Continuous - Time Survival Analysis in Mplus

Tihomir Asparouhov
Muthen & Muthen

February 9, 2006

1 Overview

Here we will describe the basic continuous time survival model implemented in Mplus and will provide some details on the basic modeling options that are available. Introduction to continuous time survival modeling can be found in Singer & Willett (2003), Hougaard (2000) or Klein & Moeschberger (1997). Survival analysis: techniques for. The survival models implemented in Mplus includes many extensions of this basic model such as mixture survival models, survival models with random effects (frailty models), multilevel survival models, time varying covariate models, competing risk models, non-proportional hazard models etc. Describing the details of these models is beyond the scope of this document. In most cases however the material presented here applies to these extensions as well. More details on the models and algorithms implemented in Mplus can be found in Larsen (2004, 2005) and Asparouhov & Muthen (2006).

Let the variable T_0 be a time-to-event variable such as time to death for example. Let C be the time when the individual leaves the target cohort due to death or other types of censoring such as lost to follow up etc. The survival variable T and the censoring indicator δ are defined by

$$T = \min\{T_0, C\} \tag{1}$$

$$\delta = \begin{cases} 1 & \text{if } T_0 > C \\ 0 & \text{if } T_0 \leq C \end{cases} . \tag{2}$$

Both variables T and δ have to be constructed and used in the survival analysis in Mplus. The T variable is specified via the *survival=* command while the δ variable is specified via the *timecensored=* command. Details on the specification options can be found in Muthen & Muthen (2006). Let X be an observed predictor of T.

2 The Proportional Hazard Model

The proportional hazard (PH) model specifies that the hazard function is proportional to the baseline hazard function, i.e.,

$$h(t) = h_0(t)Exp(\beta X) \tag{3}$$

where $h(t)$ is the hazard function and $h_0(t)$ is the baseline hazard function at time t. Two proportional hazard models are implemented in Mplus. One of the models assumes a completely non-parametric shape for the baseline hazard function. This model is known as the Cox regression model. The other model is based on a parametric model for the baseline hazard function. This model is known as the parametric PH model. The general parametric model for the baseline hazard function in Mplus is a step function with arbitrary number of steps, however through parameter constraints this parametric model can serve as an approximation to any other parametric model, including models such as Exponential, Weibull and Gompertz models. This approximation is based on the fact that any continuous function can be closely approximated by a step function. Note also that because of the parameter constraints the number of parameters that are freely estimated in the approximation model will remain the same, see Section 7 for a detailed example. First we will describe the parametric PH model implementation.

2.1 Parametric PH model

To estimate $h_0(t)$ as a step function with L intervals the survival variable is declared as *survival=T (L interval lengths)*. For example if *survival=T(2*1 2)*, the length of the intervals in the baseline hazard step function are 1, 1, 2 and ∞ in that order, i.e., the intervals in the definition of the baseline hazard function are $[0,1),[1,2),[2,4),[4,\infty)$ over which we assume that the baseline hazard is constant

$$h_0(t) = \begin{cases} h_1 & \text{if } 0 < t \leq 1 \\ h_2 & \text{if } 1 < t \leq 2 \\ h_3 & \text{if } 2 < t \leq 4 \\ h_4 & \text{if } 4 < t < \infty \end{cases} . \tag{4}$$

The analysis command option *basehazard* determines how the parameters h_1, h_2, ..., h_L are treated. If *basehazard=on* these parameters are estimated as regular parameters. Thus standard error will be computed for the baseline hazard function. Such standard errors can be used to obtain standard errors for the survival rates for example. They are also included as parameters in the model and can be held equal across class for example or they can be used in model constraint to impose certain parametric shape. Starting values can be given for these parameters and these starting values can be perturbed just as for other parameters. Acceptable starting values are between 0 and ∞, i.e., negative values are not acceptable baseline hazard function values. The *basehazard=on* option should be used if relatively few steps are used (small L) or there are enough restrictions in the model to compensate for a large number of steps. If L is large however, even with many restrictions on the h_i parameters, it may be difficult to estimate the model. The more parameters are in the model the more difficult the maximization will be, i.e., the estimation will be very computationally demanding. In addition to that the asymptotic approximation used with MLE requires larger sample size for models with larger number of parameters. Both the parameter estimates and the standard errors may have larger biases for models with larger number of parameters. These undesired effects can be avoided by specifying *basehazard=off* and in that case the parameters h_1, h_2, ..., h_L are treated as nuisance parameters. The profile likelihood is formed by explicitly maximizing the full likelihood over these parameters. The profile likelihood is then treated as regular maximum likelihood, see Murphy and van der Vaart (2000). Standard errors are not computed for the baseline hazard function, however the values of the nuisance parameters can be obtained by including *basehazard* option in the output command. In mixture models Mplus will estimate class varying baseline hazard and thus the mean of the survival variable will be unidentified. With *basehazard=off* the estimation will typically be less computationally demanding. The default setting for this option is *basehazard=off*.

2.2 Cox Regression Model

There are a number of different methods for estimating this model. The method that Mplus uses is based on PH parametric model estimation described in the previous section. To obtain a fully non-parametric baseline hazard function we just need to select sufficiently detailed step function estimation. This can be accomplished for example by settings such as *survival=T(500*0.02)* or *survival=T(1000*0.01)* if the T value ranges between 0 and 10. The exact specification of the step size typically will have a minimal effect on the estimates. The step size however affects the log-likelihood value. It is important that when LRT is conducted between two models the step function framework is the same. Mplus also implements an automatic option, *survival=T(all)*, which will construct the step intervals from the data, by making the steps as detailed as needed. With this option Mplus estimates a baseline hazard step function which is constant between every two consecutive event times. If all event times, including censored observations are $t_1 < t_2 < .... < t_n$ then Mplus estimates

$$h_0(t) = \begin{cases} h_1 & \text{if } 0 < t \leq t_1 \\ h_2 & \text{if } t_1 < t \leq t_2 \\ ... & \\ h_{n+1} & \text{if } t_n < t < \infty \end{cases} . \tag{5}$$

Equal event times are treated as one event time. There is a direct relation between the *survival=T(all)* specification of Cox regression and the *survival=T($M * h$)* specification with M large and h small. If h is smaller than the distance between any two distinct event times and Mh is greater than the biggest event time in the data, the parameter estimates and their standard errors will be the same. When estimating the Cox regression model the parameters h_i should be estimated as nuisance (unrestricted) parameters, i.e., with the settings *basehazard=off*. It is possible to estimate Cox regression with *basehazard=on* however this combination should be used with great care as the number of parameters may be too large.

For a discussion on the different ways to estimate the Cox regression model and the equivalence of the profile likelihood and the traditional partial likelihood methods see Clayton (1988).

3 The Cumulative Baseline Hazard Function

Suppose that the baseline hazard function is

$$h_0(t) = \begin{cases} h_1 & \text{if } t_0 = 0 < t \leq t_1 \\ h_2 & \text{if } t_1 < t \leq t_2 \\ ... & \\ h_{L+1} & \text{if } t_L < t < \infty \end{cases} . \tag{6}$$

The cumulative baseline hazard function at time t represents the total hazard an individual is exposed to up to time t. If $t_k < t < t_{k+1}$ the cumulative baseline hazard function is

$$H_0(t) = \int_0^t h_0(x)dx = \sum_{i=1}^{k-1} h_i(t_i - t_{i-1}) + h_k(t - t_{k-1}) \tag{7}$$

4 The Survival Function

The survival function is the probability that the survival variable T is greater than t

$$S(t) = P(T > t) = Exp(-Exp(\beta X)H_0(t)). \tag{8}$$

The survival function complements the distribution function

$$F(t) = P(T \leq t) = 1 - S(t). \tag{9}$$

5 The Likelihood Function

The likelihood function of the survival variable T is

$$L(T) = (h_0(T)Exp(\beta X))^{(1-\delta)}S(T) \tag{10}$$

where δ is the censoring variable.

6 Survival Variable in Monte Carlo Simulations

Survival variables can be used with Mplus simulation facilities. The step sizes m_1 ... m_L used for the generation process are specified in *generate = T (s* m_1 ... m_L*)*. The values of the baseline hazard function are specified in the Model Population section. These parameters are referred as $T\#1, ..., T\#L+1$ and should be specified in the Model Population section regardless of whether or not they are available in the Model section. These parameters are available in the Model section if the option *basehazard=ON*, however they are always available and should be specified in the Model Population section.

The simplest example for a survival variable specification is *generate=T(s)*. With this specification $L = 0$. The hazard function has a single step, which is the infinite interval $[0, \infty)$. In this case only one baseline hazard function value $T\#1$ has to be specified in the Model Population section. Thus the specification $T\#1 * \lambda$ defines a survival variable with constant hazard function λ over the entire $[0, \infty)$ interval. Suppose that there are no X variables in the model. In this case equation (3) says that $h(t) = h_0(t) = \lambda$. Equation (7) reduces to $H_0(t) = t\lambda$. Equations (8) and (9) then give us the distribution function of T

$$F(t) = P(T \leq t) = 1 - e^{-t\lambda}, \tag{11}$$

i.e., T is exponentially distributed with density function $\lambda e^{-t\lambda}$, mean $1/\lambda$ and variance $1/\lambda^2$, where $0 < \lambda < \infty$. This implies that the smaller the λ, the longer the survival time. Such considerations can be used for selecting proper values in the simulation study. For example if a predominant range of T between 0 and 30 is desired then the hazard should be set to $T\#1 * 0.1$. Using the distribution function of T in this case we get

$$P(0 < T < 10) = 1 - e^{-1} \approx 63\%$$

$$P(10 < T < 20) = e^{-1} - e^{-2} \approx 23\%$$
$$P(20 < T < 30) = e^{-2} - e^{-3} \approx 9\%$$
$$P(30 < T) = e^{-3} \approx 5\%.$$

7 Right Censoring of Survival Variables in Monte Carlo Simulations

The command *gentcensoring* = *T* (λ_1) specifies that the hazard for the censoring process is λ_1, i.e., an exponential variable C with mean $1/\lambda_1$ is generated as well as the uncensored survival variable T_0 following the survival variable specification. Censoring occurs if $C < T_0$. In that case we set $T = C$ and the censoring indicator δ to 1, i.e.,

$$T = \min\{T_0, C\} \tag{12}$$

$$\delta = \begin{cases} 1 & \text{if } T_0 > C \\ 0 & \text{if } T_0 \leq C \end{cases} . \tag{13}$$

Suppose that *gentcensoring* = *T* (λ_1) and the baseline hazard function is set to λ by setting *generate* = *T (s)* and within Model Population $T\#1 * \lambda$.

Then T_0 and C are independent exponentially distributed random variables with distribution $1 - e^{-\lambda t}$ and $1 - e^{-\lambda_1 t}$ respectively. In this case the variable T is also exponentially distributed with distribution function $1 - e^{-(\lambda+\lambda_1)t}$ because

$$P(T_0 > t) = e^{-\lambda t} \tag{14}$$

$$P(C > t) = e^{-\lambda_1 t} \tag{15}$$

$$P(T > t) = P(T_0 > t)P(C > t) = e^{-(\lambda+\lambda_1)t} \tag{16}$$

If $\lambda = \lambda_1$ about 50% of the observations will be censored because T_0 and C would be identically distributed and the two variables are equally likely to be the smallest.

8 Weibull PH Model Specification

The Weibull model assumes that the baseline hazard function is

$$h_0(t) = \lambda s(\lambda t)^{s-1}, \tag{17}$$

for some parameters λ and s, see Bradburn et alt. (2003). Below we describe how to set up an approximation for this model via the Mplus step function baseline model. The precision of the approximation depends on the number of intervals L used in the baseline step function. The more intervals are used the better the approximation. Typically however $L = 50$ will be sufficient. Suppose that most of the T values range from 0 to 5. We can split this range into equal intervals of length 0.1 and specify the baseline step function estimation by turning the option *basehazard=ON* and by setting *survival=T(50*0.1)*. With this setup however the baseline function will assume an unrestricted shape. We can add labels for the basehazard parameters by adding this line to the model $[T\#1 - T\#50](p1 - p50)$. Using these labels we specify the Weibull shape by adding the following *Model Constraint* section to the input file:

```
model constraint:
new (s   lambda);
p1=lambda*s*(lambda*0.05)**(s-1);
p2=lambda*s*(lambda*0.15)**(s-1);
...
p50=lambda*s*(lambda*4.95)**(s-1);
```

The value for t has been substituted with the midpoint for each of the time intervals. Take for example the first interval $[0, 0.1]$. To make the approximation as close as possible we use the midpoint of this interval 0.05 and we substitute that for t in equation (17). Better approximation can be accomplished by specifying smaller intervals for more dense time segments and larger intervals for time segments with fewer events. For example *survival=T(20*0.05 40*0.1)* will lead to better approximation if for many individuals $T < 1$. The LRT test can be used to test the model constraint equations, i.e., to test the assumption of Weibull baseline hazard.

9 References

Asparouhov, T. & Muthen, B. (2006) Continuous Time Survival Modeling with Latent Variables. In preparation.

Bradburn, M. J., Clark T.G., Love S.B., and Altman D.G. (2003), Survival Analysis Part II: Multivariate Data Analysis - an introduction to concepts and methods, British Journal of Cancer v. 89, 431-436.

Clayton, D. G. (1988), The analysis of event history data: A review of progress and outstanding problems, Statistics in Medicine, v. 7, 819-841.

Hougaard, P. (2000) Analysis of Multivariate Survival Data. Springer, New York.

Klein J.P. & Moeschberger, M.L. (1997) Survival analysis: techniques for censored and truncated data. New York: Springer.

Larsen, K. (2004) Joint Analysis of Time-to-Event and Multiple Binary Indicators of Latent Classes. Biometrics 60 (1), 85-92.

Larsen, K. (2005) The Cox Proportional Hazards Model with a Continuous Latent Variable Measured by Multiple Binary Indicators. Biometrics 61:4, 1049-1055

Murphy, S.A. & van der Vaart, A.W. (2000). On profile likelihood. Journal of the American Statistical Association. 95, 449-465.

Muthen, L.K. & Muthen, B.O. (1998-2006) Mplus User's Guide. Forth Edition. Los Angeles, CA: Muthen & Muthen.

Singer, J. D. & Willett, J. B. (2003) Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. New York: Oxford University Press.