A Mixture Model for Nuptiality Data with Long-Term Survivors*

Paraskevi Peristera and **Gebrenegus Ghilagaber**

Department of Statistics, Stockholm University, Sweden

1 Introduction and Literature Review

The tacit assumption in the analysis of duration data with censored observations is that censoring time is independent of event time. This, in turn, implies that the individuals who have not experienced the event of interest by the end of the study are a representative random sample of the population under investigation and do not differ in any systematic manner from those who have experienced the event. While this strong assumption may be valid for some types of events it is often violated in many other situations. For instance, in the analysis of data on family formation, individuals with a tendency to remain single over long periods (including their entire life) may be overrepresented among the censored. observations. Such individuals are known as long-term survivors and implementing standard survival techniques on data with long-term survivors may distort the results of the analysis.

The main approach for handling long-term survivors for failure time data is to suppose that there exist a latent subpopulation which can be considered a priori to have a zero risk of experiencing the event throughout the observation period. Empirical evidence of the existence of a surviving fraction would be heavy censoring at the end of the study period.

*Submitted for presentation at the Annual Meeting of Population Association of America (PAA): Boston, MA, 1-3 May 2014.

Mixture models that allow joint estimation of the probability of long-term survivorship and the timing of event occurrence are proposed in the literature (Maller and Zhou, 1996; Li and Choe, 1997; Brown and Ibrahim, 2003; Steele,2003; Shao and Zhou, 2004;Muthen and Masyn; 2005; Yu and Peng, 2008) for handling data with long-term survivors. Usually a logistic regression model for the event occurrence is combined with an event history model for event timing conditional on event occurrence. The advantage of mixture models is that they allow separate estimation of the effects of covariates on long-term survivorship and on event timing. This is important since the factors affecting the two processes may differ and factors that affect both processes may operate in different ways

2 Mixture Modeling

2.1 The Model

The mixture model assumes two latent subpopulations: one population with zero risk of experiencing the event (long-term survivors) and the other population with a non-zero risk (susceptible group) (Maller and Zhou, 1996). Define a binary variable Y, where $Y = 0$ indicates that an individual will never experience the event (i.e. will be a long-term survivor); $Y = 1$ indicates that the individual will eventually experience the event (susceptible individual i.e. not long-term survivor). Let T be a random variable that denotes the failure of interest, defined only when $Y = 1$. Then let $f(t)$, $h(t)$be the conditional probability density and hazard distribution functions of T, given that the event occurs ($Y = 1$). Also, let $g(y)$ is the unconditional probability function. We further assume a non-zero survival fraction given a covariate vector z

$$p(z) = P(Y = 1; z)$$

The survival function that corresponds to $g(t)$ can be expressed in terms of the mixture of susceptible and non-susceptible individuals as follows:

$$S_g(t/x; z) = (1 - p(z))S_L(t) + p(z)S_f(t)$$

where $S_L(t)$and $S_f(t) = S(t/Y = 1; x) = Pr(T > t/Y = 1; x)$ refer to the distributions of long-term survivors and susceptible respectively. Since those that are long-term survivors will never experience the event of interest then

when the $limt- > \infty : \ S_L(t) = 1$.Thus the unconditional survival function becomes

$$S_g(t/x; z) = 1 - p(z) + p(z)S(t/Y = 1; x)$$

The effect of the covariates z on the probability of being susceptible $p(z)$ can be modeled through a binary logistic regression model

$$p(z) = P(Y = 1/z) = \frac{exp(bz)}{1 + exp(bz)}$$

The conditional latency distribution $S_f(t) = S(t/Y = 1; x)$ can take the form of parametric or semiparametric distributions. Among the parametric models exponential, weibull, gompertz are commonly used to model survival data.

2.2 The Likelihood

Suppose that the data are of the form $(t_i; \delta_i; x_i; z_i)$ where δ_i is the censoring indicator with $\delta_i = 1$ if t_i is uncensored and $\delta_i = 0$ otherwise, x_i, z_i correspond the set of covariates for the incidence and survival part of the model.The likelihood contribution for individual i is:

$$\begin{aligned} p_i f(t_i/Y &= 1; xi) \ for \ \delta_i = 1 \\ &\& \\ (1 - p_i)S(t_i/Y &= 1; x_i) \ for \ \delta_i = 0 \end{aligned}$$

The observed marginal likelihood is then given by:

$$L(b) = \prod_i [p_i(z_i)f(t_i/Y = 1; x_i)]^{\delta_i}[(1 - p_i(z_i)) + p_i(z_i)S(t_i/Y = 1; x_i)]^{1-\delta_i}$$

The parameters of the model as well as the fraction of long-term survivors can be obtained through maximization of the marginal likelihood.

2.3 Scope of Analysis and Expected Outcomes

In this work we use a mixture model where parameters of a binary logistic regression model (for the conditional probability of long-term survivor given censoring) are jointly estimated with those of a continuous intensity model for family formation. The advantage of this model is that it allows simultaneous

estimation of two sets of effects of covariates: one for the probability of the event and another for the timing of the event. The model also allows for the incorporation of a frailty-term for unobserved heterogeneity. We illustrate this model in the analysis of nuptiality data. We aim to provide a comparative analysis between standard survival models, mixture models that account for long-term survivors and mixture models with frailty terms. We are interested in examining how the effect of different covariates changes for the different models and especially in the case of the mixture model. Preliminary results show that failure to account for long-term survivors may yield misleading results that could plague the purpose of the analysis.

2.4 References

Brown,E.R, and Ibrahim, J. G. (2003). A Bayesian semi-parametric joint hierarchical model for longitudinal and survival data. Biometrics, 59, 686-693

Farewell VT (1982).The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 38, 1041-1046

Li, L. and Choe, M. K. (1997). A mixture model for duration data: analysis of second births in China. Demography, 34, 189-197

Maller, R.A. and Zhou, X. (1996). Survival analysis with long-term survivors. Chichester: John Wiley and Sons

Muthen, B. and Masyn, K. (2005). Discrete-time survival mixture analysis. Journal of Educational and Behavioral Statistics, 30, 27-58

Shao, Q. and Zhou, X. (2004). A new parametric model for survival data with long-term survivors. Statistics in Medicine, 23, 3535-3543

Steel,F. (2003). A discrete-time multilevel mixture model for event history data with long-term survivors, with an application to an analysis of contraceptive sterilization in Bangladesh.

Yu, B. and Peng, Y. (2008). Mixture cure models for multivariate survival data. Computational Statistics & Data Analysis, 52, 1524-1532