

A paradox concerning nuisance parameters and projected estimating functions

BY MASAYUKI HENMI

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.
m.henmi@warwick.ac.uk

AND SHINTO EGUCHI

*Institute of Statistical Mathematics, The Graduate University of Advanced Studies,
Minami-azabu, Tokyo 106-8569, Japan*
eguchi@ism.ac.jp

SUMMARY

This paper is concerned with a paradox associated with parameter estimation in the presence of nuisance parameters. In a statistical model with unknown nuisance parameters, the efficiency of an estimator of a parameter usually increases when the nuisance parameters are known. However the opposite phenomenon can sometimes occur. In this paper, we elucidate the occurrence of this paradox by examining estimating functions. In particular, we focus on the projected estimating function, which is defined by the projection of the score function on to a given estimating function. A sufficient condition for the paradox to occur is the orthogonality of the two components of the projected estimating functions corresponding to parameters of interest and nuisance parameters. In addition, a numerical assessment is conducted in the context of a simple model to investigate the improvement of the asymptotic efficiency of estimators.

Some key words: Orthogonality of estimating functions; Propensity score; Semiparametric model.

1. INTRODUCTION

We consider a statistical model with a vector of parameters of interest β and a vector of nuisance parameters α . It is well known that in regular models the asymptotic variance matrix of the maximum likelihood estimator of a vector of smooth parameters β with unknown α is larger than or at least equal to that with known α in the sense of matrix inequality. However, it has been noted that the reversal of this relationship can occur when we use an estimation method other than maximum likelihood, as is typically necessary for examples with semiparametric models.

We present one illustrative example of such a paradoxical phenomenon. When estimating the causal effect of an exposure or treatment on an outcome of interest, we must consider a number of confounding factors which may vary with the exposure or treatment and are themselves independent predictors of the outcome. Let y , $x = (1, x_1, \dots, x_p)'$ and s

be, respectively, a continuous outcome variable of interest, a vector of variables of confounding factors, and an indicator of exposure which takes the value 1 when the subject is exposed and is otherwise 0. We assume that all of the confounding factors are observed. Robins et al. (1992) discussed the following model for estimating the causal effect:

$$y = \beta s + h(x) + \varepsilon, \quad E(\varepsilon|s, x) = 0, \quad (1)$$

$$\text{pr}(s = 1|x) = \frac{\exp(\alpha'x)}{1 + \exp(\alpha'x)}, \quad (2)$$

where h is an unknown function and α is an unknown vector of nuisance parameters. The parameter β represents the average causal effect of the exposure or treatment on the outcome. Under this semiparametric model, Robins et al. (1992) proposed the following estimating equation for β , based on a random sample $\{y_i, s_i, x_i\}$:

$$\sum_{i=1}^n \{s_i - e(x_i; \hat{\alpha})\}(y_i - \beta s_i) = 0, \quad (3)$$

where $e(x; \alpha) = \exp(\alpha'x) / \{1 + \exp(\alpha'x)\}$ and $\hat{\alpha}$ is the maximum likelihood estimator of α from the logistic regression model (2). The conditional probability $\text{pr}(s = 1|x)$ in (2) is called a propensity score and was introduced by Rosenbaum & Rubin (1983).

In this setting, a remarkable relationship with the vector of nuisance parameters α was discussed by Robins et al. (1992). Let $\hat{\beta}$ be an estimator of β defined by (3) and let $\tilde{\beta}$ be an estimator defined by (3) with the true value of α in place of $\hat{\alpha}$. Under some regularity conditions, the estimators $\hat{\beta}$ and $\tilde{\beta}$ are both consistent and asymptotically normal, but the asymptotic variances satisfy

$$\text{avar}(\tilde{\beta}) \geq \text{avar}(\hat{\beta}). \quad (4)$$

The inequality (4) implies that a more precise estimator of β can be obtained by estimating the nuisance parameter α than by using the true value of α even if the latter is known. This unusual phenomenon has been pointed out for various practical situations (Robins et al., 1994; Lawless et al., 1999; Zou & Fine, 2002).

The paradoxical relationship (4) is caused by the asymptotic independence of $\hat{\beta}$ and $\hat{\alpha}$, since we observe that

$$\sqrt{n}(\tilde{\beta} - \beta) = \sqrt{n}(\hat{\beta} - \beta) + \sqrt{n}D(\hat{\alpha} - \alpha) + o_p(1)$$

for some matrix D ; see Pierce (1982) for more formal discussion from this viewpoint. However, this is far from representing a clear understanding of the structure of the paradox. In this paper, we clarify the structure of this paradoxical phenomenon by focusing on the estimating functions rather than the estimators. The key is to consider a projected estimating function, which is defined by the orthogonal projection of the score function on to the estimating function. Since the projected and original estimating functions differ by a nonsingular multiplicative constant matrix, the solutions to the projected estimating equations and the original estimating equations are identical. The asymptotic variance matrix of the estimator obtained from an estimating function is equal to the inverse of the variance matrix of the projected estimating function. Therefore, by using this concept, we can compare geometrically the asymptotic variance matrices associated with estimating functions. This will help to clarify the phenomenon by which the paradox occurs.

2. PROJECTED ESTIMATING FUNCTIONS

In this section we introduce the projected estimating function, which is central to our discussion. Let $\mathcal{M} = \{p(x; \theta, \kappa)\}$ be a family of probability distributions, the elements of which are specified by a d -dimensional parameter θ and a nuisance parameter κ . The distribution $p(x; \theta, \kappa)$ is continuously differentiable with respect to θ for almost every x and all κ . We further assume that the parameter θ is composed of two parts, a parameter of interest β and a nuisance parameter α . Let $s(x, \theta, \kappa) = (s_\beta(x, \theta, \kappa)', s_\alpha(x, \theta, \kappa)')'$ be the score function for $\theta = (\beta', \alpha')'$, which is assumed to satisfy $E\{\|s(x, \theta, \kappa)\|^2\} < \infty$ for all θ and κ , where the expectation is taken with respect to the distribution $p(x; \theta, \kappa)$. If the nuisance parameter κ is of infinite dimension, then $\mathcal{M}$ is a semiparametric model. However, note that the following argument is still valid even if the nuisance parameter κ is assumed to be either of finite dimension or known. The latter case is referred to herein as the parametric case. Let $u(x, \theta)$ be an estimating function of d dimensions which is assumed to satisfy the conditions

$$E\{u(x, \theta)\} = 0, \quad E\{\|u(x, \theta)\|^2\} < \infty. \quad (5)$$

Furthermore, we assume that

$$\det \left\{ \frac{\partial}{\partial \theta_1} m(\theta_1, \theta_2, \kappa) \right\}_{\theta_1 = \theta, \theta_2 = \theta} \neq 0 \quad (6)$$

for all θ and κ , where $m(\theta_1, \theta_2, \kappa) = \int u(x, \theta_1) p(x; \theta_2, \kappa) dx$. We do not necessarily assume the differentiability of $u(x, \theta)$ with respect to θ . For example, the estimating function $u(x, \theta) = \text{sgn}(x - \theta)$ is not differentiable with respect to θ but satisfies condition (6), where $\text{sgn}(A)$ denotes the sign of a real number A . We have an estimator $\hat{\theta}$ of θ as the solution of the estimating equation

$$\sum_{i=1}^n u(x_i, \theta) = 0,$$

based on a given dataset $x_1, \dots, x_n$. Under some additional regularity conditions (van der Vaart, 1998, Ch. 5), $\hat{\theta}$ is consistent and asymptotically normally distributed with the asymptotic variance matrix

$$\text{avar}(\hat{\theta}) = W(\theta, \kappa)^{-1} V(\theta, \kappa) \{W(\theta, \kappa)^{-1}\}', \quad (7)$$

where $V(\theta, \kappa) = \text{var}\{u(x, \theta)\} = E\{u(x, \theta)u(x, \theta)'\}$ and $W(\theta, \kappa) = E\{u(x, \theta)s(x, \theta, \kappa)'\}$, both of which are evaluated using the true values of θ and κ . Based on the above setting, we define the projected estimating function $u_*(x, \theta)$ for an estimating function $u(x, \theta)$ by the equation

$$u_*(x, \theta) = W(\theta, \kappa)' V(\theta, \kappa)^{-1} u(x, \theta), \quad (8)$$

which depends on the unknown true value of κ in the semiparametric case. This is sometimes called the standardised estimating function (Heyde, 1997, Ch. 1). This term is, however, also used with a different meaning (Godambe, 1991) and we call $u_*(x, \theta)$ the projected estimating function according to its geometrical meaning. Note that the estimator obtained from $u_*(x, \theta)$ coincides with that obtained from $u(x, \theta)$ since $V(\theta, \kappa)$ and $W(\theta, \kappa)$ are constant matrices. The definition (8) implies that the projected estimating function $u_*(x, \theta)$ is characterised by the equation

$$u_*(x, \theta) = \arg \min_{v=Au} E\{\|s(x, \theta, \kappa) - v(x, \theta)\|^2\},$$

where A ranges over all $d \times d$ matrices, or, equivalently,

$$E[u(x, \theta) \{s(x, \theta, \kappa) - u_*(x, \theta)\}'] = 0.$$

This means that each component of $u_*(x, \theta)$ is the orthogonal projection of each component of the score function $s(x, \theta, \kappa)$ on to the linear subspace spanned by all components of the estimating function $u(x, \theta)$ in the Hilbert space

$$H_{\theta, \kappa} = \{f(x) : E[\{f(x)\}^2] < \infty\}$$

with the inner product $\langle f, g \rangle = E\{f(x)g(x)\}$ for scalar functions $f(x)$ and $g(x)$. Note that $s_*(x, \theta) = s(x, \theta)$ in the parametric case, where $s(x, \theta, \kappa) = s(x, \theta)$ with known κ . Considering the projection of the score function is also natural from a geometrical point of view. Let t be a transformation of the parameter θ to ω . Then we observe that

$$v_*(x, \omega) = \frac{\partial \theta}{\partial \omega} u_*(x, \theta),$$

where $v(x, \omega) = u\{x, t^{-1}(\omega)\}$. This implies that the projected estimating function $u_*(x, \theta)$ is well defined as a covariant vector field (Kobayashi & Nomizu, 1996, Ch. 1). We will discuss an elegant relationship between the projected estimating function and the influence function.

In terms of a projected estimating function, the asymptotic variance matrix of the estimator $\hat{\theta}$ given by (7) is represented as in the following lemma.

LEMMA 1. *The projected estimating function $u_*(x, \theta)$ satisfies*

$$\text{avar}(\hat{\theta}) = [\text{var}\{u_*(x, \theta)\}]^{-1}. \quad (9)$$

The proof directly follows from the expression (7) for the asymptotic variance matrix and the definition (8) of the projected estimating function. This formula implies that the variance matrix of the projected estimating function plays a role that is similar to the Fisher information matrix. The projection procedure extends the formula for score functions to that for general estimating functions.

3. STRUCTURE OF THE PARADOX

In this section, we examine the structure of the paradoxical phenomena mentioned in § 1 using the concept of projected estimating functions. In accordance with the partition of the parameter $\theta = (\beta', \alpha')'$, we assume that an estimating function $u(x, \theta)$ has been given in the form

$$u(x, \theta) = (u_\beta(x, \theta)', u_\alpha(x, \theta)')',$$

where $u_\beta(x, \theta)$ and $u_\alpha(x, \theta)$ are vector-valued functions of the same dimension as β and α respectively, and that $u_\beta(x, \theta)$ also satisfies condition (6), namely

$$\det \left\{ \frac{\partial}{\partial \beta_1} m_\beta(\theta_1, \theta_2, \kappa) \right\}_{\theta_1 = \theta, \theta_2 = \theta} \neq 0,$$

where $m_\beta(\theta_1, \theta_2, \kappa) = \int u_\beta(x, \theta_1) p(x; \theta_2, \kappa) dx$ and $\theta_1 = (\beta_1', \alpha_1')'$. The projected estimating function of $u(x, \theta)$ can also be divided into two components,

$$u_*(x, \theta) = (u_{*\beta}(x, \theta)', u_{*\alpha}(x, \theta)')',$$

where $u_{*\beta}(x, \theta)$ and $u_{*\alpha}(x, \theta)$ are the orthogonal projections of the components of the score functions $s_\beta(x, \theta, \kappa)$ for β and $s_\alpha(x, \theta, \kappa)$ for α , respectively, on to the linear space spanned by all components of the estimating function $u(x, \theta)$. Let $\hat{\theta} = (\hat{\beta}', \hat{\alpha}')'$ be an estimator of $\theta = (\beta', \alpha')'$ obtained from the estimating function $u(x, \theta)$. As mentioned in § 2, this is also an estimator obtained from the projected estimating function $u_*(x, \theta)$. We let $\tilde{\beta}$ be an estimator obtained from the estimating function $u_\beta(x, \beta, \alpha)$ with known α . Thus,

$$\sum_{i=1}^n u(x_i, \hat{\beta}, \hat{\alpha}) = 0, \quad \sum_{i=1}^n u_\beta(x_i, \tilde{\beta}, \alpha) = 0. \quad (10)$$

Note that $\tilde{\beta}$ is a function of the known α and is a nonfeasible estimator when α is unknown.

We now give geometrical expressions for the asymptotic variance matrices of $\hat{\beta}$ and $\tilde{\beta}$, which are key to the understanding of the structure that gives rise to the paradox. In the following discussion, $\Pi\{u(x)|v(x)\}$ denotes the vector comprising the orthogonal projection of the components of the vector-valued function $u(x)$ on to the subspace generated by the components of the vector-valued function $v(x)$ in the Hilbert space $H_{\theta, \kappa}$ defined in § 2. For symmetric matrices A and B , the notation $A \geq B$ means that $A - B$ is positive semidefinite.

LEMMA 2. The asymptotic variance matrix of $\hat{\beta}$ is

$$\text{avar}(\hat{\beta}) = (\text{var}[u_{*\beta}(x, \theta) - \Pi\{u_{*\beta}(x, \theta)|u_{*\alpha}(x, \theta)\}])^{-1}, \quad (11)$$

and therefore

$$\text{avar}(\hat{\beta}) \geq [\text{var}\{u_{*\beta}(x, \theta)\}]^{-1}. \quad (12)$$

Equality holds if and only if

$$E\{u_{*\beta}(x, \theta)u_{*\alpha}(x, \theta)'\} = 0, \quad (13)$$

which is equivalent to the conditions

$$E\{u_{*\beta}(x, \theta)s_\alpha(x, \theta, \kappa)'\} = 0, \quad E\{s_\beta(x, \theta, \kappa)u_{*\alpha}(x, \theta)'\} = 0.$$

This is a direct consequence of Lemma 1. The proof is given in the Appendix. If the estimating function $u(x, \theta)$ coincides with the projected estimating function $u_*(x, \theta)$, then the asymptotic variance matrix of $\hat{\beta}$ is equal to $[\text{var}\{u_{*\beta}(x, \theta)\}]^{-1}$ from Lemma 1. Therefore, from (12) we obtain the inequality

$$\text{avar}(\hat{\beta}) \geq \text{avar}(\tilde{\beta}), \quad (14)$$

which implies that the paradox never occurs in this case. For example, the case in which $u(x, \theta)$ is the score function for θ in the parametric model is typical. In general, we have the following lemma for the asymptotic variance matrix of $\tilde{\beta}$.

LEMMA 3. The asymptotic variance matrix of $\tilde{\beta}$ is

$$\text{avar}(\tilde{\beta}) = (\text{var}[\Pi\{u_{*\beta}(x, \theta)|u_\beta(x, \theta)\}])^{-1}, \quad (15)$$

and therefore

$$\text{avar}(\tilde{\beta}) \geq [\text{var}\{u_{*\beta}(x, \theta)\}]^{-1}. \quad (16)$$

Equality holds if and only if

$$u_{\beta*}(x, \theta) = u_{*\beta}(x, \theta), \quad (17)$$

where $u_{*\beta}(x, \theta)$ is the projected estimating function for $u_\beta(x, \theta)$, or equivalently

$$u_{*\beta}(x, \theta) = Au_\beta(x, \theta)$$

for some nonsingular matrix A .

The proof is obtained immediately through the geometrical argument. By the definition of projected estimating functions and Lemma 1, we find that the asymptotic variance matrix of $\tilde{\beta}$ is

$$\text{avar}(\tilde{\beta}) = [\text{var}\{u_{\beta*}(x, \theta)\}]^{-1}, \quad (18)$$

where $u_{\beta*}(x, \theta)$ is the projected estimating function of $u_\beta(x, \theta)$, that is the orthogonal projection of the components of the score function for β on to the linear space spanned by the components of $u_\beta(x, \theta)$. On the other hand, $u_{*\beta}(x, \theta)$ is the orthogonal projection on to a larger space which includes that linear space. Hence, by elementary Euclidean geometry, we have

$$u_{\beta*}(x, \theta) = \Pi\{u_{*\beta}(x, \theta) | u_\beta(x, \theta)\}; \quad (19)$$

see Fig. 1. Equations (18) and (19) yield equation (15). Equation (15) and the Pythagorean relationship lead to equations (16) and (17).

In general, there is no relationship between the asymptotic variance matrices of $\hat{\beta}$ and $\tilde{\beta}$. The asymptotic efficiency of $\tilde{\beta}$ relative to $\hat{\beta}$ depends on the location of $u_{*\alpha}(x, \theta)$, $u_{*\beta}(x, \theta)$ and $u_{\beta*}(x, \theta)$; see Fig. 1. However, the next theorem follows from the fact that the asymptotic variance matrices have the common lower bound implied by inequalities (12) and (16).

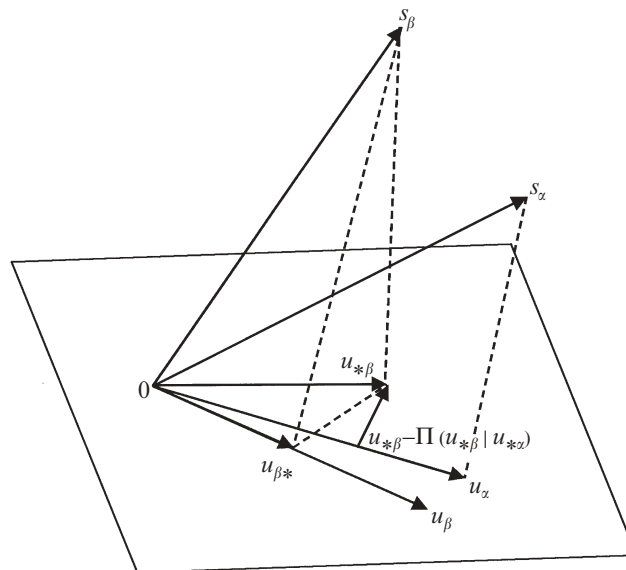


Fig. 1. The relationships between $u_{*\alpha}$, $u_{\beta*}$ and $u_{*\beta}$.

THEOREM 1. Assume that the two components of the projected estimating function $u_*(x, \theta)$ are orthogonal; that is,

$$E\{u_{*\beta}(x, \theta)u_{*\alpha}(x, \theta)'\} = 0. \quad (20)$$

Then the asymptotic variance matrix of $\hat{\beta}$ does not exceed that of $\tilde{\beta}$:

$$\text{avar}(\tilde{\beta}) \geq \text{avar}(\hat{\beta}). \quad (21)$$

Equality holds if and only if

$$E\{u_\beta(x, \theta)s_\alpha(x, \theta, \kappa)'\} = 0. \quad (22)$$

More precisely, the difference between these asymptotic variance matrices is given by

$$\text{avar}(\tilde{\beta}) - \text{avar}(\hat{\beta}) = D \text{avar}(\hat{\alpha})D', \quad (23)$$

where $D = E\{u_\beta(x, \theta)s_\beta(x, \theta, \kappa)'\}^{-1}E\{u_\beta(x, \theta)s_\alpha(x, \theta, \kappa)'\}$.

Equation (20) is condition (13) for equality of (12). Hence, by Lemmas 2 and 3, we obtain (21). Condition (22) is equivalent to condition (17) for equality of (16) under the orthogonality of $u_{*\beta}(x, \theta)$ and $u_{*\alpha}(x, \theta)$. The proof of (23) is given in the Appendix. The estimator $\hat{\beta}$ is more efficient than $\tilde{\beta}$ asymptotically if condition (22) is not satisfied. This is the paradox mentioned in § 1. Note that, if the estimating function $u(x, \theta)$ is the score function for θ in the parametric case, then equality (20) is equivalent to the orthogonality of the score functions for β and α , and equality holds in (21) (Cox & Reid, 1987).

We can give a typical situation satisfying (20). Assume that the score function for α does not depend on κ and that the following conditions hold:

$$E\{s_\beta(x, \theta, \kappa)s_\alpha(x, \theta)'\} = 0, \quad (24)$$

$$u_\alpha(x, \theta) = s_\alpha(x, \theta), \quad (25)$$

where $s_\alpha(x, \theta)$ is the score function for α . Then we see that $u_{*\alpha}(x, \theta) = s_\alpha(x, \theta)$ by condition (25). This fact and condition (24) lead to the orthogonality of $u_{*\beta}(x, \theta)$ and $u_{*\alpha}(x, \theta)$, and this implies (20).

We return to the example of § 1. Note that in this case conditions (24) and (25) are satisfied. In fact, we observe (25) because α is estimated by the maximum likelihood method based on (2). Condition (24) follows from the fact that s_α is independent of β and

$$E\{s_\beta(x, \theta, \kappa)s_\alpha(x, \theta)'\} = E\left[-\left\{\frac{\partial s_\alpha}{\partial \beta}(x, \theta)\right\}'\right] = 0.$$

Hence, the paradoxical relationship (4) can be understood from the results of this section. The difference between the asymptotic variances $\text{avar}(\tilde{\beta})$ and $\text{avar}(\hat{\beta})$ is given by

$$\frac{1}{[E\{r(x; \alpha)\}]^2} E\{h(x)r(x; \alpha)x'\} [E\{r(x; \alpha)xx'\}]^{-1} E\{h(x)r(x; \alpha)x\},$$

where $r(x; \alpha) = e(x; \alpha)\{1 - e(x; \alpha)\}$. This implies that the estimator $\hat{\beta}$ with estimated α has the same asymptotic efficiency as $\tilde{\beta}$ with known α if the unknown regression function h is identically zero. Otherwise, $\hat{\beta}$ is asymptotically more efficient than $\tilde{\beta}$. As in this example, a statistical model with a likelihood which can be written as

$$L(\theta, \kappa) = L_1(\beta, \kappa)L_2(\alpha)$$

is called a factorised likelihood model. It was first implied by Pierce (1982) that the paradoxical phenomenon (21) occurs in factorised likelihood models; see also Barndorff-Nielsen & Cox (1994, Ch. 3). The other examples mentioned in § 1 are also based on factorised likelihood models.

The orthogonality of $u_{*\beta}(x, \theta)$ and $u_{*\alpha}(x, \theta)$, which is the condition for equality of (12), implies the occurrence of the paradox as discussed in Theorem 1. On the other hand, if the condition for equality of (16),

$$u_{\beta*}(x, \theta) = u_{*\beta}(x, \theta), \quad (26)$$

is satisfied, then from Lemmas 2 and 3 we have the reverse statement of Theorem 1, namely

$$\text{avar}(\tilde{\beta}) \leq \text{avar}(\hat{\beta}). \quad (27)$$

In other words, we obtain paradoxical and nonparadoxical results in accordance with conditions (20) and (26), respectively. Note that, if the estimating function $u_{\beta}(x, \theta)$ coincides with the score function for β in the parametric case, then condition (26) is satisfied. Conditions (20) and (26) can be restated in terms of estimators as in the following proposition.

PROPOSITION 1. (i) *Condition (20) is equivalent to the asymptotic independence of $\hat{\beta}$ and $\hat{\alpha}$.*
(ii) *Condition (26) is equivalent to the asymptotic independence of $\tilde{\beta}$ and $\hat{\alpha}$.*

The proof is given in the Appendix. In § 1, we mentioned the following relationship among the estimators:

$$\sqrt{n}(\tilde{\beta} - \beta) = \sqrt{n}(\hat{\beta} - \beta) + \sqrt{n}D(\hat{\alpha} - \alpha) + o_p(1),$$

where the matrix D is the same as that defined in Theorem 1. This is obtained from the Taylor expansion of the estimating equations of (10). If $\hat{\beta}$ and $\hat{\alpha}$ are asymptotically independent, we observe (21), whereas, if $\tilde{\beta}$ and $\hat{\alpha}$ are asymptotically independent, we observe (27). Thus, the two conditions of asymptotic independence have been clarified by the geometrical conditions (20) and (26) for estimating functions, which are more easily verified than asymptotic independence.

4. EXAMPLE

In this section, we consider a simple example and confirm the paradox numerically. Let $y_1, \dots, y_n$ be a random sample, with associated density $f(y; \beta)$. However, the sample is assumed not to have been completely observed. Let r_i be a missing-data indicator, so that $r_i = 1$ if y_i is missing and $r_i = 0$ otherwise. We consider the situation in which y_i and r_i are conditionally independent given a variable $x_i = (1, x_{2i}, \dots, x_{pi})'$, which is completely observable. Such a missing-data problem has been summarised in Little & Rubin (2002, Ch. 1). The variable x_i may be associated with both the outcome and the missing indicator. In this situation it is natural to consider the estimating equation

$$\sum_{i=1}^n \frac{1 - r_i}{\text{pr}(r_i = 0 | x_i)} \frac{\partial}{\partial \beta} \log f(y_i; \beta) = 0 \quad (28)$$

for β ; see Rotnitzky & Robins (1995) for detailed discussion of this type of estimating equation. Here, the conditional probability for y_i to be observed given x_i , $\text{pr}(r_i = 0 | x_i)$, is also called a propensity score, as described in Little & Rubin (2002, Ch. 3), although

this situation is not necessarily related to causal inference. Such a model is called a missing-at-random model. If the joint distribution of y_i and x_i is specified parametrically, the parameter β can be estimated by maximising the joint likelihood function. Thus, the missing-at-random condition usually ensures ignorability of the missing-data mechanism in maximum likelihood estimation. In the current case, however, we need to modify the estimating equation for the maximum marginal likelihood estimation of β because only the marginal density of Y_i is specified parametrically, as $f(y; \beta)$. Under certain regularity conditions, we can obtain a consistent estimator of β as the solution of the estimating equation (28) if the propensity score is known. However, the propensity score is usually unknown, and a logistic regression model is often used in estimating the propensity score:

$$\text{pr}(R_i = 0|x_i) = e(x_i; \alpha) = \frac{\exp(\alpha'x_i)}{1 + \exp(\alpha'x_i)},$$

where α is an unknown vector of nuisance parameters. Let $\hat{\alpha}$ be the maximum likelihood estimator of α , and let $\hat{\beta}$ and $\tilde{\beta}$ be estimators defined as the respective solutions of (28) based on an estimated propensity score with α substituted by $\hat{\alpha}$ and based on a known propensity score with α known. Then we observe the paradoxical relationship

$$\text{avar}(\tilde{\beta}) \geq \text{avar}(\hat{\beta}).$$

The proof is the same as that for the example of § 1. The difference between these asymptotic variance matrices given by Theorem 1 is

$$I(\beta)^{-1}J(\beta, \alpha)[E\{r(x_i; \alpha)x_i x_i'\}]^{-1}J(\beta, \alpha)'I(\beta)^{-1}, \quad (29)$$

where

$$I(\beta) = E\{s_\beta(y_i; \beta)s_\beta(y_i; \beta)'\}, \quad J(\beta, \alpha) = E[\{1 - e(x_i; \alpha)\}s_\beta(y_i; \beta)x_i'], \\ r(x; \alpha) = e(x; \alpha)\{1 - e(x; \alpha)\}.$$

In statistical inference with missing data, we often encounter weighted estimating equations such as (28). For example, Robins et al. (1994) and Lawless et al. (1999) discussed parameter estimation for a regression model when both response variables and covariates may be missing, but the missing-at-random condition is satisfied. Both papers proposed an estimating equation weighted by the inverse of the propensity score and noted the paradox we have discussed herein.

We now provide a simple numerical confirmation of the main result discussed in § 3. Assume that Y_i has a normal distribution with mean μ and variance σ^2 and that the propensity score is modelled by a logistic regression with

$$e(x; \alpha) = \frac{1}{1 + \exp(\alpha_1 + \alpha_2 x)}, \quad (30)$$

so that the parameters of interest are $\beta = (\mu, \sigma^2)'$ and the nuisance parameters are $\alpha = (\alpha_1, \alpha_2)'$. Estimators $\hat{\beta}$ and $\tilde{\beta}$ are expressed in weighted forms as follows:

$$\hat{\beta} = (\hat{\mu}_\alpha, \hat{\sigma}_\alpha^2)', \quad \tilde{\beta} = (\hat{\mu}_\alpha, \hat{\sigma}_\alpha^2)',$$

where

$$(\hat{\mu}_\alpha, \hat{\sigma}_\alpha^2) = \left(\frac{\sum_{i=1}^n w_i(\alpha) Y_i}{\sum_{i=1}^n w_i(\alpha)}, \frac{\sum_{i=1}^n w_i(\alpha) (Y_i - \hat{\mu}_\alpha)^2}{\sum_{i=1}^n w_i(\alpha)} \right)$$

with $w_i(\alpha) = (1 - R_i)e(X_i, \alpha)^{-1}$. We specify that the true distribution of (Y_i, X_i, R_i) has a density decomposed into (30) together with, for the marginal distribution of (Y_i, X_i) , a bivariate normal with mean vector $(\mu, 0)$ and variance matrix

$$\begin{pmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{pmatrix}.$$

Note that the case of $\rho = 0$, which is equivalent to the independence of Y_i and X_i , corresponds to ‘missing completely at random’. We investigate the asymptotic efficiency of $\tilde{\beta}$ relative to $\hat{\beta}$, namely

$$\text{aeff} = \frac{\det \{\text{avar}(\hat{\beta})\}}{\det \{\text{avar}(\tilde{\beta})\}}.$$

Here we take $\beta = (0, 1)$ and $\alpha = (1, 1)$. As a function of ρ , aeff attains a maximum of 1 at $\rho = 0$; see Fig. 2. This is because $J(\beta, \alpha)$ in (29) vanishes when $\rho = 0$. Thus, the situation of ‘missing completely at random’ leads to asymptotic equivalence between $\hat{\beta}$ and $\tilde{\beta}$. For a variety of cases of β and α the graphs of aeff against ρ are unimodal plots similar to that in Fig. 2.

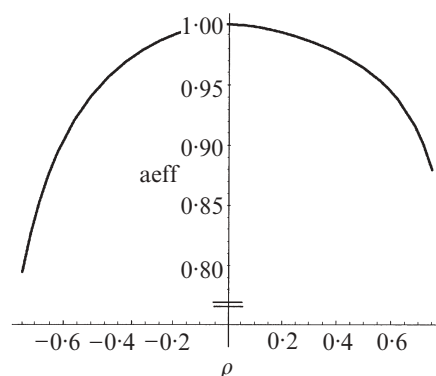


Fig. 2. The plot of asymptotic efficiency of $\tilde{\beta}$ relative to $\hat{\beta}$ against ρ in $(-0.75, 0.75)$.

5. CONCLUDING REMARKS

All the examples of the paradox we have discussed are based on the factorised likelihood model. The result in this case is relatively well known. The message of this paper is that use of the projected estimating functions makes it possible to understand the structure of the paradox clearly from a geometrical point of view. In addition, we can obtain a more general result than that for factorised likelihood models as a natural consequence of the geometrical argument.

When we discuss the asymptotic efficiency of estimators, it is common to consider the influence functions. For the estimator obtained from an estimating function $u(x, \theta)$, the influence function is given by the standardised estimating function of $u(x, \theta)$; that is,

$$u^*(x, \theta) = - \left[E \left\{ \frac{\partial u}{\partial \theta}(x, \theta) \right\} \right]^{-1} u(x, \theta).$$

Then the asymptotic variance matrix of the estimator is represented by the variance matrix of $u^*(x, \theta)$, which is just the inverse of the representation provided by the projected estimating function $u_*(x, \theta)$. This can be interpreted as a consequence of the duality of $u^*(x, \theta)$ and $u_*(x, \theta)$ in the sense that

$$E\{u^*(x, \theta)u_*(x, \theta)'\} = I,$$

where I denotes the identity matrix. The standardised estimating function $u^*(x, \theta)$, however, does not give such clear understanding of the structure of the paradox.

The paradoxical phenomenon does not occur if the estimating function for a parameter of interest is optimal in the sense that the corresponding estimator has the minimum asymptotic variance matrix. In the parametric case, the optimal estimating function is given by the score function; in the semiparametric case, the optimal estimating function is given by projecting the score function on to the linear space consisting of all the estimating functions (Amari & Kawanabe, 1997; Heyde, 1997, Ch. 2). However, this optimal estimating function may not be available because it usually depends on unknown infinite-dimensional nuisance parameters. Consequently, the phenomenon is more often observed in the semiparametric case than in the parametric case.

ACKNOWLEDGEMENT

The authors would like to thank the editor and the three referees, one in particular, for many helpful suggestions that greatly improved the presentation of this paper.

APPENDIX

Proofs

In this appendix, we omit the arguments of the estimating functions and the score functions, writing $u_{*\beta} = u_{*\beta}(x, \theta)$ and $s_\beta = s_\beta(x, \theta, \kappa)$.

Proof of Lemma 2. By Lemma 1, the asymptotic variance matrix of $\hat{\theta} = (\hat{\beta}', \hat{\alpha}')'$ is

$$\text{avar}(\hat{\theta}) = \begin{bmatrix} J_{\beta\beta} & J_{\beta\alpha} \\ J_{\alpha\beta} & J_{\alpha\alpha} \end{bmatrix}^{-1} = \begin{bmatrix} J_{\beta\beta \cdot \alpha}^{-1} & -J_{\beta\beta}^{-1} J_{\beta\alpha} J_{\alpha\alpha}^{-1} \\ -J_{\alpha\alpha}^{-1} J_{\alpha\beta} J_{\beta\beta \cdot \alpha}^{-1} & J_{\alpha\alpha \cdot \beta}^{-1} \end{bmatrix}, \quad (\text{A1})$$

where

$$J_{\beta\beta} = \text{var}(u_{*\beta}), \quad J_{\beta\alpha} = E\{u_{*\beta}(u_{*\alpha})'\}, \quad J_{\alpha\beta} = (J_{\beta\alpha})', \quad J_{\alpha\alpha} = \text{var}(u_{*\alpha}),$$

$$J_{\beta\beta \cdot \alpha} = J_{\beta\beta} - J_{\beta\alpha} J_{\alpha\alpha}^{-1} J_{\alpha\beta}, \quad J_{\alpha\alpha \cdot \beta} = J_{\alpha\alpha} - J_{\alpha\beta} J_{\beta\beta}^{-1} J_{\beta\alpha}.$$

Therefore, the asymptotic variance matrix of $\hat{\beta}$ is

$$\text{avar}(\hat{\beta}) = J_{\beta\beta \cdot \alpha}^{-1} = (J_{\beta\beta} - J_{\beta\alpha} J_{\alpha\alpha}^{-1} J_{\alpha\beta})^{-1}. \quad (\text{A2})$$

The matrix $J_{\beta\alpha} J_{\alpha\alpha}^{-1} J_{\alpha\beta}$ is the variance matrix of the orthogonal projection of $u_{*\beta}$ on to the linear space spanned by the components of $u_{*\alpha}$. Hence, by Pythagoras, we have

$$J_{\beta\beta} - J_{\beta\alpha} J_{\alpha\alpha}^{-1} J_{\alpha\beta} = \text{var}\{u_{*\beta} - \Pi(u_{*\beta} | u_{*\alpha})\}. \quad (\text{A3})$$

Equations (A2) and (A3) yield equations (11), (12) and (13). $\square$

Proof of Theorem 1. The difference between the asymptotic variance matrices of $\hat{\beta}$ and $\tilde{\beta}$ is calculated in the following way. First, the asymptotic variance matrix of $\tilde{\beta}$ is represented as

$$\text{avar}(\tilde{\beta}) = \{\text{var}(u_{\beta*})\}^{-1} = [E\{u_{\beta}(s_{\beta})'\}]^{-1} \text{var}(u_{\beta})[E\{s_{\beta}(u_{\beta})'\}]^{-1}. \quad (\text{A4})$$

Next, the asymptotic variance matrix of $\hat{\beta}$ is represented as

$$\text{avar}(\hat{\beta}) = \{\text{var}(u_{*\beta})\}^{-1}. \quad (\text{A5})$$

Since the two parts $u_{*\beta}$ and $u_{*\alpha}$ of the projected estimating function u_* are orthogonal and form a basis of the linear space spanned by $u = (u'_{\beta}, u'_{\alpha})'$, we have

$$u_{\beta} = E\{u_{\beta}(u_{*\beta})'\} \{\text{var}(u_{*\beta})\}^{-1} u_{*\beta} + E\{u_{\beta}(u_{*\alpha})'\} \{\text{var}(u_{*\alpha})\}^{-1} u_{*\alpha}.$$

This leads to the equation

$$\{\text{var}(u_{*\beta})\}^{-1} = [E\{u_{\beta}(u_{*\beta})'\}]^{-1} \text{var}(r)[E\{u_{*\beta}(u_{\beta})'\}]^{-1} = [E\{u_{\beta}(s_{\beta})'\}]^{-1} \text{var}(r)[E\{s_{\beta}(u_{\beta})'\}]^{-1}, \quad (\text{A6})$$

where $r = u_{\beta} - E\{u_{\beta}(u_{*\alpha})'\} \{\text{var}(u_{*\alpha})\}^{-1} u_{*\alpha}$. From (A4), (A5) and (A6), we obtain

$$\text{avar}(\tilde{\beta}) - \text{avar}(\hat{\beta}) = D \{\text{var}(u_{*\alpha})\}^{-1} D',$$

where $D = [E\{u_{\beta}(s_{\beta})'\}]^{-1} E\{u_{\beta}(s_{\alpha})'\}$. This yields (23) because $\text{avar}(\hat{\alpha}) = \{\text{var}(u_{*\alpha})\}^{-1}$ when $u_{*\beta}$ and $u_{*\alpha}$ are orthogonal. $\square$

Proof of Proposition 1. (i) By equation (A1), we find immediately that the asymptotic independence of $\hat{\beta}$ and $\hat{\alpha}$ is equivalent to the condition $J_{\beta\alpha} = 0$, which means the orthogonality of $u_{*\beta}$ and $u_{*\alpha}$.

(ii) The estimators $\hat{\beta}$, $\hat{\alpha}$ and $\tilde{\beta}$ are defined to be the solutions of the estimating equations

$$\sum_{i=1}^n u(x_i, \hat{\beta}, \hat{\alpha}) = 0, \quad \sum_{i=1}^n u_{\beta}(x_i, \tilde{\beta}, \alpha) = 0.$$

Hence, the usual Taylor expansion leads to

$$\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}(\tilde{\beta} - \beta) - \sqrt{n}D(\hat{\alpha} - \alpha) + o_p(1),$$

where the matrix D is the same as that defined in Theorem 1. This equation can be also written in the form

$$\sqrt{n} \begin{pmatrix} \tilde{\beta} - \beta \\ \hat{\alpha} - \alpha \end{pmatrix} = \sqrt{n} \begin{pmatrix} I & D \\ 0 & I \end{pmatrix} \begin{pmatrix} \hat{\beta} - \beta \\ \hat{\alpha} - \alpha \end{pmatrix} + o_p(1), \quad (\text{A7})$$

where I denotes an identity matrix. By (A7), we obtain the asymptotic variance matrix of $(\tilde{\beta}', \hat{\alpha}')'$ as

$$\begin{aligned} \text{avar} \begin{pmatrix} \tilde{\beta} \\ \hat{\alpha} \end{pmatrix} &= \begin{pmatrix} I & D \\ 0 & I \end{pmatrix} \begin{pmatrix} J^{\beta\beta} & J^{\beta\alpha} \\ J^{\alpha\beta} & J^{\alpha\alpha} \end{pmatrix} \begin{pmatrix} I & D \\ 0 & I \end{pmatrix}' \\ &= \begin{pmatrix} J^{\beta\beta} + DJ^{\alpha\beta} + J^{\beta\alpha}D' + DJ^{\alpha\alpha}D' & J^{\beta\alpha} + DJ^{\alpha\alpha} \\ J^{\alpha\beta} + J^{\alpha\alpha}D' & J^{\alpha\alpha} \end{pmatrix}, \end{aligned} \quad (\text{A8})$$

where $J^{\beta\beta}$, $J^{\beta\alpha}$, $J^{\alpha\beta}$ and $J^{\alpha\alpha}$ are the components of the asymptotic variance matrix of $(\hat{\beta}', \hat{\alpha}')'$ in the block matrix form, which is given by (A1). Equation (A8) gives a necessary and sufficient condition for the asymptotic independence of $\tilde{\beta}$ and $\hat{\alpha}$ as follows:

$$J^{\beta\alpha} + DJ^{\alpha\alpha} = 0.$$

Based on equation (A1), this condition is equivalent to the condition

$$\{\text{var}(u_{*\beta})\}^{-1} E\{u_{*\beta}(u_{*\alpha})'\} = [E\{u_{*\beta}(u_{*\beta})'\}]^{-1} E\{u_{*\beta}(u_{*\alpha})'\}. \quad (\text{A9})$$

Therefore, it is easy to determine that the asymptotic independence of $\tilde{\beta}$ and $\hat{\alpha}$ follows from $u_{\beta*} = u_{*\beta}$. On the other hand, when $\tilde{\beta}$ and $\hat{\alpha}$ are asymptotically independent, we define the matrix Λ such that

$$E\{u_{\beta*}(u_{*\beta})'\} = \Lambda\{\text{var}(u_{*\beta})\}^{-1}. \quad (\text{A10})$$

Equation (A9) can then be used to obtain

$$E\{u_{\beta*}(u_{*\alpha})'\} = \Lambda E\{u_{*\beta}(u_{*\alpha})'\}. \quad (\text{A11})$$

Equations (A10) and (A11) can be rewritten as

$$E\{(u_{\beta*} - \Lambda u_{*\beta})(u_{*\beta})'\} = 0, \quad E\{(u_{\beta*} - \Lambda u_{*\beta})(u_{*\alpha})'\} = 0.$$

We obtain

$$u_{\beta*} - \Lambda u_{*\beta} = 0,$$

since $u_{*\beta}$ and $u_{*\alpha}$ are linearly independent and $u_{\beta*} - \Lambda u_{*\beta}$ belongs to the linear space spanned by $u_{*\beta}$ and $u_{*\alpha}$. Hence, we have $u_{\beta*} = u_{*\beta}$ because $u_{*\beta}$ is the orthogonal projection of the score function for β on to the linear space spanned by u_{β} and u_{α} , and $u_{\beta*}$ is the orthogonal projection on to its linear subspace.

REFERENCES

- AMARI, S. & KAWANABE, M. (1997). Information geometry of estimating functions in semiparametric statistical models. *Bernoulli* **3**, 29–54.
- BARNDORFF-NIELSEN, O. E. & COX, D. R. (1994). *Inference and Asymptotics*. London: Chapman and Hall.
- COX, D. R. & REID, N. (1987). Parameter orthogonality and approximate conditional inference (with Discussion). *J. R. Statist. Soc. B* **49**, 1–39.
- GODAMBE, V. P. (1991). Estimating functions: an overview. In *Estimating Functions*, Ed. V. P. Godambe, pp. 3–20. New York: Oxford University Press.
- HEYDE, C. C. (1997). *Quasi-Likelihood and its Application*. New York: Springer.
- KOBAYASHI, S. & NOMIZU, K. (1996). *Foundations of Differential Geometry*, **1**. New York: Wiley.
- LAWLESS, J. F., KALBFLEISCH, J. D. & WILD, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. R. Statist. Soc. B* **61**, 413–38.
- LITTLE, R. J. A. & RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: Wiley.
- PIERCE, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Ann. Statist.* **10**, 475–8.
- ROBINS, J. M., MARK, S. D. & NEWBY, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48**, 479–95.
- ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.
- ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
- ROTNITZKY, A. & ROBINS, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82**, 805–20.
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.
- ZOU, F. & FINE, J. P. (2002). A note on a partial empirical likelihood. *Biometrika* **89**, 958–61.

[Received March 2003. Revised February 2004]