Survival models in Stan

D.W. Bester

1 Model descriptions

Model 1

This is a survival model, where events happen according to the the hazard rate:

$$\begin{aligned} h(t) &= Be^{\theta t} \\ H(t) &= \int_0^t h(x)dx \\ &= \frac{B}{\theta}\left(e^{\theta t} - 1\right) \end{aligned}$$

The events are simulated using $H^{-1}(t)$, and then subjected to random (uninformative) censoring.

Model 2

This is the same as model 1, but we have multiple gompertz parameters. We try to use a `group` variable to assign the right gompertz parameters within Stan. Thus, this model fits the hazard rate

$$h(t) = B_j e^{\theta_j t} \tag{1}$$

where j is the group number of the subject in question. It is a covariate in our model.

Model 3

This is the same as model 2, but we add a "linear predictor":

$$\eta = \beta x \tag{2}$$

where x is a covariate. The hazard rate then becomes

$$\begin{aligned} h(t) &= B_j e^{\theta_j t} e^{\eta} \\ &= B_j e^{\theta_j t} e^{\beta x} \end{aligned}$$

where j is the group number of the subject in question. This is similar to the cox proportional hazards model. Notice that the linear predictor does not contain an intercept term. This is since the gompertz parameter B_j already acts as an intercept for that group. The β parameter measures the influence of of covariate x measured against the baseline. That is, we can rewrite the hazard rate as:

$$h(t) = e^{\theta_j t + \beta x + log(B_j)}$$

Thus, $log\,(B_j)$ is the intercept for group j, θ_j measures the influence of time on the hazard rate for group j, and β measures the influence of covariate x, assuming the influence is global over all groups.

2 Likelihood, estimation, and choice of prior parameters

All of our models assume that survival times come from a process with hazard rate $h(t)$, and we can use a Poisson process to show the likelihood contribution of each observation. Let T be the event-time of a subject in our model, then

$$T > t \quad \longleftrightarrow \quad N(t) < 1$$

where $N(t)$ is the number of events at time t, which will always be 0 or 1 in our model, then we have

$$N(t) \sim \text{Poisson}\left(\int_0^t h(x)dx\right).$$

We don't observe T_i directly, since our observations are subject to censoring. We observe $V_i = \min(T_i, C_i)$ where C_i is a censoring time, assumed to be random and independent of T (non-informative censoring). Our observations are the pairs (v_i, δ_i) where δ_i is an indicator taking 1 if $T_i \leq C_i$ and 0 otherwise. Thus, δ_i is an event indicator, and it will be equal to the number of events at time t.

The likelihood contribution of the observation pair (v_i, δ_i) is

$$\begin{aligned} \mathcal{L}\left(\theta|(v_i, \delta_i)\right) &= S(v_i)^{\delta_i} f(v_i)^{(1-\delta_i)} \\ &= h(v_i)^{\delta_i} \exp\left(-\int_0^{v_i} h(x)dx\right) \\ &= h(v_i)^{\delta_i} \exp\left(-H(v_i)\right) \end{aligned}$$

where θ is the set of all parameters. We will derive this likelihood in appendix A. This is the likelihood we need to add to MCMC software if we want to fit survival models.[1] The point of this repository is to show how this can done using Stan's `functions` block.

Stan was able to recover the chosen parameters for all of the above models. For model 3, the more complicated of the three models in terms of number of parameters, convergence was very sensitive to the choice of priors assumed on the Gompertz parameters, $\boldsymbol{B}$ and $\boldsymbol{\theta}$.

Add discussion here about the nuance of the prior parameters, how it effected the convergence of model 2 vs model 3, and how to avoid this problem in general by thinking carefully about prior parameters.

A Survival likelihood derivation

Let T, $T \geq 0$ be a continuous random variable representing a random future lifetime. Denote $F(\cdot)$ as the *cumulative distribution function*:

$$F(t) = P(T \leq t),$$

and $S(T)$ as the *survivor function*

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - F(t), \end{aligned} \tag{3}$$

We have $f(\cdot)$ as the probability density function (pdf), where

$$F(t) = \int_0^t f(x)dx,$$

which can also be expressed as

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\ &= \frac{d}{dt}F(t) = -\frac{d}{dt}S(t). \end{aligned} \tag{4}$$

[1] MCMC software like Stan and JAGS do provide Poisson densities, but we can only use them in special cases — if the hazard rate is constant or piecewise constant. For more complicated hazard rate functions, we have to resort to more tedious measures. Since the likelihood contains both $h(t)$ and $H(t)$, we need to add these time-varying functions to the sampler.

This has an intuitive meaning; the probability of an event happening at time t (or rather, in the infinitesimally small time period t to $t + dt$ is):

$$P(T = t) \approx dF(t) = f(t)dt. \tag{5}$$

For survival studies, it is useful to also define the *hazard rate*:

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left(t < T \leq t + \Delta t \mid T > t\right)}{\Delta t}.$$

We can also express this as

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{P(T > t)\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} \cdot \frac{1}{P(T > t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{6}$$

The hazard rate is conditioned on survival to time t, and should not be confused with the pdf, which is not conditioned. Whereas the approximate probability of an event happening at time t (or rather, in the infinitesimally small time period t to $t + dt$) is given by (5), the probability that a subject alive at time t will die at time t (or rather, within the infinitesimally small instant between time t to $t + dt$) is approximately:

$$h(t)dt. \tag{7}$$

The difference is subtle and lies in the conditionality: $h(t)$ is conditioned on the subject being alive at time t. Therefore, we say $h(t)$ is the mortality experienced by a subject alive at time t.

Each of the functions $f(t)$, $F(t)$, $S(t)$, or $h(t)$ uniquely defines the distribution of T (under some regularity conditions).[2] That is, by specifying one of these functions, you automatically specify the others, as all relationships (3), (4), (6) must hold. There is one final relationship that ties this all together. From (6) we have:

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{-\frac{d}{dt}S(t)}{S(t)} \quad \text{from (4)} \\ &= -\frac{d}{dt}\log(S(t)). \end{aligned}$$

[2] Add the note from Finkelstein's book that shows when $h(t)$ *doesn't* uniquely define the process T.

Integrating both sides:

$$\begin{aligned}\int_0^t h(x)dx &= -\int_0^t \left(\frac{d}{dx}\log(S(x))\right)dx \\ &= -\Big[\log(S(x))\Big]_{x=0}^{x=t} \\ &= -\big[\log(S(t)) - \log(1)\big] \\ &= -\log(S(t))\end{aligned}$$

finally giving

$$\exp\left(-\int_0^t h(x)dx\right) = \exp\big(-H(t)\big) = S(t).$$

Using earlier results, we also have

$$\begin{aligned}f(t) &= h(t)S(t) \qquad \text{from (6)} \\ &= h(t)\exp\big(-H(t)\big)\end{aligned}$$

Now that we have all the necessary definitions, we can derive the likelihood contribution for observation pairs (v, δ). The likelihood is defined as a function of the parameters of interest, which is proportional to the probability of the observations.

$$\begin{aligned}P(V = v, \delta = 1) &= P(Y = v, Y \leq C) \\ &= P(Y = v, v < C) \\ &= P(Y = v)P(v > C) \qquad \text{(independence)} \\ &\propto P(Y = v) \\ &= f(v)dv \\ &\propto f(v) \\ &= h(v)\exp\big(-H(v)\big)\end{aligned}$$

The term $P(v > C)$ is absorbed into the proportionality since it is non-informative; it contains none of the parameters of interest. The above method can also be used to construct likelihoods in cases where censoring is informative, such as competing risks models.

$$\begin{aligned}P(V = v, \delta = 0) &= P(C = v, Y > C) \\ &= P(C = v, Y > v) \\ &= P(C = v)P(Y > v) \qquad \text{(independence)} \\ &\propto P(Y > v) \\ &= S(v) \\ &= \exp\big(-H(v)\big)\end{aligned}$$

We can combine these two expressions to form the joint distribution of (Y, δ)

$$\begin{aligned} P(V = v, \delta = 0) &= \big(f(v)\big)^{\delta}\big(S(v)\big)^{1-\delta} \\ &= \Big(h(v)\exp\big(-H(v)\big)\Big)^{\delta}\Big(\exp\big(-H(v)\big)\Big)^{1-\delta_i} \\ &= h(v)^{\delta}\exp\big(-H(v)\big) \end{aligned}$$

For a set of $i = 1 \ldots N$ i.i.d observation pairs of (v_i, δ_i), the likelihood is

$$\begin{aligned} &\prod_{i=1}^{N}\Big(h(v_i)^{\delta_i}\exp\big(-H(v_i)\big)\Big) \\ &= \Big(\prod_{i=1}^{N}h(v_i)^{\delta_i}\Big)\exp\Big(-\sum_{i=1}^{N}H(v_i)\Big) \end{aligned}$$

with a log likelihood of

$$\begin{aligned} &\sum_{i=1}^{N}\Big(\delta_i\log\big(h(v_i)\big) - H(v_i)\Big) \\ &= \Big(\sum_{i=1}^{N}\delta_i\log\big(h(v_i)\big)\Big) - \Big(\sum_{i=1}^{N}H(v_i)\Big) \end{aligned} \tag{8}$$

This derivation is outlined in Tableman and Kim (2003) and a similar — but shorter — derivation appears in Klein and Moeschberger (2003, ch. 3.5). It is adequate for our purposes.

In section 2, however, we stated our distributional assumption in terms of the Poisson process.[3] Klein and Moeschberger (2003) explain that counting processes is an alternative way to develop inference techniques for censored and truncated data and it is more general than the derivation we used to arrive at (8). It requires a fair amount background, however, including stochastic integration, martingale theory, measure-theoretic probability, and counting processes, but the result is a general theory for event-time processes. Klein and Moeschberger (2003, ch. 3.6) walk through the necessary prerequisites to illustrate how (8) can be derived using counting processes. Although their derivation is certainly not devoid of rigour, they still insist that their argument is heuristic, and point to Chapter 2 of Andersen et al. (1993) for a precise proof. In this last text the reader can find all of the theory behind the proof of the Non-homogeneous Poisson Process likelihood — which regularly makes an appearance in survival analysis.

[3]Our model — where each subject can only have a single event that causes the subject to exit the study — is a special case of the Poisson process that allows multiple events per subject.

References

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Countin Processes*. Springer-Verlag. 6

Klein, J. and M. Moeschberger (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer. 6

Tableman, M. and J. Kim (2003). *Survival Analysis Using S: Analysis of Time-to-Event Data*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. 6