Survival Analysis: Time To Event Modelling

Sudipta Das

Assistant Professor,
Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

Outline I

1. Parametric Estimation

2. Regression Models for Survival Data
   - Accelerated failure-time (AFT) model
     - Weibull Model
     - Log-Normal Model
     - Log-Logistic Model

Parametric Estimation I

- Parametric estimation of Type I censored data
- Data: $(t_1, \delta_1), \ldots, (t_n, \delta_n)$.
- Assumption: Samples are i.i.d.
- Likelihood:

$$L(\underline{\theta}) = \prod_{i=1}^{n} f(t_i|\underline{\theta})^{\delta_i} S(t_i|\underline{\theta})^{1-\delta_i}$$

- Example: Find the *m.l.e.* if the failure distribution is $Exp(\lambda)$.
- Likelihood

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \left[\lambda e^{-\lambda t_i}\right]^{\delta_i} \left[e^{-\lambda t_i}\right]^{1-\delta_i} \\
&= \lambda^{\sum_{i=1}^{n} \delta_i} e^{-\lambda \sum_{i=1}^{n} t_i} \\
&= \lambda^{k} e^{-\lambda \sum_{i=1}^{n} t_i},
\end{aligned}
$$

where k is number of failed items.
- Log likelihood

$$
l(\lambda) = k \log \lambda - \lambda \sum_{i=1}^{n} t_i
$$

- M.L.E.

$$
\hat{\lambda} = \frac{k}{\sum_{i=1}^{n} t_i}
$$

Parametric Estimation III

- Parametric estimation of Type II censored data
- Data: $t_{(1)}, \ldots, t_{(r)}$.
- Assumption: Samples are i.i.d.
- Likelihood:

$$L(\underline{\theta}) = \frac{n!}{(n-r)!}\left[\prod_{i=1}^{r} f(t_{(i)}|\underline{\theta})\right]\left[S(t_{(r)}|\underline{\theta})\right]^{n-r}$$

- Example: Find the *m.l.e.* if the failure distribution is $Exp(\lambda)$.
- Likelihood

$$
\begin{aligned}
L(\lambda) &\propto \left[\prod_{i=1}^{r} \lambda e^{-\lambda t_{(i)}}\right]\left[e^{-\lambda t_{(r)}}\right]^{n-r} \\
&= \lambda^r e^{-\lambda \sum_{i=1}^{r} t_{(i)}} e^{-\lambda(n-r)t_{(r)}} \\
&= \lambda^r e^{-\lambda\left[\sum_{i=1}^{r} t_{(i)}+(n-r)t_{(r)}\right]}
\end{aligned}
$$

- Log likelihood

$$
l(\lambda) = r\log\lambda - \lambda\left[\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}\right]
$$

- M.L.E.

$$
\hat{\lambda} = \frac{r}{\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}}
$$

Parametric Estimation V

- Example:- German banks credit (part) data

Duration	Amount	Installment Rate in %	Age	No. of Credits	No. of people Maintenance	$\cdots$ $\cdots$	Type
6	1169	4	67	2	1	$\vdots$	Good
48	5951	2	22	1	1	$\vdots$	Bad
12	2096	2	49	1	2	$\vdots$	Good
42	7882	2	45	1	2	$\vdots$	Good
24	4870	3	53	2	2	$\vdots$	Bad
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
45	1845	4	23	1	1	$\vdots$	Bad
45	4576	3	27	1	1	$\vdots$	Good

Parametric Estimation VI

- Recall the likelihood function for right censored data

$$L(\underline{\theta}) = \prod_{i=1}^{n} \left[f^{\delta_i}(T_i|\underline{\theta}) S^{1-\delta_i}(T_i|\underline{\theta}) \right]$$

- Maximum Likelihood Estimator of $\underline{\theta}$ is $\hat{\underline{\theta}}$
- Numerical solver
  - Good initial choice should be needed
  - Hint from non-parametric estimator function

Parametric Estimation VII

- Lifetime: $X \sim Gamma(\beta, \lambda)$
  - Probability density function: $f_X(t, \beta, \lambda) = \frac{\lambda(\lambda t)^{\beta-1}}{\Gamma(\beta)} e^{-\lambda t}$
  - Survival function: $S(t) = 1 - \frac{1}{\Gamma(\beta)}\gamma(\beta, \lambda t)$
    - $\gamma(\beta, \lambda t) = \int_0^{\lambda t} u^{\beta-1} e^{-u} du$
- Estimated parameter set: $\underline{\hat{\theta}} = [\hat{\beta}, \hat{\lambda}]'$, where

$$\hat{\beta} = \hat{a} = 3.4945 \text{ and } \hat{\lambda} = \frac{1}{\hat{s}} = 0.0896$$

  - Estimated survival function:

$$\hat{S}(t) = 1 - \frac{1}{\Gamma(\hat{\beta})}\gamma(\hat{\beta}, \hat{\lambda} t).$$

Parametric Estimation VIII

- Lifetime: $X \sim Weibull(\beta, \lambda)$
  - Probability density function: $f_X(t, \beta, \lambda) = \beta\lambda(\lambda t)^{\beta-1} e^{-(\lambda t)^{\beta}}$
  - Survival Function: $S(t) = e^{-(\lambda t)^{\beta}}$
- Estimated parameter set: $\underline{\hat{\theta}} = [\hat{\beta}, \hat{\lambda}]'$, where

$$\hat{\beta} = \hat{a} = 2.2836 \text{ and } \frac{1}{\hat{\lambda}} = \hat{s} = 42.4011$$

  - Estimated survival function:

$$\hat{S}(t) = e^{-(\hat{\lambda} t)^{\hat{\beta}}}.$$

- Lifetime: $X \sim LN(\mu, \sigma)$
  - Probability density function: $f_X(t, \mu, \sigma) = \frac{1}{t\sigma}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln t - \mu}{\sigma}\right)^2}\right]$,
  - Survival Function: $S(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right) = \Phi\left(\frac{-\ln t + \mu}{\sigma}\right)$,
    - where $\Phi(t) = \int_{-\infty}^{t} \phi(z)dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}z^2} dz$.
- Estimated parameter set: $\underline{\hat{\theta}} = [\hat{\mu}, \hat{\sigma}]'$, where

$$\hat{\mu} = 3.5605 \text{ and } \hat{\sigma} = 0.6456$$

  - Estimated survival function:

$$\hat{S}(t) = 1 - \Phi\left(\frac{\ln t - \hat{\mu}}{\hat{\sigma}}\right) = \Phi\left(\frac{-\ln t + \hat{\mu}}{\hat{\sigma}}\right).$$

    - *meanlog* means *mean of* $\log X$
    - *sdlog* means *standard deviation of* $\log X$

Parametric Estimation X

- Lifetime: $X \sim LL(\beta, \lambda)$
  - Probability density function: $f_X(t, \beta, \lambda) = \frac{\beta\lambda(\lambda t)^{\beta-1}}{[1+(\lambda t)^\beta]^2}$
  - Survival Function: $S(t) = \frac{1}{1+(\lambda t)^\beta}$
- Estimated parameter set: $\underline{\hat{\theta}} = [\hat{\beta}, \hat{\lambda}]'$, where

$$\hat{\beta} = \hat{a} = 2.7672 \text{ and } \frac{1}{\hat{\lambda}} = \hat{s} = 35.1290$$

  - Estimated survival function:

$$\hat{S}(t) = \frac{1}{1 + (\hat{\lambda} t)^{\hat{\beta}}}$$

Parametric Estimation XI

- Estimated survival functions

$$\hat{S}(t, \underline{\hat{\theta}})$$

- **FIGURE 4**

Parametric Regression: Introduction I

- Previously, we have stressed the importance of modeling the survival function, hazard function, or some other parameter associated with the failure-time distribution.
- Often a matter of greater interest is to ascertain the relationship between the failure time X and one or more of the explanatory variables.
- In most studies there are explanatory variables or covariates such as treatments, group indicators, individual characteristics, or environmental conditions, whose relationship to lifetime is of interest.
- This leads to a consideration of regression models.

Parametric Regression: Introduction II

- Consider a failure time $X > 0$, and a vector $Z^T = (Z_1, \ldots, Z_p)$ of explanatory variables associated with the failure time X.
- Covariate vector, Z^T may include
  - quantitative variables
    - such as blood pressure, temperature, age, and weight,
  - qualitative variables
    - such as gender, race, treatment, and disease status

  and/or
  - time-dependent variables, in which case $Z^T(x) = [Z_1(x), \ldots, Z_p(x)]$
    - Typical time-dependent variables include whether some intermediate event has or has not occurred by time x,
    - the amount of time which has passed since the same intermediate event,
    - serial measurements of covariates taken since a treatment commenced.

Parametric Regression: Introduction III

- Two approaches to the modeling of covariate effects on survival have become popular in the statistical literature
  - Accelerated failure-time (AFT) model
  - Cox's proportional hazard (PH) model

Accelerated failure-time (AFT) model I

- This approach is analogous to the classical linear regression approach.
- In this approach, the natural logarithm of the survival time $Y = \ln(X)$ is modeled.
  - This is the natural transformation made in linear models to convert positive variables to observations on the entire real line.

Accelerated failure-time (AFT) model II

- Accelerated failure time (AFT) regression model

$$\underbrace{\text{log of failure time for given covariate profile}}_{\log X|\mathbf{Z}}$$
$$= \underbrace{\text{location determined by covariates}}_{\gamma_0+\gamma^T\mathbf{Z}} + \underbrace{\text{scale}}_{\sigma} \times \underbrace{\text{error}}_{W}$$

i.e.

$$\begin{aligned} \log X|\mathbf{Z} &= \gamma_0 + \gamma^T\mathbf{Z} + \sigma W \\ \Rightarrow Y|\mathbf{Z} &= \mu + \sigma W \end{aligned}$$

  - Coefficient of regression $[\gamma_0, \gamma^T]^T$
  - Location parameter $[\gamma_0 + \gamma^T\mathbf{Z}]$
  - Scale parameter σ
- This is also called log-linear model

Accelerated failure-time (AFT) model III

- Common choices of error distribution W
  - Standard normal distribution for W yields
    - a Log-normal regression model,
  - Standard extreme value distribution for W yields
    - a Weibull regression model,
  - Logistic distribution for W yields
    - a Log-logistic regression model.

Accelerated failure-time (AFT) model IV

- Why is this model called the accelerated failure-time model?
  - Let $S_0(x)$ denote the survival function of $X = e^Y$ when $Z = 0$, i.e.,
    - $S_0(x)$ is the survival function of $e^{\gamma_0 + \sigma W}$.
- Now, the survival function for any covariate Z

$$\begin{aligned}
S(x|Z) &= Pr[X > x|Z] \\
&= Pr[Y > \ln x|Z] \\
&= Pr[\gamma_0 + \sigma W > \ln x - \gamma^T Z|Z] \\
&= Pr[e^{\gamma_0 + \sigma W} > xe^{-\gamma^T Z}|Z] \\
&= S_o[xe^{-\gamma^T Z}].
\end{aligned}$$

Accelerated failure-time (AFT) model V

- Notice that the effect of the explanatory variables in the original time scale is to change the time scale by a factor $\exp(-\gamma^T Z)$.
- Depending on the sign of $\gamma^T Z$, the time is either
  - Accelerated by a constant factor ($\gamma^T Z < 0$) or
    - Survival function decays at faster rate
    - $\log X$, equivalently X tends to be low
  - Degraded by a constant factor ($\gamma^T Z > 0$)
    - Survival function decays at slower rate
    - $\log X$, equivalently X tends to be high

Accelerated failure-time (AFT) model VI

- Let $h_0(x)$ be the baseline hazard at $Z = 0$, thus

$$h_o(x) = -\frac{d}{dx}\ln[S_0(x)]$$

- Let $h(x)$ be the arbitrary hazard thus

$$\begin{aligned}
h(x) &= -\frac{d}{dx}\ln[S(x)] \\
&= -\frac{d}{dx}\ln\left[S_0\left(xe^{-\gamma^T Z}\right)\right] \\
&= -\frac{d}{dxe^{-\gamma^T Z}}\ln\left[S_0\left(xe^{-\gamma^T Z}\right)\right] \times \frac{d}{dx}\left[xe^{-\gamma^T Z}\right] \\
&= h_0\left(xe^{-\gamma^T Z}\right)e^{-\gamma^T Z}
\end{aligned}$$

  - Notice the above relation as the hazard rate of an individual with a covariate value Z for this class of models is related to a baseline hazard rate h_0.

AFT: Weibull Model I

- AFT Weibull model

$$\begin{aligned} \log X|Z &= Y|Z \\ &= \gamma_0 + \gamma^T Z + \sigma W \end{aligned}$$

where $W \sim EV(0, 1)$.

- Thus, $Y \sim EV(\gamma_0 + \gamma^T Z, \sigma)$

AFT: Weibull Model II

- Therefore,

$$
\begin{aligned}
X|Z &= e^{\gamma_0+\gamma^T Z+\sigma W} \\
&= e^{\mu+\sigma W}
\end{aligned}
$$

- $X|Z$ follows Weibull distribution with
  - shape parameter: $\beta = \frac{1}{\sigma}$ and
  - scale parameter: $\lambda = e^{-\mu} = \frac{1}{e^{\gamma_0+\gamma^T \mathbf{z}}}$
- The survival function

$$
\begin{aligned}
S(x|Z) &= \exp\left[-(\lambda x)^{\beta}\right] \\
&= \exp\left[-\left(\frac{x}{e^{\gamma_0+\gamma^T \mathbf{z}}}\right)^{1/\sigma}\right]
\end{aligned}
$$

AFT: Weibull Model III

- Example: Bank credit data
- Construct the likelihood function for right-censored data

$$L = \prod_{j=1}^{n} \left[\frac{1}{\sigma} f_W \left(\frac{y_j - \mu}{\sigma} \right) \right]^{\delta_j} \left[S_w \left(\frac{y_j - \mu}{\sigma} \right) \right]^{1-\delta_j},$$

  where
  - $f_W(w) = e^{w-e^w}$ and $S_W(w) = e^{-e^w}$
- Find the maximum likelihood estimates of μ, (i.e., $[\gamma_0, \gamma^T]$) and σ, along with their standard errors.
- Thus, the estimates of β and λ are obtained.
- Then, one can estimate the survival function $S(x)$ along with its standard error.

AFT: Weibull Model IV

- Results
  - Using 5 covariates; (Age, Amount, InstallmentRatePercentage, NumberExistingCredits and NumberPeopleMaintenance)
    - $\hat{\gamma}_0 = 3.08$
    - $\hat{\gamma} = [2.80\times10^{-3}, 8.36\times10^{-5}, 4.56\times10^{-2}, -1.22\times10^{-2}, 1.88\times10^{-3}]^T$
    - $\hat{\sigma} = 0.357$
  - Using 2 covariates; (Amount and InstallmentRatePercentage)
    - $\hat{\gamma}_0 = 3.14$
    - $\hat{\gamma} = [8.45\times10^{-5}, 5.13\times10^{-2}]^T$
    - $\hat{\sigma} = 0.356$

AFT model: Log-Normal Model I

- AFT Log-Normal Model

$$\begin{aligned} \log X|Z &= Y|Z \\ &= \gamma_0 + \gamma^T Z + \sigma W \end{aligned}$$

$W \sim N(0, 1)$ [i.e., Standard normal]

- Thus, $Y \sim N(\gamma_0 + \gamma^T Z, \sigma)$

AFT model: Log-Normal Model II

- Therefore,

$$\begin{aligned} X|Z &= e^{\gamma_0+\gamma^T Z+\sigma W} \\ &= e^{\mu+\sigma W} \end{aligned}$$

- $X|Z$ follows Log-Normal distribution with
  - location parameter (mean of log X): $\mu = \gamma_0 + \gamma^T \mathbf{Z}$
  - scale parameter (sd of log X): σ
  - Mean (of X) is $e^{\left(\mu+\frac{\sigma}{2}\right)}$ and
  - Variance (of X) is $[e^{\sigma} - 1]\, e^{(2\mu+\sigma)}$.
- $S(x|Z) = 1 - \Phi\left[\frac{\log x - \gamma_0 - \gamma^T \mathbf{Z}}{\sigma}\right]$

AFT model: Log-Normal Model III

- Example: Bank credit data
- Construct the likelihood function for right-censored data

$$L = \prod_{j=1}^{n} \left[\frac{1}{\sigma} f_W \left(\frac{y_j - \mu}{\sigma} \right) \right]^{\delta_j} \left[S_W \left(\frac{y_j - \mu}{\sigma} \right) \right]^{1-\delta_j},$$

  where
  - $f_W(w) = \phi(w)$ and $S_W(w) = \Phi(-w)$
- Find the maximum likelihood estimates of μ, (i.e., $[\gamma_0, \gamma^T]$) and σ, along with their standard errors.
- Then, one can estimate the survival function $S(x)$ along with its standard error.

AFT model: Log-Normal Model IV

- Results
  - Using 5 covariates; (Age, Amount, InstallmentRatePercentage, NumberExistingCredits and NumberPeopleMaintenance)
    - $\hat{\gamma}_0 = 2.90$
    - $\hat{\gamma} = [1.66\times10^{-3}, 7.59\times10^{-5}, 6.38\times10^{-2}, 5.80\times10^{-2}, -2.75\times10^{-2}]^T$
    - $\hat{\sigma} = 0.552$
  - Using 2 covariates; (Amount and InstallmentRatePercentage)
    - $\hat{\gamma}_0 = 3.00$
    - $\hat{\gamma} = [7.71\times10^{-5}, 6.58\times10^{-2}]^T$
    - $\hat{\sigma} = 0.551$

AFT model: Log-Logistic Model I

- AFT Log-Logistic Model

$$\begin{aligned} \log X|Z &= Y|Z \\ &= \gamma_0 + \gamma^T Z + \sigma W, \end{aligned}$$

  W follows a standard logistic distribution.

- Thus $Y \sim \mathit{logistic}(\mu = \gamma_0 + \gamma^T Z, \sigma)$

AFT model: Log-Logistic Model II

- Therefore,

$$\begin{aligned} X|Z &= e^{\gamma_0+\gamma^T Z+\sigma W} \\ &= e^{\mu+\sigma W} \end{aligned}$$

  - $X|Z$ follows Log-Logistic distribution with
    - shape parameter: $\beta = \sigma^{-1}$ and
    - scale parameter: $\lambda = e^{-\mu}$
- The survival function

$$\begin{aligned} S(x|Z) &= \frac{1}{1+(\lambda x)^{\beta}} \\ &= \frac{1}{1+e^{-\frac{\gamma_0+\gamma^T Z}{\sigma}} x^{\frac{1}{\sigma}}} \end{aligned}$$

AFT model: Log-Logistic Model III

- Example: Bank credit data
- Construct the likelihood function for right-censored data

$$L = \prod_{j=1}^{n} \left[\frac{1}{\sigma} f_W \left(\frac{y_j - \mu}{\sigma} \right) \right]^{\delta_j} \left[S_w \left(\frac{y_j - \mu}{\sigma} \right) \right]^{1-\delta_j},$$

where

- $f_W(w) = \frac{e^w}{(1+e^w)^2}$ and $S_W(w) = \frac{1}{1+e^w}$

- Find the maximum likelihood estimates of μ, (i.e., $[\gamma_0, \gamma^T]$) and σ, along with their standard errors.
- Then, one can estimate the survival function $S(x)$ along with its standard error.

AFT model: Log-Logistic Model IV

- Results
  - Using 5 covariates; (Age, Amount, InstallmentRatePercentage, NumberExistingCredits and NumberPeopleMaintenance)
    - $\hat{\gamma}_0 = 2.88$
    - $\hat{\gamma} = [2.75\times10^{-3}, 8.34\times10^{-5}, 6.23\times10^{-2}, 1.99\times10^{-2}, -3.38\times10^{-2}]^T$
    - $\hat{\sigma} = 0.295$
  - Using 2 covariates; (Amount and InstallmentRatePercentage)
    - $\hat{\gamma}_0 = 2.95$
    - $\hat{\gamma} = [8.47\times10^{-5}, 6.62\times10^{-2}]^T$
    - $\hat{\sigma} = 0.294$