Estimation from Current-Status Data in Continuous Time

NIELS KEIDING
University of Copenhagen, Denmark

KAMILLA BEGTRUP
University of Copenhagen, Denmark

THOMAS H. SCHEIKE
University of Copenhagen, Denmark

GÜNTHER HASIBEDER
Technical University of Vienna, Austria

Received December 14, 1995; accepted February 26, 1996

Abstract. The nonparametric maximum likelihood estimator for current-status data has been known for at least 40 years, but only recently have the mathematical-statistical properties been clarified. This note provides a case study in the important and often studied context of estimating age-specific immunization intensities from a seroprevalence survey. Fully parametric and spline-based alternatives (also based on continuous-time models) are given. The basic reproduction number $R_0$ exemplifies estimation of a functional. The limitations implied by the necessarily rather restrictive epidemiological assumptions are briefly discussed.

Keywords: Age-specific incidence, basic reproduction number, epidemiology, smoothing splines, Weibull survival distribution.

1. Introduction

Age-specific immunization rates are basic building blocks in any detailed epidemiological study of diseases with life-long immunity (measles, mumps, rubella, hepatitis). Direct estimation of these rates would be available from a follow-up study, in which a population of susceptibles would be followed for a period of time with all infections recorded, but this survey design is unfortunately rarely feasible. It is much more common to have a cross-sectional sample where for each person, current age and current immunization status is known. In the survival analysis context such data may be said to be all censored, either to the right (if immunization has not yet happened) or to the left (if the person is immune).

There has been considerable recent interest in the analysis of such current-status data, see Jewell and Shiboski (1990), Diamond and McDonald (1992), Shiboski and Jewell (1992), Grummer-Strawn (1993), Sun and Kalbfleisch (1993), Andersen and Rønn (1995), Rabinowitz et al. (1995), Jewell and van der Laan (1995) and Rossini and Tsiatis (1996). In this note we assume continuous time and focus on the nonparametric maximum likelihood estimator (NPMLE) first derived by Ayer et al. (1955). Early mathematical-statistical studies of the large-sample properties of such estimators were by Prakasa Rao (1969) and...
Brunk (1970), and a definitive treatment has been given by Groeneboom, see Groeneboom and Wellner (1992) for a survey. We review the statistical properties of this estimator and illustrate them in a particular example. We also compare the NPMLE to a fully parametric (Weibull) approach and an intermediate approach based on splines.

An important parameter for infectious diseases is the basic reproduction number or transmission potential \( R_0 \), intuitively the number of secondary cases which one case would produce in a susceptible population. Dietz (1993) recently surveyed methods for estimation of \( R_0 \) for infectious diseases. Dietz and Schenzle (1985) derived a formula for \( R_0 \) for populations with age-specific immunization rates, and allowing for particular age-specific vaccination policies. Dietz and Schenzle made the further assumptions of stationary population and proportionate mixing, and even though both of these may be somewhat crude in practice, it is still at least of methodological interest to study the estimation of \( R_0 \) based on Dietz-Schenzle assumptions.

The methods are illustrated throughout on rubella seroprevalence data for 223 males older than 3 months (for which exact date of birth was known) that were tested at the Institute of Virology, Vienna, in the period 1-25 March 1988. The Austrian vaccination policy against rubella had then for some time been to routinely immunize girls just before puberty but not to vaccinate males. In addition various particular risk groups of females (such as pre-school teachers) were offered immunization. It is likely that the herd immunity effect of this policy is minor so that the males may be taken as approximating an unvaccinated population.

2. Estimation of the Distribution Function

It is assumed throughout that the mortality is the same for susceptibles and immunes and that the immunity is life-long. The immunization intensity \( \lambda_0(a) \) is defined from \( \lambda_0(a) da = P \{ \text{immunization in } [a, a + da] \} \text{ person susceptible at age } a \}. \) Let \( \Lambda_0(a_0) = \int_0^{a_0} \lambda_0(a) da; \) our first task is to estimate the distribution function \( G(a) = 1 - \exp\{-\Lambda_0(a)\} \), the cumulative probability of being immunized by age \( a \).

Let \( Y_i \) denote the age of immunization of \( i \) (using the convention \( Y_i = \infty \) if \( i \) never gets immunized), \( Z_i \) the current age of person \( i \) and \( \delta_i = I\{Z_i > Y_i\} = 1 \) if \( i \) is immunized, \( \delta_i = 0 \) if not. Available are i.i.d. replications \((Z_1, \delta_1), \ldots, (Z_n, \delta_n)\), so that all \( Y_i \) are either right censored or left censored. The censoring pattern (i.e. the sampling method) is assumed to correspond to an assumption of independence of current age \( Z_i \) and age of immunization \( Y_i \). Let \( F \) and \( f \) denote distribution and density function of \( Z_1; G \) and \( F \) are assumed to be distinctly parametrized.

The likelihood for \( G \) in this situation (conditional on \( Z_i = z_i, i = 1, \ldots, n \)) is

\[
L = \prod_{i=1}^{n} G(z_i)^{\delta_i} [1 - G(z_i)]^{1-\delta_i},
\]

a product of \( n \) binomial likelihoods, each with 1 unit.

The nonparametric maximum likelihood estimator (NPMLE) in this particular situation has a well-known explicit representation (Ayer et al., 1955, Groeneboom and Wellner, 1992), see also Keiding (1991) for an exposition. Let \( Z_{(1)}, \ldots, Z_{(n)} \) denote the ordered