Accelerated Failure Time models (parametric survival regression)

toby@sg.cs.titech.ac.jp
Toby Dylan Hocking
30 Sept 2013

This is a class of models for

- several inputs $x_i \in \mathbb{R}^p$,
- using incomplete information about the outputs $y_i \in \mathbb{R}^+$, meaning that sometimes we know only an interval of possible values $y_i \in (\underline{y}_i, \overline{y}_i)$,
- assuming a distribution for the outputs such as $y_i \sim \text{LogNormal}(\mu_i, 1)$ or $y_i \sim \text{LogLogistic}(\alpha_i, 2)$ in the examples plotted below,
- where the real-valued scale parameter $\mu_i = w'x_i$ or $\log \alpha_i = w'x_i$ is assumed to be a linear combination of the inputs,
- and the center (mean or median) of the distribution is used as a predicted value $\hat{y} = f(x) \in \mathbb{R}^+$.

Maximum likelihood inference is then performed on the variable weights w. The only trick is that there is a non-standard likelihood function, since we have incomplete information for the outputs y_i:

- If we observe the complete output value y_i, as in the right panel in the plot below, then the likelihood function is the **density** $d(y_i, \hat{y})$, e.g. for the LogNormal.σ distribution,

$$d(y_i, \hat{y}) = \frac{1}{\hat{y}\sigma\sqrt{2\pi}} \exp\left(\frac{(\log y_i - \log \hat{y})^2}{-2\sigma^2}\right) \quad (1)$$

  (more functions on the back for those who are interested)

- If we observe only an interval $y_i \in (\underline{y}_i, \overline{y}_i)$, as in the left panel in the plot below, then the likelihood function is the **cumulative distribution function** $\int_{\underline{y}_i}^{\overline{y}_i} d(y, \hat{y})\, dy$, which is often available in closed form (it is not closed-form for the LogNormal, but we can still evaluate it numerically).

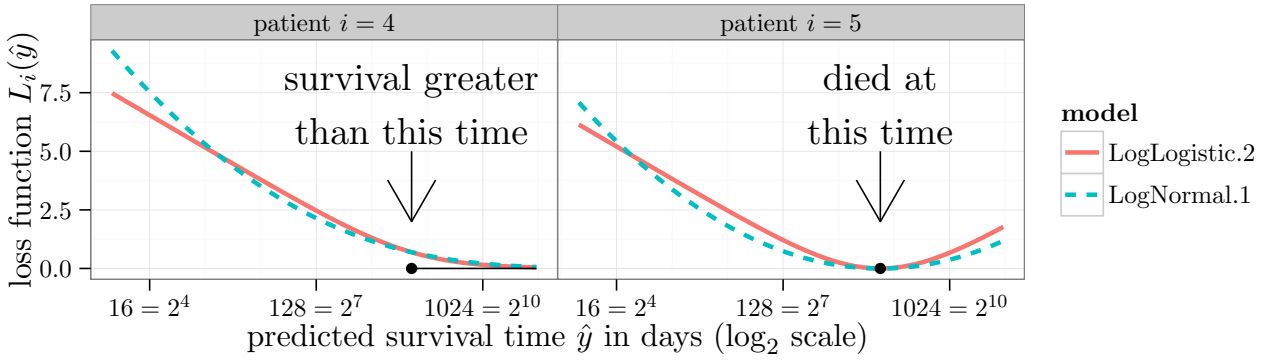
These models can be used for predicting

- Survival times of patients i that are treated for the same disease. Inputs x_i are age at diagnosis, blood pressure, tumor size, etc. Output y_i is the survival time after treatment. There are two cases at the time of the statistical analysis: (a) patient i has died, so we observe y_i, or (b) patient i is still alive, so we observe $y_i \in (t_i, \infty)$ where t_i is the time patient i has survived so far.
- Penalties for related segmentation problems i, such as $\lambda_i \in \mathbb{R}^+$ in $\min_z \text{Cost}_i(z) + \lambda_i \text{ Penalty}(z)$. Weak/incomplete labels about breakpoint locations do not indicate any single output λ_i value, but instead provide an interval $\lambda_i \in (\underline{\lambda}_i, \overline{\lambda}_i)$ of optimal output values, and I used features x_i of the segmentation problem such as number of points to segment and estimated variance. For more info see my ICML 2013 paper "Learning Sparse Penalties for Change-point Detection..."

Every concave maximum likelihood problem can also be thought of as a minimization of some convex loss function. In machine learning we focus on getting accurate predictions $\hat{y}$, so in the plot below we show the loss function in terms of the predicted values $L_i(\hat{y})$. Note that there is a nuisance/shape parameter, such as the standard deviation σ in (1), which affects the shape of the plotted loss function but does not affect the stationary point of the optimization with respect to w.

Which distribution is better, LogNormal or LogLogistic? There are other distributions (Weibull, any other positive-valued distribution, see back of this page), so how to choose which one is best? It depends on your data, so first define an evaluation metric such as the zero-one loss, and then use the distribution which gives the lowest prediction error on a held-out test set of data.

Free/open-source implementation: `survreg()` function in R package `survival`.



In survival analysis we have data (t_i, x_i, δ_i) for a set of patients $i \in \{1, \ldots, n\}$. Like in usual regression, $x_i \in \mathbb{R}^m$ is a vector of input variables. However, we observe a time $t_i \in \mathbb{R}^+$ and $\delta_i \in \{1 = \text{death}, 0 = \text{censor}\}$ which are related to the actual survival time $y_i \in \mathbb{R}^+$ as follows:

$$t_i = \begin{cases} \text{the amount of time patient } i \text{ lived} & \Rightarrow y_i = t_i \text{ if } \delta_i = 1 = \text{death} \\ \text{the amount of time until the study ended} & \Rightarrow y_i > t_i \text{ if } \delta_i = 0 = \text{censor.} \end{cases} \tag{2}$$

So if $\delta_i = 0 = \text{censor}$, all we know is that the survival of patient i is at least t_i. So the likelihood can be written as

$$\prod_{\delta_i=1=\text{death}} d(t_i) \prod_{\delta_i=0=\text{censor}} s(t_i) = \prod_{i=1}^{n} s(t_i) h(t_i)^{\delta_i} \tag{3}$$

where d is the density function, $s(t) = 1 - F(t)$ is the survival function, F is the cumulative distribution function, and h is the hazard function ($h = d/s$). Thus the log likelihood is

$$\sum_{i=1}^{n} \underbrace{\log s(t_i) + \delta_i \log h(t_i)}_{\log \text{lik}_i}$$

Distribution	$-\log \text{lik}_i$	Link
Exponential(λ_i)	$\lambda_i t_i - \delta_i \log \lambda_i$	$\lambda_i = \exp w' x_i$
Log-Logistic(α_i, β)	$(1+\delta_i) \log\left[1 + (t_i/\alpha_i)^\beta\right] - \delta_i \log(\beta t^{\beta-1} \alpha_i^{-\beta})$	$\alpha_i = \exp w' x_i$
Weibull(γ_i, k)	$(t_i/\gamma_i)^k - \delta_i \left[\log(k/\gamma_i) + (k-1)\log(t_i/\gamma_i)\right]$	$\gamma_i = \exp w' x_i$
Log-Normal(μ_i, σ)	$-\log s(t_i) = 1/2 + \text{erf}\left[\log(t_i/\mu_i)/(\sigma\sqrt{2})\right]$	$\mu_i = w' x_i$

Distribution	Prediction	$\hat{t}_i$	Surrogate loss $L(\delta, t, \hat{t})$
Exp(λ_i)	Mean	λ_i^{-1}	$t/\hat{t} + \delta \log \hat{t}$
Log-Logistic(α_i, β)	Median	α_i	$(1+\delta) \log\left[1 + (t/\hat{t})^\beta\right] - \delta \log(\beta t^{\beta-1} \hat{t}^{-\beta})$
Weibull(γ_i, k)	Mean	γ_i	$(t/\hat{t})^k - \delta \log(\hat{t}^{-1} k t^{k-1})$
Log-normal(μ_i, σ)	Mean	$\exp \mu_i$	$-\log s(t) = 1/2 + \text{erf}\left[\log(t/\hat{t})/(\sigma\sqrt{2})\right]$

Exercise for the reader: consider the more general case of real-valued interval output data, e.g. $y_1 = (-\infty, 4)$, $y_2 = (-3, \infty)$, $y_3 = (-1, 2)$, $y_4 = (5, 5)$. Instead of using one of these positive-valued distributions, consider an equivalent real-valued distribution (e.g. logistic rather than log-logistic). Write the maximum likelihood problem, derive the equivalent convex minimization problem, and write the surrogate loss $L(\underline{y}_i, \overline{y}_i, \hat{t})$. What is the gradient of the surrogate loss with respect to the optimization variables w?