Bayesian approaches to survival modeling

Patrick Breheny

October 24

Introduction

- In today's lecture, we will see how survival analysis works from the Bayesian perspective, beginning with one-parameter models and continuing through multiparameter models and then looking at semiparametric modeling
- This shift in perspective is not as dramatic as it might first appear, in the sense that we have spent a great deal of time talking about likelihood, which is also an integral component of Bayesian analysis
- The notion of a prior, however, is unique to Bayesian analysis, and I will provide a quick overview

Bayesian inference: Main idea

The central idea of the Bayesian framework is that if we treat θ as a random variable, then

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)},$$

where

- $p(x|\theta)$ is the likelihood
- $p(\theta)$ is the *prior*: Our beliefs about the plausible values of our parameter before seeing any data
- $p(\theta|x)$ is the *posterior*: Our updated beliefs about the plausible values for our parameter after seeing the data
- $p(x)$ is a normalizing constant typically not of interest

Priors

- To carry out Bayesian inference, therefore, we need to specify both a prior as well as a likelihood
- Broadly speaking, there are two main ways of specifying priors:
  - *Informative priors* attempt to incorporate knowledge from other sources such as past studies in order to realistically capture one's state of knowledge about θ
  - *Reference priors* attempt to represent a vague, uninformed baseline, so that all conclusions will be based on the data alone, not from any external sources

Inference

- Once the model has been specified, all inference is based on the posterior $p(\theta|x)$
- For example, we can obtain point estimates via the posterior mean $\int \theta p(\theta|x)\,d\theta$ or posterior mode $\max_\theta p(\theta|x)$
- We can obtain 95% posterior intervals $[a, b]$ such that $\int_a^b p(\theta|x)\,d\theta = 0.95$
- We can calculate tail probabilities: $\mathbb{P}(\theta < 0) = \int_{-\infty}^0 p(\theta|x)\,d\theta$
- Note that with the Bayesian approach, no asymptotic arguments are required, although the integrals involved may be complicated, and thus, numerical integration methods are typically crucial to Bayesian methodology

Pike rat example

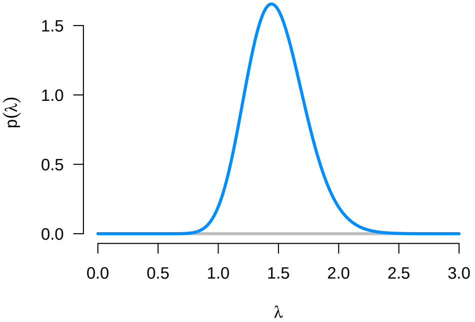
- To illustrate, let's analyze the Pike rat data using an exponential distribution
- Recall that in the frequentist version of this analysis,
  - The Score/Wald test of $H_0 : \lambda = 1$ yielded $p = 0.07$, while the LRT p-value was 0.04
  - However, the exponential fit isn't very good
- For the exponential distribution, $\lambda \sim \Gamma(\alpha, \beta)$ is a convenient (*conjugate*) prior, resulting in a closed form for the posterior:

$$\lambda | v \sim \Gamma(\alpha + d, \beta + v),$$

where $d = \sum_i d_i$ and $v = \sum_i t_i$

Bayesian approach: Reference prior

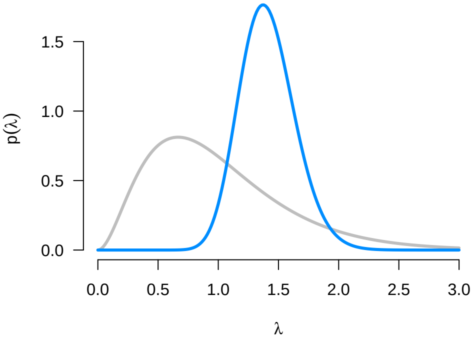
We will look at two potential prior distributions; first, an uninformative flat prior:



- $\mathbb{P}(\lambda < 1|d, v) = 0.014$
- 95% PI: (1.04, 2.00)

Bayesian approach: Gamma(3,3) prior

Suppose prior studies suggested that λ was likely between 0 and 2, and could reasonably be represented by a Gamma(3,3) distribution:



- $\mathbb{P}(\lambda < 1|d, v) = 0.028$
- 95% PI: (0.99, 1.87)

Nuisance parameters in the Bayesian setting

- As we have seen, nuisance parameters are a thorny problem in frequentist statistics, with several ways of addressing the issue (score, Wald, LRT, plus lots of others we didn't talk about)
- The Bayesian approach deals with nuisance parameters in a very different way
- Since inference is based on the posterior:

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta}),$$

to obtain the marginal posterior for θ_j, we simply integrate over the possible values of $\boldsymbol{\theta}_{-j}$:

$$f(\theta_j|\mathbf{x}) = \int f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}_{-j}$$

Monte Carlo integration

- In multi-parameter problems, we almost always rely on numerical integration
- This can be done in multiple ways, but the most common way is to generate random samples from the posterior (*Monte Carlo integration*); with such a sample,
    - Posterior means can be approximated by sample means
    - Posterior quantiles can be approximated by sample quantiles, etc.
    - Integrating over nuisance parameters can be approximated by simply looking at the marginal distribution of interest (one must still, of course, generate the random sample from the full posterior)
- Sounds nice...how are these random samples generated?

MCMC software

- The dominant method for generating such samples is via Markov chains (*Markov chain Monte Carlo*, or MCMC)
- A detailed discussion of MCMC methodology is beyond the scope of this course, but it involves generating new draws from conditional distributions $\theta^{(m+1)} \sim f(\text{Data}, \theta^{(m)})$ in such a way that the distribution of $\{\theta^{(m)}\}_{m=1}^{\infty}$ converges to the posterior distribution
- There are three commonly used programs for MCMC:
  - OpenBUGS (ancestor: WinBUGS)
  - JAGS (which we will be using)
  - STAN (impressive, but underdeveloped for survival . . . for now)

  all of which let the user specify the model and take care of the MCMC details for you

JAGS

- To install JAGS on a Windows machine:
  https://sourceforge.net/projects/mcmc-jags
  download and run the installer, clicking through to accept all the defaults (for install instructions on Linux/Mac, e-mail me)
- JAGS syntax is fairly intuitive; to implement fitting a gamma distribution to right-censored data with reference priors, the JAGS model specification would look like:

```
model {
  for (i in 1:n) {
    cens[i] ~ dinterval(t[i], tos[i]) # 1 if censored
    t[i] ~ dgamma(shape, rate)        # NA if censored
  }
  shape ~ dunif(0, 1000)
  rate ~ dunif(0, 1000)
}
```

rjags

JAGS can be run directly, but it's more convenient to run through its companion R package `rjags`:

```
library(rjags)
jagsData <- list(n = nrow(Data),
                 t = ifelse(Death==1, Time, NA),
                 tos = Time,
                 cens= 1-Death)
model <- jags.model(model_file,
                    data = jagsData,
                    n.chains = 4,
                    n.adapt = 1000)
post <- jags.samples(model, c('rate', 'shape'), 10000)
```

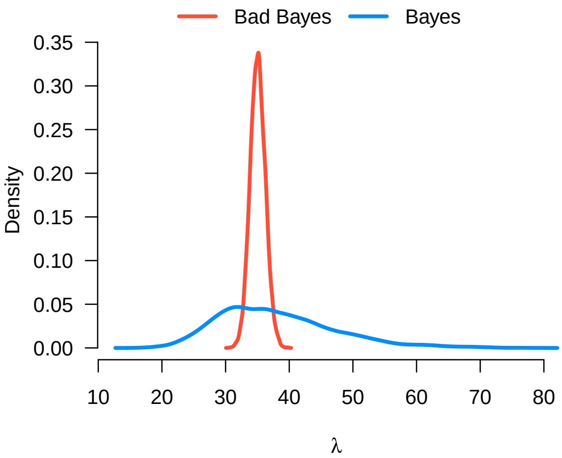
Pike rat data: Gamma model

- To illustrate, let's re-analyze the Pike rat data using a Gamma model (code given on previous two slides; it's online as well)
- Recall that in the frequentist version of this analysis,
  - The Gamma distribution was vastly superior to the exponential in terms of fitting the data
  - When carrying out inference for the rate parameter, taking into account uncertainty regarding the shape parameter was critical
- In the interest of time, I'm skipping some of the implementation details
  - I'm not going to go over every line of code, but all the code is provided online, with comments
  - Also, some additional code is provided for things like checking MCMC diagnostics (they all look fine)

Empirical Bayes

- A similar phenomenon happens in Bayesian inference
- Suppose that we simply replace α in the model with $\hat{\alpha}$ and treat $\hat{\alpha}$ as a constant (or, depending on your perspective, put a point prior with infinite strength on $\alpha = \hat{\alpha}$)
- This is known as an *empirical Bayes* approach
- Empirical Bayes certainly has its applications and can be a very useful statistical method, although this is an example of using it badly

Bayesian posterior for λ in the Pike rat study



Confidence/posterior intervals

	Nuisance parameters	
	Ignored	Accounted for
SE	1.2	8.4
Wald	(32.7, 37.4)	(18.6, 51.4)
Likelihood ratio	(32.7, 37.4)	(21.1, 54.1)
Bayes	(32.7, 37.4)	(23.6, 53.6)

Introduction

- Recall the proportional hazards model:

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{x}_i^T \boldsymbol{\beta}),$$

  where different choices of $\lambda_0(t)$ lead to different parametric models (exponential, Weibull, etc.)
- As we have discussed, however, parametric models often provide unsatisfactory fits to real data
- Our primary interest is in the regression coefficients; it would be unfortunate if misspecifying λ_0 led us to incorrect inference for $\boldsymbol{\beta}$, so in principle, we'd like to make as few assumptions about λ_0 as possible

Piecewise exponential

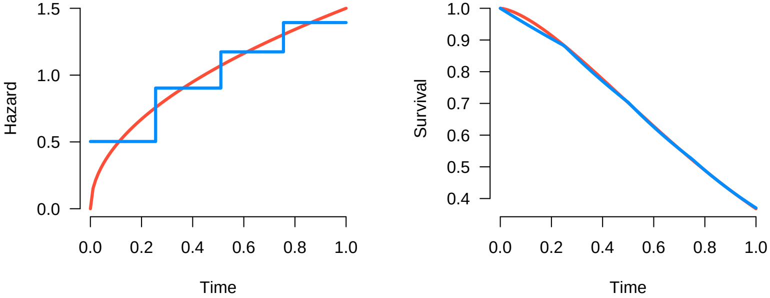
- Last time, we introduced one approach for doing so called Cox regression; today, we will examine a Bayesian model that behaves similarly
- As we saw earlier in the course with the Kaplan-Meier estimator, there is sometimes a fine line between "nonparametric" and "having a lot of parameters"
- With this in mind, let's consider modeling λ_0 as a piecewise constant function:

$$\lambda_0(t) = \lambda_j \text{ for all } t \in [a_{j-1}, a_j)$$

with $0 = a_0 < a_1 < \cdots < a_K$, where K denotes the total number of intervals; the resulting distribution could be thought of as piecewise exponential

Piecewise exponential: Hazard

Note that the hazard is piecewise constant, but the survival function is not



Equivalent Poisson

- OK, but piecewise constant hazard isn't exactly a standard distribution; how can we encode the equivalent of `t[i] ~ dgamma(shape, rate)`?
- To do so, we can use a clever rearrangement of the data such that its likelihood matches that of a Poisson distribution
- Let N_{ij} indicate whether subject i failed in interval j:

$$N_{ij} = 1\{t_i \in (a_{j-1}, a_j) \text{ and } d_i = 1\};$$

  in what follows, I will assume that the cutpoints $\{a_j\}$ are chosen such that $a_j \neq t_i \, \forall \, i, j$; cutpoints can always be chosen in this way

Equivalent Poisson (cont'd)

- Subject i's contribution to the likelihood can then be written

$$L_i = \prod_{j=1}^{K} (e^{\eta_i}\lambda_j)^{N_{ij}} \exp\{-e^{\eta_i} H_{ij}\lambda_j\},$$

  where

$$H_{ij} = \begin{cases} \min(t_i, a_j) - a_{j-1} & \text{if } t_i > a_j \\ 0 & \text{if } t_i < a_j \end{cases}$$

- This looks quite similar to the Poisson likelihood with rate parameter $\theta_{ij} = e^{\eta_i} H_{ij}\lambda_j$

Equivalent Poisson (cont'd)

- Indeed, the ratio between the two likelihoods is H_{ij*}, where $j*$ is the interval such that $N_{ij*} = 1$ (the two are identical for a censored observation)
- Since H_{ij} does not involve any parameters, the likelihoods are therefore proportional and sampling from one posterior is equivalent to sampling from the other
- Two technical notes:
  - This argument doesn't hold if $t_i = a_j$; the ratio would be $1/0$
  - There is a limit to the number of intervals we can choose: if there are no events in two adjacent intervals j and $j+1$, then λ_j and λ_{j+1} are not identifiable

  In practice, then, it is usually wise to select cut points from values in between the unique failure times

Piecewise exponential model specification

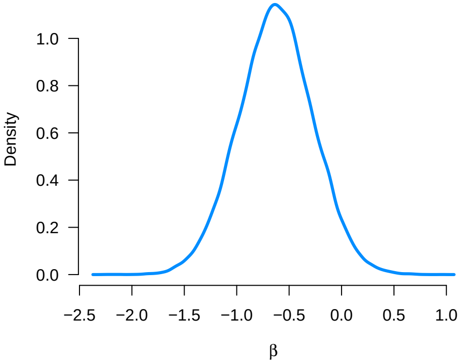
- To illustrate this model in action, let's apply it to the Pike rat data with a single predictor, `Group`
- We will use the reference priors:

$$\beta \sim \mathrm{N}(0, \tau^2)$$
$$\lambda_j \sim \Gamma(\alpha, \beta)$$

  where τ^2 is very large and α, β very small
- To begin, we will just set K (the number of pieces in our piecewise model) as large as possible (the number of unique failure times); we will then explore what our results look like if we lower K

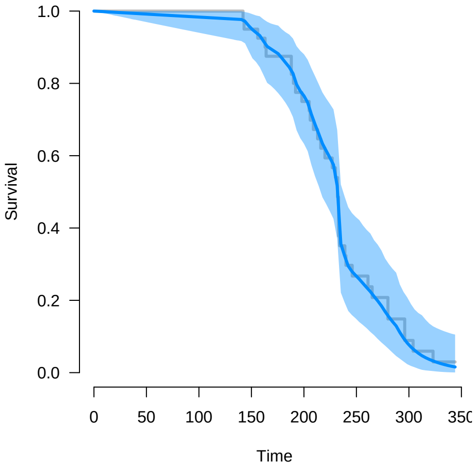
Results: β

Posterior density of β:

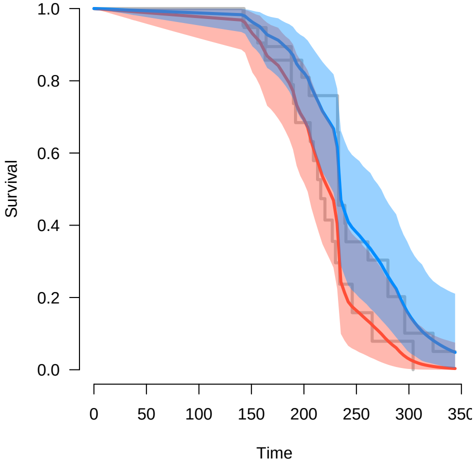


- PM: -0.63
- 95% PI: (-1.31, 0.08)

Results: Baseline survival



Results: Survival for each group

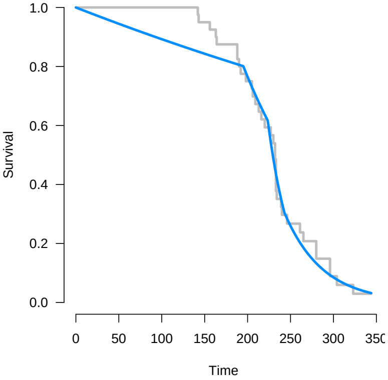


Nuisance parameters

- These confidence intervals are a nice illustration of the advantages Bayesian inference offers with respect to handling nuisance parameters
- As we will discuss in a future lecture, it is possible to go back and estimate the baseline survival in a Cox model
- It is also possible to calculate confidence intervals for the baseline survival
- However, there is *not* an easy way to calculate confidence intervals for the baseline survival in a way that takes into account uncertainty with regard to β

Changing K

Posterior mean of baseline survival with $K = 4$; the plot on the right attempts to choose the piecewise intervals a bit more intelligently given the low hazard over the first 140 days or so



Comparison of results for β

	Est	Lower	Upper
Exponential	-0.09	-0.75	0.56
Weibull	-0.72	-1.38	-0.07
Cox	-0.57	-1.25	0.11
BPE, K=29	-0.63	-1.31	0.08
BPE, K=4	-0.55	-1.24	0.13

Exponential/Weibull = Frequentist versions (`survreg`)
BPE = Bayesian piecewise exponential
Est = MLE / posterior mean
Lower/Upper = endpoints of 95% CI/PI