theta_1)\}.$$

A function g^* in a class of functions is called conditionally optimal if it maximizes $I(g,\theta_1 \mid S(\theta_1))$ for all θ_1 and $S(\theta_1)$.

We now restrict to a certain linear class of functions. This restriction will soon be motivated and explained. Suppose the functions $h_i(x_i,\theta_1)$ are such that

$$E\{h_i \mid S_i(\theta_1)\} = 0, \qquad i = 1,2,\ldots,m. \tag{4.1}$$

We consider finding the optimum in the class $\sum_i a_i h_i$ where the a_i are allowed to be functions of $S_i(\theta_1)$ and θ_1. By allowing the a_i to depend on θ_1 we mean that if two values θ_{10} and θ_{11} lead to the same conditional sample space then it may be that $a_i(\theta_{10}, S_i(\theta_{10})) \neq a_i(\theta_{11}, S_i(\theta_{11}))$. However, if θ_{10} and θ_{11} lead to different conditional sample spaces then $a_i(\theta_{10}, S_i(\theta_{11}))$ would not be defined. Now a slight modification of Theorem 1 of Godambe (1985) gives us that the conditional optimum in this class is given by

$$a_i^* = E\{(\partial h_i/\partial\theta_1) \mid S_i(\theta_1)\}/E\{h_i^2 \mid S_i(\theta_1)\}, \qquad i = 1,\ldots,m. \tag{4.2}$$

It should be noted that by $(\partial a_i/\partial\theta_1)$ we mean $\partial a_i(\theta_1, S_i(\theta_{10}))/\partial\theta_1$ evaluated at $\theta_{10} = \theta_1$. Now for a given set of functions h_i, we let

$$g_1^* = \sum_i a_i^* h_i. \tag{4.3}$$

There still remains the problem of choosing the functions h_i in the first place. In the examples we will consider there is a natural choice. For these examples, w of (3.2) is of the form $\sum_i b_i(\theta_1,\theta_{2i})v_i(x_i,\theta_1)$ with $E\{v_i \mid S_i(\theta_1)\} = 0$. It is then natural to take $h_i = v_i$, since v_i is the globally optimal estimating function for θ_1 based on only x_i.

4.3 *A compromise*

The original model was not conditional. Conditioning was motivated by a desire to eliminate the nuisance parameters. Let

$$g(c) = \sum_i c_i a_i^* h_i. \tag{4.4}$$

As a compromise between conditional and unconditional optimality we want to choose $c = (c_1, \ldots, c_m)$ in (4.4), if possible, to minimize

$$E\{g(c) - w\}^2 \tag{4.5}$$

where w is as in (3.2). In minimizing (4.5) we do not allow c_i to depend on the data for the following two reasons: (I) Any dependence of c_i on $S_i(\theta_1)$ would tend to negate the effect of a_i^* on g and hence render conditional optimality ineffective. (II) Any dependence of c_i on the other part of the data would tend to negate our initial choice of the functions h_i. Hence the c_i are allowed to be functions of θ_1 only. Assuming (4.5) is minimized for $c = c^*$ we denote

$$g_2^* = g(c^*). \tag{4.6}$$

In some cases the above compromise may not be available. An alternative is to use, in place of w in (4.5), the unconditional optimum in the class $\sum_i b_i(\theta) h_i$ which is given by $b_i = b_i^* = E\{\partial h_i/\partial\theta_1\}/E\{h_i^2\}$. In the examples we will consider this linear unconditional optimum is in fact equal to w.

4.4 *Hypothesis testing*

Significance tests for particular values of θ_1 may be obtained by examining the distribution of g_1^* in (4.3) conditional on $S(\theta_1)$. If m is large this distribution may be difficult to obtain. Since g_1^* is a sum of independent components a normal approximation may be reasonable. We have

$$g_{1s}^* = g_1^*/[E\{(g_1^*)^2 \mid S(\theta_1)\}]^{1/2} \tag{4.7}$$

is approximately $N(0,1)$ conditional on $S(\theta_1)$. Similarly, from (4.6) we have

$$g_{2s}^* = g_2^*/[E\{(g_2^*)^2 \mid S(\theta_1)\}]^{1/2} \tag{4.8}$$

is approximately $N(0,1)$ conditional on $S(\theta_1)$.

The approximate pivotals in (4.7) and (4.8) may also be used to construct approximate significance intervals for θ_1 which would also be approximate confidence intervals. A question of interest is whether tests based on g_1^* or g_2^* are preferable in the sense of being more powerful against false alternatives or in the sense of having actual significance levels nearer to nominal significance levels. This question is discussed in more detail in Sections 5 and 6.

The idea of basing confidence intervals for a parameter on the distribution of a function of the data and the parameter is not new. See, for example, Boos (1980).

5. Examples

In this section the procedures discussed in Section 4 will be investigated via some examples.

Example 1. This example is also known as the Neyman-Scott problem. Suppose

$$x_{ij} \sim N(\theta_1, \theta_{2i}) \quad (j = 1, \ldots, n_i, \quad i = 1, \ldots, m)$$

are all mutually independent. Let

$$\bar{x}_i = \sum_j x_{ij}/n_i, \qquad S_i^2 = \sum_j (x_{ij} - \bar{x}_i)^2.$$

Now in (3.3) $S_i(\theta_1) = S_i^2 + n_i(\bar{x}_i - \theta_1)^2$, and is complete. Now w_i in (3.1) is given by $w_i = n_i(\bar{x}_i - \theta_1)/\theta_{2i}$ and we will take $h_i = \bar{x}_i - \theta_1$. Now a_i^* in (4.2) is given by $a_i^* = n_i^2/S_i(\theta_1)$ and in (4.3) we have

$$g_1^* = \sum_i n_i^2(\bar{x}_i - \theta_1)/S_i(\theta_1), \tag{5.1}$$

that is, g_1^* is equal to w with θ_{2i} replaced by its maximum likelihood estimate $S_i(\theta_1)/n_i$.

In (4.6) we obtain

$$g_2^* = \sum_i (n_i - 2)n_i(\bar{x}_i - \theta_1)/S_i(\theta_1)$$

where the sum is over all strata for which $n_i \geq 3$. If n_k is 1 or 2 then (4.5) is infinite if $c_k \neq 0$. The estimating function g_2^* was also obtained by Lindsay (1982), Kumon and Amari (1984) and Neyman and Scott (1948).

The actual distributions of g_1^* and g_2^* conditional on $S(\theta_1)$ are quite complicated for this example. In (4.7) we obtain

$$g_{1s}^* = \left[\sum_i n_i^2(\bar{x}_i - \theta_1)/S_i(\theta_1)\right] \Big/ \left[\sum_i n_i^2/S_i(\theta_1)\right]^{1/2} \tag{5.2}$$

and in (4.8)

$$g_{2s}^* = \left[\sum_i (n_i - 2)n_i(\bar{x}_i - \theta_1)/S_i(\theta_1)\right] \Big/ \left[\sum_i (n_i - 2)^2/S_i(\theta_1)\right]^{1/2}. \tag{5.3}$$

It should be noted that when $m = 1$ g_{1s}^* in (5.2) is equal to g_{2s}^* in (5.3) and the significance test based on their exact conditional distribution is equivalent to the usual t-test.

A small simulation study was conducted to estimate the probabilities of rejecting various values of θ_1 at the .05 and .10 significance levels based on a standard

normal approximation to the distributions of (5.2) and (5.3) for various values of m, n_i and θ_2. The agreement between the actual and the nominal significance levels seems quite good, especially at the nominal .10 level, except in some examples where the total number of observations, $\sum_i n_i$, is small, say less than 30. It does not seem possible to draw any general conclusions regarding the superiority of (5.2) or (5.3), but note that as the n_i increase the difference between (5.2) and (5.3) decreases.

Example 2. Suppose x_i are gamma variates with shape parameter α_i and rate parameter $\theta_1\theta_{2i}$ while y_i are gamma with shape β_i and rate θ_{2i} $(i = 1, \ldots, m)$. Here α_i and β_i are assumed to be known. The situation here is what we would have after a reduction by sufficiency if we had samples of α_i and β_i observations from exponential distributions with rates $\theta_1\theta_{2i}$ and θ_{2i} respectively. Now $S_i(\theta_1) = y_i + \theta_1 x_i$ is complete sufficient for θ_{2i} given θ_1 and w_i in (3.1) is given by

$$w_i = (\alpha_i y_i - \theta_1 \beta_i x_i)\theta_{2i}/\theta_1(\alpha_i + \beta_i).$$

Taking $h_i = \alpha_i y_i - \theta_1 \beta_i x_i$ we find in (4.3)

$$g_1^* = \sum_i (\alpha_i + \beta_i + 1)(\alpha_i y_i - \theta_1 \beta_i x_i)/\{(\alpha_i + \beta_i)\theta_1 S_i(\theta_1)\}$$

and in (4.6)

$$g_2^* = \sum_i (\alpha_i y_i - \theta_1 \beta_i x_i)/\{\theta_1 S_i(\theta_1)\}.$$

Note that g_2^* is equal to w with θ_{2i} replaced by its maximum likelihood estimate $(\alpha_i + \beta_i)/S_i(\theta_1)$.

Lindsay (1982) also obtains the estimating function g_2^*, as do Kumon and Amari (1984) for the special case $\alpha_i = \beta_i$.

Note that $y_i/S_i(\theta_1) = (h_i + \beta_i S_i(\theta_1))/\{(\alpha_i + \beta_i)S_i(\theta_1)\}$ has a standard beta distribution with parameters (β_i, α_i) conditional on $S_i(\theta_1)$. For approximate significance tests we obtain in (4.7)

$$g_{1s}^* = \left[\sum_i (\alpha_i + \beta_i + 1)h_i/\{(\alpha_i + \beta_i)\theta_1 S_i(\theta_1)\}\right] \Big/ \left[\sum_i \alpha_i\beta_i(\alpha_i + \beta_i + 1)/\{\theta_1^2(\alpha_i + \beta_i)^2\}\right]^{1/2} \tag{5.4}$$

and in (4.8)

$$g_{2s}^* = \left[\sum_i h_i/\{\theta_1 S_i(\theta_1)\}\right] \Big/ \left[\sum_i \alpha_i\beta_i/\{\theta_1^2(\alpha_i + \beta_i + 1)\}\right]^{1/2}. \tag{5.5}$$

As in Example 1, a small simulation study was conducted. In this example the actual and nominal significance levels agreed very well, even when the total

number of observations, $\sum_i(\alpha_i + \beta_i)$, was as small as 16. In none of the cases considered were the differences in the performances of (5.4) and (5.5) large enough to be of practical importance and they did not definitively favour one estimating function over the other.

Example 3. Suppose

$$x_{ij} \sim N(\theta_{2i}, 1) \quad (j = 1, \ldots, m_i)$$
$$y_{ij} \sim N(\theta_1\theta_{2i}, 1) \quad (j = 1, \ldots, n_i, \ i = 1, \ldots, m)$$

are all mutually independent. Let $x_i = \sum_j x_{ij}$, $y_i = \sum_j y_{ij}$. Then $\{x_i, y_i, i = 1, \ldots, m\}$ is minimal sufficient for $(\theta_1, \theta_{21}, \ldots, \theta_{2m})$ and $S_i(\theta_1) = x_i + \theta_1 y_i$ is complete sufficient for θ_{2i} given θ_1. Now w_i in (3.1) is given by

$$w_i = (m_i y_i - \theta_1 n_i x_i)\theta_{2i}/(m_i + n_i\theta_1^2)$$

and we will take $h_i = m_i y_i - \theta_1 n_i x_i$. Now g_1^* in (4.7) is just w with θ_{2i} replaced by its maximum likelihood estimate, $S_i(\theta_1)/(m_i + n_i\theta_1^2)$.

$$g_1^* = \sum_i (m_i y_i - \theta_1 n_i x_i) S_i(\theta_1)/(m_i + n_i\theta_1^2)^2.$$

For this example g_2^* in (4.6) does not exist since c minimizing (4.5) depends on θ_2. Note that, conditionally on h_i, g_1^* has expectation equal to w.

Kumon and Amari (1984) obtain the estimating function g_1^* for the special case $m_i = n_i$.

Conditionally on $S_i(\theta_1)$, h_i is normal with mean 0 and variance $m_i n_i(m_i + n_i\theta_1^2)$. From this the distribution of g_1^* conditional on $S(\theta_1)$ is easily calculated and exact conditional significance tests and confidence intervals are relatively easy to obtain.

6. Conditional vs unconditional optimality

Unlike in the examples of Section 5, when the complete sufficient statistic $S(\theta_1)$ is independent of θ_1, $S(\theta_1) = S$, the conditional score provides globally optimum estimating function both conditionally and unconditionally (Godambe (1976)). Otherwise there is a basic conflict. In Section 5, for Example 1, the estimating function $w = \sum w_i$ which depends on θ_2 is unconditionally optimum locally at θ_2. "w" cannot be conditionally optimum for the conditional distribution is independent of θ_2. The estimating function g_2^* is also conditionally inferior to g_1^*. For large samples the conditional variance of the estimate $\hat{\theta}_{11}$ where $[g_1^*(\theta_1) = 0] \Rightarrow [\hat{\theta}_{11} = \theta_1]$, would generally be smaller than that of $\hat{\theta}_{12}$ where $[g_2^*(\theta_1) = 0] \Rightarrow [\hat{\theta}_{12} = \theta_1]$. But the situation is just the opposite unconditionally as implied by Neyman and Scott (1948). These results imply that the asymptotic unconditional variance of the estimate $\hat{\theta}_{12}$ is smaller than that of $\hat{\theta}_{11}$. That conditionally the situation is otherwise, follows from the conditional superiority of the estimating

function g_1^* to g_2^*. This apparent paradox can be attributed to the fact that the conditioning statistic depends on θ_1.

The derivation of the estimating function g_2^* depends in an important way on the distribution of the sufficient statistic $S(\theta_1)$. This distribution, on the other hand plays no role in the derivation of the estimating function g_1^*. For instance, if in Example 1 the χ^2 distribution of $S(\theta_1)$ is replaced by some other distribution (preserving completeness) the estimating function g_2^* could be affected but not g_1^*.

Godambe (1991) provided two large sample approximations for optimal estimating functions in presence of nuisance parameters. These approximations, for a semi-parametric version of Example 1, coincide with the estimating functions g_1^* and g_2^*. It is evident (see equations (14), (38), (39) of Godambe (1991)) that the derivation of g_2^* depends more on the entire underlying distribution than that of g_1^*. This is also clear from Lindsay (1982) and Kumon and Amari (1984).

The conditionality considerations also play a very important part in the choice of the functions h_i in (4.1). As said before, the functions h_i are modifications of the unconditionally locally optimum estimating functions w_i in (3.2); the latter being dependent on the nuisance parameter θ_{2i}. The modifications are based on the following considerations. (1) The functions h_i should be independent of the nuisance parameter θ_2 and should be both conditionally and unconditionally unbiased. (2) The conditionally optimum combination of the functions h_i, namely g_1^* in (4.3) should be approximately unconditionally optimum for large samples.

7. Discussion

We have seen that for significance tests based on (4.7) and (4.8) it is not possible to generally recommend either of g_{1s}^* or g_{2s}^* over the other. However, g_{1s}^* does have the advantage of being more generally available.

For the Neyman-Scott problem, g_{2s}^* ignores strata which have fewer than 3 observations where g_{1s}^* can use the information from strata for which the subsample size is 2. There is a serious problem with g_1^* in (5.1) if any of the n_i's are equal to 1. If $n_k = 1$ then the contribution to g_1^* from stratum k is $(x_{k1} - \theta_1)^{-1}$ which approaches infinity as θ_1 approaches x_{k1}. Therefore g_1^* should not be used in that case. Strata for which $n_i = 1$ could be ignored, but that is not suggested by our theory and such strata do contain some information about θ_1. As was mentioned in Section 5, for the Neyman-Scott problem our procedure reduces to the t-test when the number of strata, m, is 1.

We noted in Examples 1 and 3 that g_1^*, and in Example 2 g_2^*, was equal to w with θ_2 replaced by its maximum likelihood estimate. We emphasize that our approach is not to merely replace θ_2 in w by an estimate.

We have made the assumption that the minimal sufficient statistic is complete. This was only to ensure that $w(\theta_{20})$ in (3.2) was the locally optimum unbiased estimating function (Definition 1.1) and that h_i, if proportional to the w_i, would be the globally optimal estimating function based on x_i only. However, none of the development in Section 4 depends on this; all that is really needed are functions h_i satisfying (4.1).

In the examples we have considered the conditional approach has worked very well. The approach may be unsatisfactory when the minimal sufficient statistic for

the nuisance parameter is too fine, that is, when the conditioning is very extreme and the conditional distribution is not of much use for inference. In this case the conditioning would have to be relaxed. Cox and Reid (1987) and Liang (1987) have discussed conditioning on the maximum likelihood estimate of a nuisance parameter which is orthogonal to the parameter of interest. For our purposes we would want to condition on a function $T(x, \theta_1)$ which is such that the conditional distribution of some interesting function $f(x, \theta_1)$ is not too heavily dependent on θ_2. Somewhat indirectly we suggest conditioning on a function $T(x, \theta_1)$, that captures the information in the data about θ_2. A natural starting point is to condition on $\hat{\theta}_2(\theta_1)$. If desirable, we could further condition on other functions $T(x, \theta_1)$.

In relation to the above discussion it should be emphasized that conditioning on the minimal sufficient statistic for the nuisance parameter θ_2, as in Godambe (1980), plays a dual role: (1) It provides a definition of information in the distribution about the interesting parameter θ_1 in presence of θ_2. (2) It defines the optimal estimating function for θ_1.

The conditioning suggested by Cox and Reid (1987), discussed above, depends on orthogonal parameter transformation while, as shown in Subsection 2.1, the definitions (1) and (2), just mentioned, are independent of any transformations. One possible conclusion seems to be that Cox and Reid, by implication, suggest replacing or approximating the present model, where the distribution conditional on the minimal sufficient statistic is degenerate, by a more manageable model. Such approximating replacement models clearly underlay the modern theory of "quasi-likelihood estimation" (Godambe and Heyde (1987)). It is important to note here that the locally optimum estimating function for θ_1, namely w in (2.5), is already orthogonal to $\partial \log p / \partial \theta_2$, that is $E_\theta[w(\partial \log p / \partial \theta_2)] = 0$; assuming of course the parameter θ_2 to be real. Further we have $E_\theta(\partial w / \partial \theta_2) = 0$. Hence we can replace in w, θ_2 by its maximum likelihood estimate without much affecting the optimality for large samples (Godambe (1991)).

We can identify two broad approaches to the elimination of nuisance parameters. First, there is the conditional approach where inference for θ_1 is based on a conditional distribution which does not depend too heavily on the nuisance parameter. Second, there is the method of integration as in Bayesian or empirical Bayesian methods. Conditional methods have the advantage of depending only on the conditional part of the model; integration methods may use all of the data. The question of interest is can we identify the situations where one or the other approach is preferable? Of course, ideally we would like the two approaches to agree.

REFERENCES

Boos, D. D. (1980). A new method for constructing approximate confidence intervals from M estimates, *J. Amer. Statist. Assoc.*, **75**, 142–145.

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

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.

Godambe, V. P. (1984). On ancillarity and Fisher information in the presence of a nuisance parameter, *Biometrika*, **71**, 626–629.

Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes, *Biometrika*, **72**, 419–428.

Godambe, V. P. (1991). Orthogonality of estimating functions and nuisance parameters, *Biometrika*, **78**, 143–151.

Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation, *Internat. Statist. Rev.*, **55**, 231–244.

Godambe, V. P. and Thompson, M. E. (1974). Estimating equations in the presence of a nuisance parameter, *Ann. Statist.*, **2**, 568–571.

Kumon, M. and Amari, S. (1984). Estimation of a structural parameter in the presence of a large number of nuisance parameters, *Biometrika*, **71**, 445–459.

Lehmann, E. L. and Sheffé, H. (1955). Completeness, similar regions, and unbiased estimation—part II, *Sankhyā*, **15**, 219–236.

Liang, K. (1987). Estimating functions and approximate conditional likelihood, *Biometrika*, **74**, 695–702.

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

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