A NOTE ON PARAMETER ESTIMATION IN REGRESSION MODELS FOR CASE COHORT DATA

CHARLES UNIVERSITY PRAGUE
faculty of mathematics and physics

P. KLÁŠTERECKÝ & M. KULICH
Department of Probability and Statistics
Charles University, Prague
petr.klasterecky@matfyz.cz

INTRODUCTION

The main idea of the ***case-cohort design*** is to sample individuals from the study population (the full cohort) in order to reduce data collection costs, especially when there are relatively few events compared to the size of the cohort. Only the subjects experiencing the event (cases) and a subcohort consisting of the sampled individuals are followed during the study and have the covariate values recorded. Methods for analysing case-cohort data are usually modifications of the corresponding procedures for complete data which attempt to eliminate the unobserved information and to account properly for the sampling scheme. Most often the sampled controls' contributions are weighted in some way by the inverse selection probabilities.

For the survival time T, an unknown (vector) regression parameter β_0 and a vector of covariates Z, the ***linear transformation models*** assume

$$H_0(T) = -\beta_0' Z + \varepsilon, \qquad (1)$$

where H_0 is an unknown monotone transformation function such that $H_0(0) = -\infty$ and ε is a random variable with a known distribution, independent of Z. Specific models of this quite general class are obtained by choosing a particular distribution of ε, such as the extreme-value distribution for the proportional hazards model or the standard logistic distribution for the proportional odds model. We shall be interested in estimation of the regression parameters in the above model (1) applied to case–cohort data.

PARAMETER ESTIMATION

Using the standard counting process notation, Lu and Tsiatis [1] suggested the estimating equations

$$\sum_{i=1}^{n} \int_0^{\infty} Z_i \pi_i [dN_i(t) - Y_i(t) d\Lambda\{\beta' Z_i + H(t)\}] = 0 \quad \text{and} \quad \sum_{i=1}^{n} \pi_i [dN_i(t) - Y_i(t) d\Lambda\{\beta' Z_i + H(t)\}] = 0, \qquad (2)$$

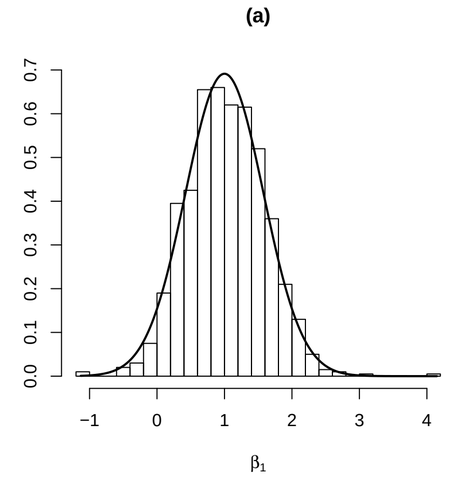
where the weights π_i are the inverse selection probabilities for each individual in the full cohort, i.e. $\pi_i = \delta_i + (1 - \delta_i)\xi_i/p$, where ξ_i is the subcohort indicator and $p = P(\xi_i = 1)$ is the subcohort sampling probability. The resulting estimator $\hat{H}$ of H_0 is a step function with jumps in the observed failure times. The estimator of β_0 is shown to be consistent and asymptotically normal in [1]. The computational algorithm we used in our simulations "shuttles" between estimation of H and β until convergence is reached.

SIMULATION STUDIES AND RESULTS

We generated failure times from the proportional odds model, i.e. the linear transformation model with hazard function of ε given by $\lambda(t) = \exp(t)/(1+\exp(t))$. Please see the paper for further details concerning the general simulation methodology.

In the first set of simulations, two independent covariates $Z_1 \sim \mathrm{U}(0,1)$ and $Z_2 \sim \mathrm{Alt}(0.35)$ were generated with regression parameters $\beta_1 = 1$ and $\beta_2 = -1$. Our results confirm good performance of the estimator reported by Lu and Tsiatis in this case. The parameter estimates are only slightly biased, the standard errors are well estimated, the confidence interval coverage is good, and the asymptotic normal distribution well approximates the empirical distribution of the simulated estimates (see Figures (a) and (b)).

In the second set of simulations, we considered three mutually correlated covariates: a dichotomous covariate $Z_1 \sim \mathrm{Alt}(0.35)$ and two continuous covariates Z_1 and Z_2 from truncated normal distributions with mean and variance depending on Z_1 and Z_1 and Z_2, respectively. The true parameter values were $\beta_1 = 2.3$, $\beta_2 = 0.7$, and $\beta_3 = 2.9$. This time, the estimator suffers from substantial biases, underestimated standard errors, and poor confidence interval coverage. The histograms of the simulated estimates (see Figures (c), (d) and (e)) reveal noticeable skewness towards the upper tail of the distribution for all the three parameter estimates. The performance of the estimator was generally worse when the subcohort sampling fraction was lower, regardless of the absolute subcohort size. However, the case-cohort design is most useful when a very large population yields just a few cases (rare diseases etc.), which is exactly the setting where the investigated estimator encounters serious performance problems.



SUMMARY

We studied an estimation procedure for case-cohort linear transformation models based on (2) and we found this method to be seriously biased when applied in practically relevant settings. Therefore we argue that a new estimation procedure should be developed.

Reference: [1] Lu, W. and Tsiatis, A.A., Semiparametric transformation models for the case-cohort study. *Biometrika*, to appear in 2006.
Acknowledgement: This work was partially supported by the grant GAČR 201/05/H007 and by the research project MSM 0021620839.