Chan, Kwun Chuen Gary; **Wang, Mei-Cheng**

Estimating incident population distribution from prevalent data. (English) Zbl 1251.62040

Biometrics 68, No. 2, 521-531 (2012).

Summary: A prevalent sample consists of individuals who have experienced disease incidence but not failure event at the sampling time. We discuss methods for estimating the distribution function of a random vector defined at the baseline for an incident disease population when data are collected by prevalent sampling. Prevalent sampling designs are often more focused and economical than incident study designs for studying the survival distribution of a diseased population, but prevalent samples are biased by design. Subjects with longer survival time are more likely to be included in a prevalent cohort, and other baseline variables of interests that are correlated with survival time are also subject to sampling bias induced by the prevalent sampling scheme. Without recognition of the bias, applying empirical distribution functions to estimate the population distribution of baseline variables can lead to serious bias. In this article, nonparametric and semiparametric methods are developed for distribution estimation of baseline variables using prevalent data.

MSC:

62P10 Applications of statistics to biology and medical sciences; meta analysis
62G07 Density estimation
65C60 Computational problems in statistics (MSC2010)
92C50 Medical applications (general)

Keywords:

accelerated failure time model; cross-sectional sampling; left truncation; proportional hazards model

Full Text: DOI

References:

[1] Andersen, Statistical Models Based on Counting Process (1993) · doi:10.1007/978-1-4612-4348-9

[2] Asgharian, Length-biased sampling with right censoring: An unconditional approach, Journal of the American Statistical Association 97 pp 201– (2002) · Zbl 1073.62561 · doi:10.1198/016214502753479347

[3] Bach, Patient demographic and socioeconomic characteristics in the SEER-Medicare database: Applications and limitations, Medical Care 40 pp IV-19– (2002) · doi:10.1097/00005650-200208001-00003

[4] Bergeron, Covariate bias induced by length-biased sampling of failure times, Journal of the American Statistical Association 103 pp 737– (2008) · Zbl 05564527 · doi:10.1198/016214508000000382

[5] Brookmeyer, Biases in prevalent cohorts, Biometrics 43 pp 739– (1987) · Zbl 0715.62221 · doi:10.2307/2531529

[6] Carroll, Generalized partially linear single-index models, Journal of the American Statistical Association 92 pp 477– (1997) · Zbl 0890.62053 · doi:10.1080/01621459.1997.10474001

[7] Chen, On a general class of semiparametric hazards regression models, Biometrika 88 pp 687– (2001) · Zbl 0985.62086 · doi:10.1093/biomet/88.3.687

[8] Chen, Attributable risk function in the proportional hazards model for censored time-to-event, Biostatistics 7 pp 515– (2006) · Zbl 1170.62370 · doi:10.1093/biostatistics/kxj023

[9] Cox, A general definition of residuals, Journal of the Royal Statistical Society, Series B (Methodological) 30 pp 248– (1968) · Zbl 0164.48903

[10] Crowley, Covariance analysis of heart transplant survival data, Journal of the American Statistical Association 72 pp 27– (1977) · doi:10.1080/01621459.1977.10479903

[11] Gross, Nonparametric estimation and regression analysis with left-truncated and right-censored data, Journal of the American Statistical Association 91 pp 1166– (1996) · Zbl 0882.62037 · doi:10.1080/01621459.1996.10476986

[12] Gürler, A bivariate distribution function estimator and its variance under left truncation and right censoring (1996)

[13] Huang, Cox regression with accurate covariates unascertainable: A nonparametric-correction approach, Journal of the American Statistical Association 95 pp 1209– (2000) · Zbl 1008.62040 · doi:10.1080/01621459.2000.10474321

[14] Kalbfleisch, The Statistical Analysis of Failure Time Data (1980) · Zbl 0504.62096

[15] Keiding, Age-specific incidence and prevalence: A statistical perspective, Journal of the Royal Statistical Society, Series A (Statistics in Society) 154 pp 371– (1991) · Zbl 1002.62504 · doi:10.2307/2983150

[16] Keiding, Event history analysis and the cross-section, Statistics in Medicine 25 pp 2343– (2006) · doi:10.1002/sim.2579

[17] Lai, Rank regression methods for left-truncated and right-censored data, The Annals of Statistics 19 pp 531– (1991) · Zbl 0739.62031 · doi:10.1214/aos/1176348110

[18] Lin, Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators, Journal of the American Statistical Association 86 pp 725– (1991) · Zbl 0733.62048 · doi:10.1080/01621459.1991.10475101

[19] Schoenfeld, Chi-squared goodness-of-fit tests for the proportional hazards regression model, Biometrika 67 pp 145– (1980) · Zbl 0446.62039 · doi:10.1093/biomet/67.1.145

[20] Shen, Analyzing length-biased data with semiparametric transformation and accelerated failure time models, Journal of the American Statistical Association 104 pp 1192– (2009) · Zbl 1388.62294 · doi:10.1198/jasa.2009.tm08614

[21] Therneau, Martingale-based residuals for survival models, Biometrika 77 pp 147– (1990) · Zbl 0692.62082 · doi:10.1093/biomet/77.1.147

[22] Tsai, A note on the product-limit estimator under right censoring and left truncation, Biometrika 74 pp 883– (1987) · Zbl 0628.62101 · doi:10.1093/biomet/74.4.883

[23] van der Varrt, Weak Convergence and Empirical Processes (1996) · doi:10.1007/978-1-4757-2545-2

[24] Vardi, Empirical distributions in selection bias models, The Annals of Statistics 13 pp 178– (1985) · Zbl 0578.62047 · doi:10.1214/aos/1176346585

[25] Wang, A semiparametric model for randomly truncated data, Journal of the American Statistical Association 84 pp 742– (1989) · Zbl 0677.62037 · doi:10.1080/01621459.1989.10478828

[26] Ying, A large sample study of rank estimation for censored regression data, The Annals of Statistics 21 pp 76– (1993) · Zbl 0773.62048 · doi:10.1214/aos/1176349016

[27] Vardi, Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation, Biometrika 76 pp 751– (1989) · Zbl 0678.62051 · doi:10.1093/biomet/76.4.751

[28] Wang, Nonparametric estimation from cross-sectional survival data, Journal of the American Statistical Association 86 pp 130– (1991) · Zbl 0739.62026 · doi:10.1080/01621459.1991.10475011

[29] Wang, Hazards regression analysis for length-biased data, Biometrika 83 pp 343– (1996) · Zbl 0864.62080 · doi:10.1093/biomet/83.2.343

[30] Wang, Statistical models for prevalent cohort data, Biometrics 49 pp 1– (1993) · Zbl 0771.62079 · doi:10.2307/2532597

[31] Warren, Overview of the SEER-Medicare data: Content, research applications, and generalizability to the United States elderly population, Medical Care 40 pp 3– (2002) · doi:10.1097/00005650-200208001-00002