THE GRASS GROUP (Grup de Recerca en Anàlisi eStadística de la Supervivència)

Guadalupe Gómez
Departament d'Estadística i Investigació Operativa
Facultat de Matemàtiques
Universitat Politècnica de Catalunya

This group was formally born during a SEIO meeting in Sevilla in 1995. Its aim is to bring together researchers from Catalunya and abroad to collaborate in the theoretical developments, as well as in the applied methodologies, of the analysis of survival data. Even before its conception, the group had been actively collaborating with Professor Stephen W. Lagakos from the Biostatistics Department of the Harvard School of Public Health. During these years, our applied work, as well as some of the methods developed, have been based in problems proposed by clinicians and epidemiologists from the Institut Municipal de la Salut (IMS), from the Institut Municipal d'Investigacions Mèdiques (IMIM), from the Hospital Germans Trias i Pujol (Can Ruti), as well as from the AIDS Clinical Trial Group (ACTG) from Harvard University.

The main topics in our methodological research are the analysis of interval-censored data and the study of missing data problems. The first has been approached using frequentist and Bayesian methods, while the second has benefit from parametric as well as semiparametric developments. Among others, and as a consequence of the collaboration with the above mentioned Catalan groups, several papers have been published where survival analysis techniques have been applied to several complex scenarios such as the modelling of breast cancer survival as a function of the elapsed time between symptom and treatment (Gómez *et al.*, 1996), the study of the impact of the functional capacity into survival in the elderly population of Barcelona (Lamarca *et al.*, 1998) and the short term survival of individuals with tuberculosis who are infected with HIV (Falqués *et al.*, 1999).

Analysis in survival/sacrifice experiments: The analysis of the distribution of the time-to-tumor in experiments with rodents where several sacrifice times are scheduled has been investigated. The use of a four-variate counting process imbedded in their corresponding martingale framework is used to derive a new estimator and to study its asymptotic properties (Gómez and Julià, 1990; Gómez and Van Ryzin, 1992).

Properties for left-censored data: A theoretical study is undertaken for the asymptotic properties of the left-censored Kaplan-Meier estimator derived as the solution of a backward Dóleans equation (Gómez, Julià and Utzet, 1992; Gómez, Julià and Utzet, 1994).

Self-consistency approaches for interval-censored data: The self-consistency concept was first developed by Efron (1967) and it represented a different way of deriving the Kaplan-Meier estimator for the survival function. Turnbull used this same idea to estimate the distribution function from a sample containing interval-censored observations. Based on that, an algorithm is derived which is appropriate for other censoring mechanisms. We have developed this idea when interest lies in the elapsed latency time from an interval-censored origin to a right-censored end-point (double-censored data). The method has been applied to a cohort of hemophiliacs that became infected with HIV in the early 80's and subsequently developed AIDS (Gómez, 1992; Gómez and Lagakos, 1994; Gómez and Calle, 1999). These ideas are further developed and extended to estimate the parameters in a linear regression model when the covariate is interval-censored. The methodology is applied to the analysis of the viral load baseline history of an HIV+ group of patients that were subsequently randomized to six different therapies (Gómez, Espinal and Lagakos, 2001).

Nonparametric Bayesian analysis: We approach the estimation of the survival function based on interval-censored data from a nonparametric Bayesian point of view. We propose a methodology that accommodates the theory for right-censored data based on Dirichlet processes to an interval-censoring scheme by using Markov Chain Monte Carlo methods. We apply this methodology to analyze the risk of HIV infection in a cohort of injecting drug users in Barcelona. The nonparametric Bayesian approach allows, not only the incorporation of prior believes about the survival function, but also, the analysis of the risk of seroconversion without assuming restrictive parametric models. Furthermore, the estimator for the distribution function is smooth and thus differences between groups can be easily interpreted. (Gómez *et al.*, 2000; Calle and Gómez, 2001*a*). We extend these ideas to the analysis of regression models when one of the covariates is interval-censored. We encountered this situation in an AIDS clinical trial where the goal was to assess the association between duration of viral load suppression on a failed regimen and subsequent viral load. A completely parametric approach to this problem is, in general, not appropriate because it requires a model for the interval-censored covariate which cannot be validated. We propose the use of a mixture of Dirichlet processes. This mixture enables us to specify parametrically every component in the model except the distribution of the interval-censored covariate, which is treated nonparametrically. We develop in detail the proposed methodology for the linear regression model (Calle and Gómez, 2001*b*).

Survival analysis with missing covariates: In this topic the goal is to estimate the survival when part of the covariates of interest are missing. The main statistical problem is, on one hand, that any complete-case based inference is potentially biased and, on the other, in general it is not possible to assess the ignorability of the non-response mechanism, nor to test the assumptions on the distributions. Under a missing at random non-response pattern, we have developed some imputation techniques in order to complete the unobserved subsample and to apply a standard methodology. In the most general framework, we have adapted maximum likelihood based strategies to evaluate the impact of the model assumptions on the resulting estimates (Serrat *et al.*, 1998; Gómez and Serrat, 1999).

Semiparametric methods for missing covariates: We develop semiparametric strategies to deal with missing covariates in the context of a survival analysis study. We define a grouped Kaplan-Meier estimator as a discrete time version of the usual Kaplan-Meier estimator, when the covariates of interest are categorical and completely observed. We also developed inverse probability weighted generalized estimating equations. We use them to estimate the proportion of individuals at risk/censored/events in each one of the categories, in presence of missing data. The resulting estimator for the survival function is asymptotically unbiased and normal distributed. Its properties for finite samples have been also illustrated by simulation. To summarize the resulting inferences we propose a sensitivity analysis perspective on a rank of plausible values for the non-ignorability parameters in the non-response pattern. We have applied this methodology to the analysis of a cohort of pulmonary tuberculosis HIV-infected patients with a large proportion of missingness in the main predictors –CD4+ lymphocytes count and tuberculin skin test– (Serrat and Gómez, 2001*a*, 2001*b*).

References

1. Berger, A., Gómez, G. and Wallenstein, S. (1988). «A Homogeneity Test for Follow-Up Studies». *IMA Journal of Mathematics Applied in Medicine and Biology*, 5, 101-112.
2. Calle, M. L. and Gómez, G. (2001*a*). «Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods». *Journal of Statistical Planning and Inference*, 98, 73-87.
3. Calle, M. L. and Gómez, G. (2001*b*). «Semiparametric Bayesian analysis of regression models with an interval-censored covariate». *Technical Report, 2001/04, Department of Statistics and Operations Research.* Universitat Politècnica de Catalunya.
4. Falqués, M., Langohr, K., Gómez, G., Garcia de Olalla, P., Jansà, J. M. and Caylà, J. (1999). «Supervivencia en pacientes con tuberculosis infectados con VIH. Estudio de los fallecimientos en los primeros nueve meses de tratamiento». *Revista Española de Salud Pública*, 73, 549-562.
5. Gómez, G. (1992). «Estimation of Induction Distribution with Doubly Censored Data and Application to AIDS». *Teoría de la Probabilidad y sus Aplicaciones*, 37, 36-45 (published version is in russian).
6. Gómez, G. and Calle, M. L. (1999). «Nonparametric Estimation with Doubly Censored Data». *Journal of Applied Statistics*, 26, 45-58.
7. Gómez, G., Calle, M. L., Egea, J. M. and Muga, R. (2000). «Estimation of the risk of HIV infection as a function of the length of intravenous drug use. A nonparametric Bayesian approach». *Statistics in Medicine*, 19, 2641-2656.

8. Gómez, G., Espinal, A. and Lagakos, S. (2001). «Inference for a linear regression model with an interval-censored covariate». *Technical Report, 2000/18, Department of Statistics and Operations Research.* Universitat Politècnica de Catalunya.
9. Gómez, G. and Julià, O. (1990). «Estimation and Asymptotic Properties of the Distribution of Time-to-Tumor in Carcinogenesis Experiments». *IMA Journal of Mathematics Applied in Medicine and Biology*, 7, 109-123.
10. Gómez, G., Julià, O. and Utzet, F. (1992). «Survival Analysis for Left-Censored Data in Survival Analysis: State of the Art (editors: J. P. Klein and P. Goel): 269-288, Kluwer Academic Publishers, Dordrecht, Holanda.
11. Gómez, G., Julià, O. and Utzet, F. (1994). «Asymptotic Properties of the Left Kaplan-Meier Estimator». *Communications in Statistics: Theory and Methods*, 23, 123-135.
12. Gómez, G. and Lagakos, S. (1994). «Estimation of the Infection Time and Latency Distribution of AIDS with Doubly Censored Data». *Biometrics*, 50, 204-212.
13. Gómez., G., Porta, M., Griful, E., Maguire, A., Calle, M. L., Malats, N., Fernández, E., Piñol, J. L. and Gallén, M. (1996). «Modelling breast cancer survival and the symptom-to-treatment interval». *Journal of Epidemiology and Biostatistics*, 1, 175-182.
14. Gómez, G. and Serrat, C. (1999). «Estudios de supervivencia con datos no observados. Dificultades inherentes al enfoque paramétrico». *Qüestiió*, 23, 365-392.
15. Gómez, G. and Van Ryzin, J. (1992). «Estimation of the Subsurvival Function for Time-to-Tumor in Survival/Sacrifice Experiments». *Statistics and Probability Letters*, 13, 5-13.
16. Lamarca, R., Alonso, J., Gómez, G. and Muñoz, A. (1998). «Left-Truncated Data with Age as Time Scale: An Alternative for Survival Analysis in the Elderly». *Journal of Gerontology: Medical Sciences*, 53A, M337-M343.
17. Serrat, C. and Gómez, G. (2001*a*). «Estimating the stratified survival with missing covariates: I. A semiparametric approach». *Technical Report, 2001/12, Department of Statistics and Operations Research.* Universitat Politècnica de Catalunya.
18. Serrat, C. and Gómez, G. (2001*b*). «Estimating the stratified survival with missing covariates: II. A simulation study». *Technical Report, 2001/13, Department of Statistics and Operations Research.* Universitat Politècnica de Catalunya.
19. Serrat, C., Gómez, G., García, P. and Caylà, J. (1998). «CD4+ lymphocytes and tuberculin skin test as survival predictors in pulmonary tuberculosis HIV-infected patients». *International Journal of Epidemiology*, 27, 703-712.