Regularized AFT models

Anuj Khare

A standard AFT model is defined as follows:

$$\log(Y_i) = \beta_0 + x_i^T \beta + \sigma \epsilon_i \tag{1}$$

Where x_i are the covariates, Y_i is the observed time (output). ϵ_i F, where F is the distribution function.
Consider F to be the logistic cdf, given as:

$$F(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

The probability density function is given as:

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2} \tag{3}$$

Survival function can be calulated from the above:

$$1 - F(x) = \frac{1}{1 + e^{x}} \tag{4}$$

Here on, we assume that σ is constant for each observation i, and is ignored, and that the covariate matrix is appended with a column of ones. Hence,

$$\epsilon_i = \frac{\log(y_i) - (x_i^T \beta)}{\sigma} \tag{5}$$

For interval regression with censored data, we are given time intervals $\{\underline{t}_i, \bar{t}_i\}$ and covariates x_i for $i = 1 : n$, where $\underline{t}_i$ may be $-inf$ (left censoring) and $\bar{t}_i$ may be inf (right censoring).

Likelihood

For calculating likelihood, in the observations with no censoring, the pdf is used, and in censored observations, the cdf is used. The likelihood is given as:

$$L(\beta) = \prod_{i=1}^{n} \zeta(\beta, \underline{t}_i, \bar{t}_i) \tag{6}$$

where,

$$\zeta_i = \zeta(\beta, \underline{t}_i, \overline{t}_i) = \begin{cases} \zeta_{i,1} = F(R_i) & \text{if: } -\infty = \underline{t}_i \ , \overline{t}_i < \infty \\ \zeta_{i,2} = 1 - F(L_i) & \text{if: } -\infty < \underline{t}_i \ , \overline{t}_i = \infty \\ \zeta_{i,3} = F(R_i) - F(L_i) & \text{if: } -\infty < \underline{t}_i \neq \overline{t}_i < \infty \\ \zeta_{i,4} = d(R_i) & \text{if: } -\infty < \underline{t}_i = \overline{t}_i < \infty \end{cases} \tag{7}$$

where, $R_i = \log(\overline{t}_i) - \eta_i$, and $L_i = \log(\underline{t}_i) - \eta_i$. $\eta_i = x_i^T\beta$ is the vector of linear predictors. Clearly, for left censored observations, $\underline{t}_i = -inf$, hence we can substitute $L_i = 0$ in ζ_3 to get ζ_2. Simlarly, for right censored observations, $\overline{t}_i = inf$, hence we can substitute $R_i = inf$ in ζ_3 to get ζ_1.

Hence the log likelihood is given as:

$$l(\beta) = \sum_{i=1}^{n} \log \zeta_i = \sum_{i=1}^{n} \gamma_i \tag{8}$$

where, $\gamma_i = \log \zeta_i$.

Score

For now, we consider that $\underline{t}_i = \overline{t}_i$ is not possible. Hence we are left with the types of censoring.
We consider the parital derivative of $\gamma_{i,3}$ wrt to the parameters β. We can simply obtain derivatives for $\gamma_{i,1}$ and $\gamma_{i,2}$ by setting $L_i = 0$ and $R_i = inf$ respectively.
We use the following:

$$\frac{\partial F(R_i)}{\partial \beta_j} = -x_{ij} F(R_i)[1 - F(R_i)] \tag{9}$$

$$\frac{\partial \gamma_{i,3}}{\partial \beta_j} = -x_{ij} \frac{[F(R_i)(1 - F(R_i)) - F(L_i)(1 - F(L_i))]}{F(R_i) - F(L_i)} = x_{ij}[F(L_i) + F(R_i) - 1] \tag{10}$$

Define $\mu_i = (F(L_i) + F(R_i) - 1)$, $\mu = [\mu_1, ...\mu_n]^T$. Then, partial derivative of log-likelihood is given as:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}\mu_i \tag{11}$$

Hence, the score (gradient of llikelihood) is given as:

$$g = \nabla_\beta l(\beta) = X^T\mu = \sum_{i=1}^{n} \mu_i \overline{x}_i \tag{12}$$

Hessian

Hessian is the second derivative of loglikelihood:

$$H = \frac{\partial g(\beta)^T}{\partial \beta} = \sum_{i=1}^{n} (\nabla_\beta \mu_i) \overline{x}_i^T \tag{13}$$

Using 9, we have:

$$\frac{\partial \mu_i}{\partial \beta_j} = -x_{ij}[F(R_i)(1 - F(R_i)) + F(L_i)(1 - F(L_i))] \tag{14}$$

$$\nabla_\beta \mu_i = -\overline{x}_i[F(R_i)(1 - F(R_i)) + F(L_i)(1 - F(L_i))] \tag{15}$$

Hence, the hessian is given as:

$$H = \sum_{i=1}^{n} -w_i \overline{x}_i \overline{x}_i^T \tag{16}$$

where, w_i is given as:

$$\begin{aligned} w_i &= [F(R_i)(1 - F(R_i)) + F(L_i)(1 - F(L_i))] \\ &= -\mu_i[F(R_i) - F(L_i)] \end{aligned} \tag{17}$$

Define $W = diag(w_1, ...w_n)$.

$$H = -X^T W X \tag{18}$$

Clearly, the hessian is negative definite. Thus, the loglikelihood is strictly convex, and a unique global maximum exists.

IRLS

We use Newton's algorithm to find MLE for the AFT model, using negative loglikelihood (NLL). The Newton update is as follows:

$$\begin{aligned} \beta &= \widetilde{\beta} - H^{-1}\widetilde{g} \\ &= \widetilde{\beta} - (X^T \widetilde{W} X)^{-1} X^T \widetilde{\mu} \\ &= (X^T \widetilde{W} X)^{-1}((X^T \widetilde{W} X)\widetilde{\beta} - X^T \widetilde{\mu}) \\ &= (X^T \widetilde{W} X)^{-1} X^T (\widetilde{W} X \widetilde{\beta} - \widetilde{\mu}) \\ &= (X^T \widetilde{W} X)^{-1} X^T (\widetilde{W} X \widetilde{\beta} - \widetilde{\mu}) \end{aligned} \tag{19}$$

where, define the working response $\widetilde{z} = X\widetilde{\beta} - \widetilde{W}^{-1}\widetilde{\mu}$. Here, the tilde denotes that the respective values are evaluated using the parameters from the previous step.
Hence, at each step we are solving a weighted least squares problem, which is a minimizer of:

$$\sum_{i=1}^{n} \widetilde{w}_i(\widetilde{z}_i - \overline{x}_i^T\beta)^2 \tag{20}$$

We can rewrite z as:

$$\begin{aligned} \widetilde{z}_i &= x_i^T\widetilde{\beta} + \frac{\widetilde{\mu}_i}{\widetilde{w}_i} \\ &= x_i^T\widetilde{\beta} - \frac{1}{F(R_i) - F(L_i)} \end{aligned} \tag{21}$$

This algorithm is the iteratively reweighted least squares (IRLS), since at each iteration we solve a weighted least squares problem.

Elastic net penalty and coordinate descent

We define the elastic net (L1 + L2) penalty as follows:

$$\lambda P_\alpha(\beta) = \lambda(\alpha\|\beta\|_1 + 1/2(1-\alpha)\|\beta\|_2^2) \tag{22}$$

Adding the elastic net (L1 + L2) penalty, we get the following penalized weighted least squares objective:

$$M = \sum_{i=1}^{n} \widetilde{w}_i(\widetilde{z}_i - \overline{x}_i^T\beta)^2 + \lambda P_\alpha(\beta) \tag{23}$$

The subderivative of the optimization objective is given as:

$$\frac{\partial M}{\partial \beta_k} = \sum_{i=1}^{n} \widetilde{w}_i x_{ik}(\widetilde{z}_i - \overline{x}_i^T\beta) + \lambda\alpha \text{ sgn}(\beta_k) + \lambda(1-\alpha)\beta_k \tag{24}$$

where, sgn(β_k) is 1 if $\beta_k > 1$, -1 if $\beta_k < 0$ and 0 if $\beta_k = 0$.
Using the subderivative, three cases of solutions for β_k may be obtained. The solution is given by:

$$\hat{\beta}_k = \frac{S\left(\sum_{i=1}^{n} \widetilde{w}_i x_{ik}\left[\widetilde{z}_i - \sum_{j\neq k} x_{ij}\beta_j\right], \lambda\alpha\right)}{\sum_{i=1}^{p} \widetilde{w}_i x_{ik}^2 + \lambda(1-\alpha)} \tag{25}$$

where, S is the soft thresholding operator, and w_i and z_i are given in 17 and 21 respectively.
The coordinate descent algorithm works by cycling through each β_j in turn, keeping the others constant, and using the above estimate to calculate the optimal value $\hat{\beta}_j$. This is repeated until convergence.

Algorithm

As explained in the coxnet paper, the algorithm to be followed for fitting the distribution is:

1. Initialize $\widetilde{\beta}$.
2. Compute $\widetilde{w}_i$ and $\widetilde{z}_i$.
3. Find $\hat{\beta}$ by solving the penalized weighted least square problem defined in 23 using coordinate descent.
4. Set $\widetilde{\beta} = \hat{\beta}$
5. Repeat steps 2-4 until convergence of $\hat{\beta}$

Sources:

- Machine Learning: A Probabilistic Perspective - Kevin Murphy
- Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent - Simon, Friedman, Hastie, Tibshirani
- AFT - TD Hocking