University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2004 *Paper* 163

Choice of Monitoring Mechanism for Optimal Nonparametric Functional Estimation for Binary Data

Nicholas P. Jewell* Mark J. van der Laan†
Stephen Shiboski‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, jewell@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

‡Department of Epidemiology and Biostatistics, University of California, San Francisco, steve@biostat.ucsf.edu

http://biostats.bepress.com/ucbbiostat/paper163

Choice of Monitoring Mechanism for Optimal Nonparametric Functional Estimation for Binary Data

Nicholas P. Jewell, Mark J. van der Laan, and Stephen Shiboski

Abstract

Optimal designs of dose levels in order to estimate parameters from a model for binary response data have a long and rich history. These designs are based on parametric models. Here we consider fully nonparametric models with interest focused on estimation of smooth functionals using plug-in estimators based on the nonparametric maximum likelihood estimator. An important application of the results is the derivation of the optimal choice of the monitoring time distribution function for current status observation of a survival distribution. The optimal choice depends in a simple way on the dose response function and the form of the functional. The results can be extended to allow dependence of the monitoring mechanism on covariates.

1 Introduction

A common problem in dose response experiments is estimation the relationship between the level of a dose, C and the probability of a binary response, denoted by $F(C)$. Suppose the function $F = F_\theta$ is parametrically modeled by say a logistic or probit function, and that n_i observations are taken at a set of k dose levels $c_1, \ldots, c_k$. A natural design question relates to the optimal choice of $c_1, \ldots, c_k$ with regard to efficient estimation of all or some components of θ. See, for example, Sitter (1992) and the references therein. Such optimization often leads to two or three point designs and depend on the unknown value of θ.

Sitter (1992) tackles the issue that the optimal design depends on unknown values of θ using a minimax approach over a region of possible values for θ, but does not consider that the parametric model for F_θ is also assumed to be known in advance. Here we consider optimal choice of the dose levels where the form of F is unspecified and interest focuses on estimation of a single functional of F.

The results have immediate application to estimation of functionals of the distribution, F, of a survival random variable, T, where estimation is based on current status data; here, observation of T is restricted to knowledge of whether or not T exceeds a random independent monitoring time C. Nonparametric estimation of the survival function, and semi-parametric techniques for related regression models, based on current status data, are reviewed in Jewell & van der Laan (2004). In detail, let T be the survival random variable of interest, with associated distribution function F. Assume that the monitoring time, C, is randomly selected from a distribution function G, independently of T. An independent and identically distributed sample of n individuals is therefore drawn from the

joint distribution of (T, C); however, only $\{(\Delta_i, C_i : i = 1, \ldots, n\}$ is observed where $\Delta = I(T \leq C)$. In this context, the design question relates to optimal choice of G for estimation of a given functional of F, based on such current status data. In some settings, choice of the monitoring times may not be under the control of the investigator; however, in many applications in carcinogenecity testing and cross-sectional disease incidence estimation, monitoring times may be pre-selected. We use current status notation in what follows below.

2 Optimal Choice of G with F unspecified

Nonparametric maximum likelihood estimation of the distribution function, F, of T from current status data is easily implemented using the pool-adjacent-violator algorithm (Ayer et al, 1955). Here we wish to select the distribution function, G, of C, in terms of minimizing the asymptotic variance of a specific functional estimate.

The properties of the nonparametric maximum likelihood estimator F_n of F, based on current status data, were established by Groeneboom & Wellner (1992) who, in particular, considered the efficiency of smooth functionals of F_n as estimators of the corresponding functional of F. The estimator, F_n, is known to converge only at the rate $n^{-1/3}$. However, plug-in estimates of smooth functionals are asymptotically Gaussian, converging at the standard rate $n^{-1/2}$.

In detail, consider the parameter $\mu = \int (1 - F(u))r(u)du$ for some function r, and the corresponding estimator $\mu_n = \int (1 - F_n(u))r(u)du$. Suppose there is a constant $M < \infty$ so that (i) r is bounded on $[0, M]$, (ii) F is continuous with a density $f > 0$ on $[0, M]$

and zero elsewhere, and (iii) $g(c) = dG/dc > 0$ on $[0, M]$. Huang & Wellner (1995) proved that, for any pair (F, G) and function r that satisfy (i)–(iii), the estimator μ_n is regular and asymptotically linear with the variance of its influence curve given by

$$VAR(IC) = \int \frac{r^2(c)}{g(c)} F(c)(1 - F(c))dc. \tag{1}$$

The question we pose here is that, for a given r and F, what choice of the monitoring time distribution G minimizes the variance of the influence function for μ_n? That is we seek the G that minimizes the right hand side of (1).

To solve this optimization problem, we perform an "infinite dimensional differentiation" of (1) with respect to the density g corresponding to G. Specifically, let h be any function in ${L_0}^2(G)$, the set of all square-integrable functions with respect to the measure dG that satisfy $\int h(c)dG(c) = 0$; then, for any g_0 and for a small enough positive number ϵ, $(1 + \epsilon h)g_0$ describes a one-dimensional family of densities that passes through g_0 at $\epsilon = 0$. If g_0 minimizes (1), it follows that the function

$$\epsilon \to \int \frac{r^2(c)}{(1 + \epsilon h(c))g_0(c)} F(c)(1 - F(c)dc \tag{2}$$

has a minimum at $\epsilon = 0$. That is,

$$\frac{d}{d\epsilon} \int \frac{r^2(c)}{(1 + \epsilon h(c))g_0(c)} F(c)(1 - F(c))dc \Big|_{\epsilon=0} = 0.$$

This yields $\int \frac{r^2(c)}{g_0(c)} F(c)(1 - F(c)h(c)dt = 0$. This is equivalent to saying that

$$\int \frac{r^2(c)}{[g_0(t)]^2} F(c)(1 - F(c)h(c)dG(c) = 0$$

Since this is true for all h in ${L_0}^2(G)$, it follows that

$$\frac{r^2(c)}{[g_0(c)]^2} F(c)(1 - F(c)) = K,$$

for some constant K. Solving for K by normalizing then yields

$$g_0(c) = \frac{|r(c)|F(c)^{1/2}(1-F(c))^{1/2}}{K^*}, \tag{3}$$

where the constant $K^* = \int |r(c)|F(c)^{1/2}(1-F(c))^{1/2}dc$. To complete this analysis, we must show that this g_0 in fact yields a minimum of (2). This is seen by taking the second derivative of (2), and evaluating at $\epsilon = 0$; this yields

$$2\int \frac{r^2(c)}{g_0(c)}F(c)(1-F(c)h(c)^2dc = 2K^*\int |r(c)|F(c)^{1/2}(1-F(c))^{1/2}h(c)^2dc > 0,$$

as desired.

We have thus shown that the optimal g_0 depends on the function r and F through (3).

We briefly consider two simple examples where interest focuses on (i) the mean, and (ii) the variance of F. For the mean, take $r(c) \equiv 1$. Here, the optimal choice is $g_0 \propto F^{1/2}(1-F)^{1/2}$; thus monitoring times (or doses) should be concentrated around the median of F. Alternatively, for the variance, take $r(c) = 2c - E(F)$, with the subsequent optimal choice given by $g_0(c) \propto |2c - E(F)|F^{1/2}(1-F)^{1/2}$. In this case, monitoring times (doses) will be much less concentrated around the median of F with more weight given to values in the tails of F.

For illustration, suppose the unknown F is described by an exponential distribution with mean 1, conditional on being less than 10. Figure 1 illustrates the optimum choice of g for estimation of the mean and variance of F, based on the nonparametric maximum likelihood estimator.

FIGURE 1 ABOUT HERE

3 Allowing the Optimal Choice of G to Depend on Covariates

We extend the result of §2 to allow for monitoring designs that are allowed to depend on a k dimensional fixed covariate Z. The assumption that C and T are independent is now loosened to C being independent of T, given Z. In nonparametrically estimating the functional μ, based on observed data $\{(\Delta_i, C_i, Z_i) : i = 1, \ldots, n\}$, the efficient influence curve is given by

$$\begin{aligned} IC_{eff}(c) &= \frac{r(c)\{F(c|Z) - \Delta\}}{g(c|Z)} + \int_0^\infty r(u)\{1 - F(u|Z)\}du - \mu \\ &= \frac{r(c)\{F(c|Z) - \Delta\}}{g(c|Z)} + \int_0^\infty r(u)\{\bar{F}(u|Z) - \bar{F}(u)\}du, \end{aligned}$$

with $\bar{F} = 1 - F$, a special case of (4.12) in van der Laan & Robins (2003, p. 242). The variance of this influence curve is then

$$\begin{aligned} E\left(IC_{eff}{}^2\right) &= E\left[\frac{r^2(c)\left\{F(c|Z) - \Delta\right\}^2}{g^2(c|Z)}\right] + E\left[\left(\int_0^\infty r(u)\{\bar{F}(u|Z) - \bar{F}(u)\}du\right)^2\right] \\ &\quad +2E\left[\frac{r(c)\{F(c|Z) - \Delta\}}{g(c|Z)}\int_0^\infty r(u)\{\bar{F}(u|Z) - \bar{F}(u)\}du\right] \\ &= E\left[\frac{r^2(c)\left\{F(c|Z - \Delta\right\}^2}{g^2(c|Z)}\right] + \phi(F_Z), \end{aligned}$$

where $E(\cdot)$ is the expectation with respect to the data generating distribution, and F_Z is the marginal distribution of Z, which does not depend on g. The second step in this derivation follows from taking conditional expectations in the right order. We now seek the optimal set of conditional densities $g(c|Z)$ that minimizes the expectation

$$E\left[\frac{r^2(c)\left\{F(c|Z - \Delta\right\}^2}{g^2(c|Z)}\right] = E_{F_Z}\left[\int_0^\infty \frac{r^2(c)F(c|Z)\left\{1 - F(c|Z)\right\}}{g(c|Z)}dc\right].$$

For a fixed Z, an identical argument to §2 shows that the density that optimizes $\left[\int_0^\infty \frac{r^2(c)F(c|Z)\{1-F(c|Z)\}}{g(c|Z)}dc\right]$ is given by

$$g_0(c|Z) = \frac{|r(c)|F(c|Z)^{1/2}(1-F(c|Z))^{1/2}}{K^*(Z)}, \tag{4}$$

with the normalizing constant $K^* = \int |r(c)|F(c|Z)^{1/2}(1-F(c|Z))^{1/2}dc$, as before. It immediately follows that the densities (4), for all Z, provide the optimal conditional monitoring densities.

4 Further Extensions

In many examples, particularly in the presence of covariates, interest focuses on functionals that are not merely based on the marginal distribution F. For example, if we assume a regression model linking T with Z of the form $E(T|Z) = \beta Z$, we may wish to select a monitoring distribution to optimize estimation of β. A simple example of this occurs in a two group comparison of the mean of T. As before, van der Laan & Robins (2003, p. 242) provides the relevant efficient influence curve for estimation of a smooth functional $\mu(F_{T,Z})$ of the joint distribution $F_{T,Z}$ of (T, Z). In particular, suppose $D(T, Z)$ is the efficient influence curve for $\mu(F_{T,Z})$ in the full data world where $\{(T_i, Z_i) : i = 1, \ldots, n\}$ is observed. Let $a_{g,Z}$ be the left end point of the support of the density $g(\cdot|Z)$. Then, the analogous efficient influence curve based on $\{(\Delta_i, C_i, Z_i) : i = 1, \ldots, n\}$ is given by

$$IC_{eff} = \frac{D'(c, Z)\{F(c|Z) - \Delta\}}{g(c|Z)} + \int_0^\infty D'(u, Z)\{1 - F(u|Z)\}du + D(a_{g,Z}),$$

where $D'(t, Z) = \frac{\partial D(t,Z)}{\partial t}$. For simplicity, we now assume that $a_{g,Z}$ does not vary with g and, in particular, agrees with the left end point of the support of F. Then, the same approach

as in §3 shows that the optimal conditional monitoring densities are given by

$$g_0(c|Z) = \frac{|D'(c,Z)|F(c|Z)^{1/2}(1-F(c|Z))^{1/2}}{K^*(Z)},$$

with normalizing constant $K^* = \int |D'(c,Z)|F(c|Z)^{1/2}(1-F(c|Z))^{1/2}dc$.

For the results in §2–3, it is desirable to allow that some of the components of Z be time-dependent. In this case, the efficient influence curve is the implicit solution to an integral equation, and so it is not easy to see how optimization can proceed straightforwardly. In practice, discrete sequential choice of future monitoring times might be based on current values of the time dependent covariates using the results of §3.

5 Discussion

In practice, of course, F is no more known a priori than θ in the parametric setting. Thus, an optimum design based on a presumed F may be somewhat different than the true F in the experimental setting. We suggest therefore that a series of plausible Fs be considered along with the associated optimum design. Then, for each such F, the relevant variance of the desired functional can be calculated from (1) over the range of possible optimal designs under consideration. As in Sitter (1992) a minimax criterion could then be used to select a particular design that is robust to some misspecification of F. At the very least, optimum nonparametric and parametric designs can be compared to illuminate how much the design depends on a particular parametric model choice. Similarly, to exploit the role of covariates a plausible regression model for $F(c|Z)$ must be invoked to derive the optimal monitoring densities $g(c|Z)$.

We have focused here on estimation of a single functional. In many examples, investi-

gators may wish to estimate several functionals efficiently and simultaneously. In principle, the joint influence curve can be calculated as in (1) although now we have several possible optimality criteria, including D-, A-, and E-optimality (see Sitter, 1992). Any of these approaches can serve as the basis of optimal choice of g.

References

Ayer, M, Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26**, 641-7.

Groeneboom, P, Wellner, J.A. (1980) *Nonparametric Maximum Likelihood Estimators for Interval Censoring and Denconvolution.* Boston: Birkhäuser.

Huang, J., Wellner, J.A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I. *Statist. Neer.* **49**, 153-63.

Jewell, N.P., van der Laan, M.J. (2004). Current status data: Review, recent developments and open problems. In *Advances in Survival Analysis* , Handbook in Statistics #23, 625-42, Amsterdam: Elsevier.

van der Laan, M.J., Robins, J.M. (2003) *Unified Methods for Censored Longitudinal Data and Causality.* New York: Springer.

Sitter, R.R. (1992). Robust designs for binary data. *Biometrika* **48**, 1145-55.

Figure 1: Optimal Choice of Monitoring Time Density, g_0, for Nonparametric Estimates of the Mean (dotted line), and Variance (dash-dotted line) of the Distribution Function F (with density given by the solid line)