

MULTIPHASE DESIGNS AND ESTIMATION IN BIOSTATISTICS

J.F. Lawless¹

ABSTRACT

Multistage sampling schemes are often used in biostatistics and epidemiology for reasons of economy or convenience. We consider contexts where the relationship between a response variable Y and a vector of covariates x is of interest. At the first stage, individuals are selected and certain measurements are taken. The initial sample is then subsampled in subsequent stages and additional measurements are taken on selected individuals. As a result, different amounts of data are missing on different individuals. We consider approaches to estimation in such settings and discuss efficiency.

KEYS WORDS: Estimation; Multistage Sampling; Subsampling.

RÉSUMÉ

Les plans d'échantillonnage à plusieurs degrés sont souvent utilisés en biostatistique et en épidémiologie, et ce pour des raisons d'économie et de commodité. Dans ce travail, notre intérêt se porte sur la relation entre la variable réponse Y et un vecteur de covariables x . À la première étape, les individus sont sélectionnés et certaines mesures sont prises sur eux. L'échantillon initial subit ensuite une séquence de ré-échantillonnages fournissant des mesures additionnelles sur les individus sélectionnés. En conséquence, différentes quantités de données sont manquantes sur plusieurs individus. Nous considérons des approches d'estimation pour de telles situations et nous discutons des problèmes d'efficacité.

MOTS-CLÉS : Échantillonnage à plusieurs degrés; estimation; ré-échantillonnage.

EXTENDED ABSTRACT

Multiphase (often called "multistage") sampling schemes are often used in biostatistics and epidemiology for reasons of economy or convenience. Recent examples are given by Breslow and Holubkov (1997), Clayton et al. (1998), Lawless et al. (1998), Reilly and Pepe (1995) and references cited therein. We consider contexts where the relationship between a response variable Y and a vector of covariates x is of interest. At the first phase (stage), individuals are selected and certain measurements are taken. The initial sample is then subsampled and additional measurements are taken on selected individuals at this second phase. Selection probabilities for phase two in general depend on Y and x information from phase one. A third phase would consist of subsamples of individuals selected for phase two, and so on. As a

result different amounts of data are observed (or, conversely, missing) on different individuals.

We consider approaches to estimation in such settings, based on recent work by Lawless, Wild and Kalbfleisch (1998). Parametric models are used for the distribution Y given x but no parametric assumptions are made for the distribution of x . Semiparametric maximum likelihood is considered, along with several pseudo likelihood methods which have been proposed in the literature: these include weighted estimation (e.g., Zhao and Lipsitz 1992, Whittemore 1997), estimated likelihood (e.g. Pepe and Fleming 1991), mean score estimation (e.g., Reilly and Pepe 1995), and conditional pseudo likelihood (e.g. Breslow and Holubkov 1997). Lawless et al. (1998) unify and compare these methods.

¹ J.F. Lawless, Department of Statistics & Actuarial Science, University of Waterloo, Waterloo, Canada, N2L 3G1
E-mail: jlawless@uwaterloo.ca

Efficiency comparisons are made for a two-phase setting which involves a logistic regression model for a binary response. A simulation study of different scenarios indicates that maximum likelihood is the only method that performs well in all cases. The pseudo likelihood methods that are easiest to use are the weighted and mean score methods; they perform well in certain settings but not in others. Lawless et al. (1998) provide a more detailed discussion.

REFERENCES

Breslow, N.E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. Roy. Statist. Soc. B.* 59, 447-461.

Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998). Analysis of longitudinal binary data from multi-phase sampling. *J. Roy. Statist. Soc. B.* 60, 71-81.

Lawless, J.F., Wild, C.J. and Kalbfleisch, J.D. (1998). Semiparametric methods for response-selective and missing data problems in regression. To appear in *J. Roy. Statist. Soc. B.*

Pepe, M.S. and Fleming, T.R. (1991). A nonparametric method for dealing with mismeasured covariate data. *J. Amer. Statist. Assoc.* 86, 108-113.

Reilly, M. and Pepe, M.S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* 82, 299-314.

Whittemore, A.S. (1997). Multistage sampling designs and estimating equations. *J. Roy. Statist. Soc. B.* 59, 589-602.

Zhao, L.P. and Lipsitz, S. (1992). Design and analysis of two-stage studies. *Statistics in Med.* 11, 769-782.