BSTT536: Survival Data Analysis

Instructor: Hua Yun Chen, PhD

Division of Epidemiology and Biostatistics
School of Public Health
University of Illinois at Chicago

Table of Content

Two-sample comparison by parametric models

Parametric regression model for survival time

Fit Weibull Model

Accelerated failure time Model

SAS Proc LifeReg

Comparison of Survival Functions in parametric model

In the myeloma data, we may want to compare the survival distributions for male and female.

1. For male (sex=1), the exponential model fit

$$\begin{aligned}\hat{\lambda}_m &= \frac{\sum_{i=1}^{n} \delta_i * (sex_i = 1)}{\sum_{i=1}^{n} X_i * (sex_i = 1)} = 0.0292 \\ \hat{V}(\hat{\lambda}_m) &= \frac{\sum_{i=1}^{n} \delta_i * (sex_i = 1)}{\{\sum_{i=1}^{n} X_i * (sex_i = 1)\}^2} = 3.88 \times 10^{-5}.\end{aligned}$$

2. For female (sex=2), the exponential model fit

$$\begin{aligned}\hat{\lambda}_f &= \frac{\sum_{i=1}^{n} \delta_i * (sex_i = 2)}{\sum_{i=1}^{n} X_i * (sex_i = 2)} = 0.0379 \\ \hat{V}(\hat{\lambda}_f) &= \frac{\sum_{i=1}^{n} \delta_i * (sex_i = 2)}{\{\sum_{i=1}^{n} X_i * (sex_i = 2)\}^2} = 1.03 \times 10^{-4}.\end{aligned}$$

3. Is the male survival distribution significantly different from the female survival distribution?

Compare Two Parametric Survival Distributions

1. Survival distribution for male

$$S_m(t) = \exp(-\lambda_m t)$$

2. Survival distribution for female

$$S_f(t) = \exp(-\lambda_f t) = \exp\left(-\gamma \lambda_m t\right),$$

   where $\gamma = \frac{\lambda_f}{\lambda_m}$.

3. We can test the hypothesis

$$H_0 : \gamma = 1 \text{ vs. } H_A : \gamma \neq 1.$$

Perform the Test

1. From the model fit

$$\hat{\gamma} = \frac{\hat{\lambda}_f}{\hat{\lambda}_m} = \frac{0.0379}{0.0292} = 1.298$$

2. *The variance for the $\hat{\gamma}$ can be obtained by δ-method

$$V(\hat{\gamma}) = \hat{\gamma}^2 \left[\frac{V(\hat{\lambda}_f)}{\hat{\lambda}_f^2} + \frac{V(\hat{\lambda}_m)}{\hat{\lambda}_m^2} \right].$$

3. The estimated variance

$$\hat{V}(\hat{\gamma}) = 1.298^2 \times \left[\frac{1.03 \times 10^{-4}}{0.0379^2} + \frac{3.88 \times 10^{-5}}{0.0292^2} \right] = 0.197$$

Perform the Test (continuing)

1. A test statistic for $H_0 : \gamma = 1$,

$$\frac{\hat{\gamma} - 1}{\sqrt{\hat{V}(\hat{\gamma})}} = \frac{0.298}{\sqrt{0.197}} = 0.671.$$

   p-value based on normal distribution is 0.251.

2. No significant difference between the two survival functions.

More on Two-sample Comparison under Exponential Model

1. γ can only take positive values. The model can be written as

$$\begin{aligned} S_m(t) &= \exp(-\lambda_m t) \\ S_f(t) &= \exp(-\lambda_f t) = \exp\left(-e^{\alpha}\lambda_m t\right), \end{aligned}$$

   where $\alpha = \log\gamma$ and α can take any values: positive or negative.

2. The two survival distribution can be combined into one form as

$$S(t|\text{gender}) = \exp\left(-e^{\alpha\times\text{gender}}\lambda_m t\right),$$

   where gender $= 0$ for male, and gender $= 1$ for female.

Conditional exponential model

1. In the myeloma dataset, aside from sex(or gender),age, blood urea nitrogen (BUN), serum calcium (CA), haemoglobin (HB), the percentage of plasma cells in the bone marrow (PC), and the presence or absence of Bence-Jones protein (BJ) can also potentially affect the survival.
2. Let Z denote the vector of covariates, the covariate effects can be modeled by

$$S(t|z) = \exp\left(-e^{\alpha z}\lambda_m t\right),$$

3. It means that, conditional on the covariate value z, the survival time is assumed to follow an exponential distribution with hazard $\lambda_m \exp(\alpha z)$.

Regression interpretation of the conditional exponential model

1. The conditional exponential model can be interpreted in terms of regression model as

$$\log T = \beta_0 + \beta Z + \epsilon,$$

where $\beta_0 = -\log \lambda_m$, $\beta = -\alpha$, and ϵ has a distribution function $1 - \exp\{-\exp(u)\}$.

2. To see that,

$$\begin{aligned} P(T > t \mid Z = z) &= P(\log T > \log t \mid Z = z) \\ &= P(\beta_0 + \beta Z + \epsilon > \log t \mid Z = z) \\ &= P(\epsilon > \log t - \beta_0 - \beta Z \mid Z = z) \\ &= \exp\{-\exp(\log t - \beta_0 - \beta z)\} \\ &= \exp\{-\exp(\alpha z)\lambda_m t\}, \end{aligned}$$

Generalization of the regression model

1. It is natural to add a scale parameter $\sigma > 0$ to the exponential regression model

$$\log T = \beta_0 + \beta Z + \sigma\epsilon,$$

   where ϵ has a distribution function $1 - \exp\{-\exp(u)\}$.

2. This results

$$\begin{aligned} P(T > t \mid Z = z) &= P(\log T > \log t \mid Z = z) \\ &= P(\epsilon > (\log t - \beta_0 - \beta Z)/\sigma \mid Z = z) \\ &= \exp[-\exp\{(\log t - \beta_0 - \beta z)/\sigma\}] \\ &= \exp[-\exp\{(-\beta/\sigma)z\}(\lambda_m t)^{1/\sigma}], \end{aligned}$$

Weibull Family of Distributions

1. When $\beta = 0$, the survival time distribution

$$P(T > t) = \exp\{-(\lambda t)^{1/\sigma}\},$$

   is called Weibull family of distributions. More often $b = 1/\sigma$ is used as the scale parameter in Weibull distribution.
2. Weibull distribution reduce to exponential distribution when $b = 1$.
3. When $b > 1$, compared with the exponential survival function, the weibull survival function decreases slower for $t < 1/\lambda$ and decreases faster for $t > 1/\lambda$.
4. When $b < 1$, compared with the exponential survival function, the weibull survival function decreases faster for $t < 1/\lambda$ and decreases slower for $t > 1/\lambda$.

Weibull Distribution for $\lambda = 1$ and $b = 2$

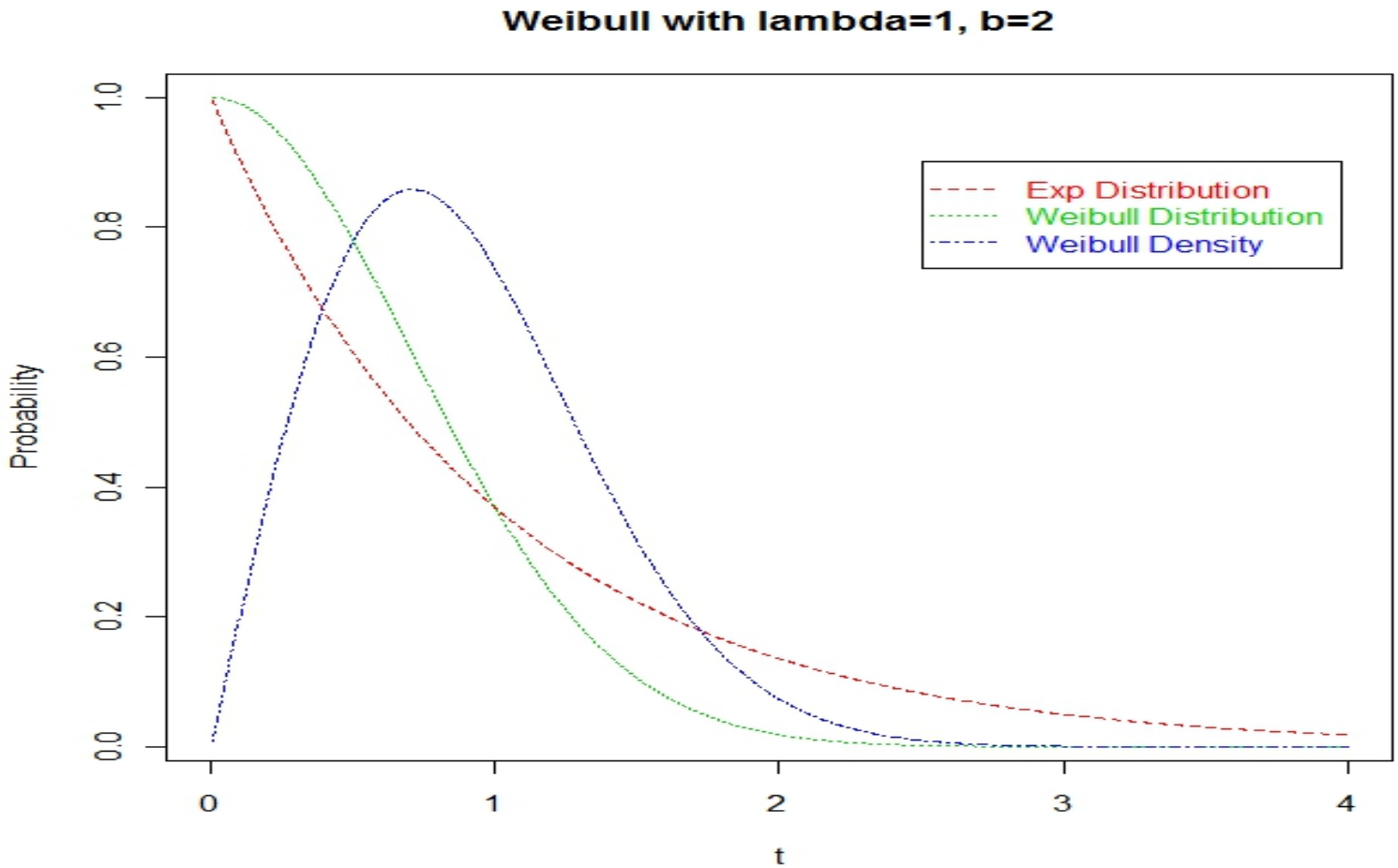

Figure: Weibull distribution compared with exponential distribution

Weibull Distribution for $\lambda = 1$ and $b = 0.5$

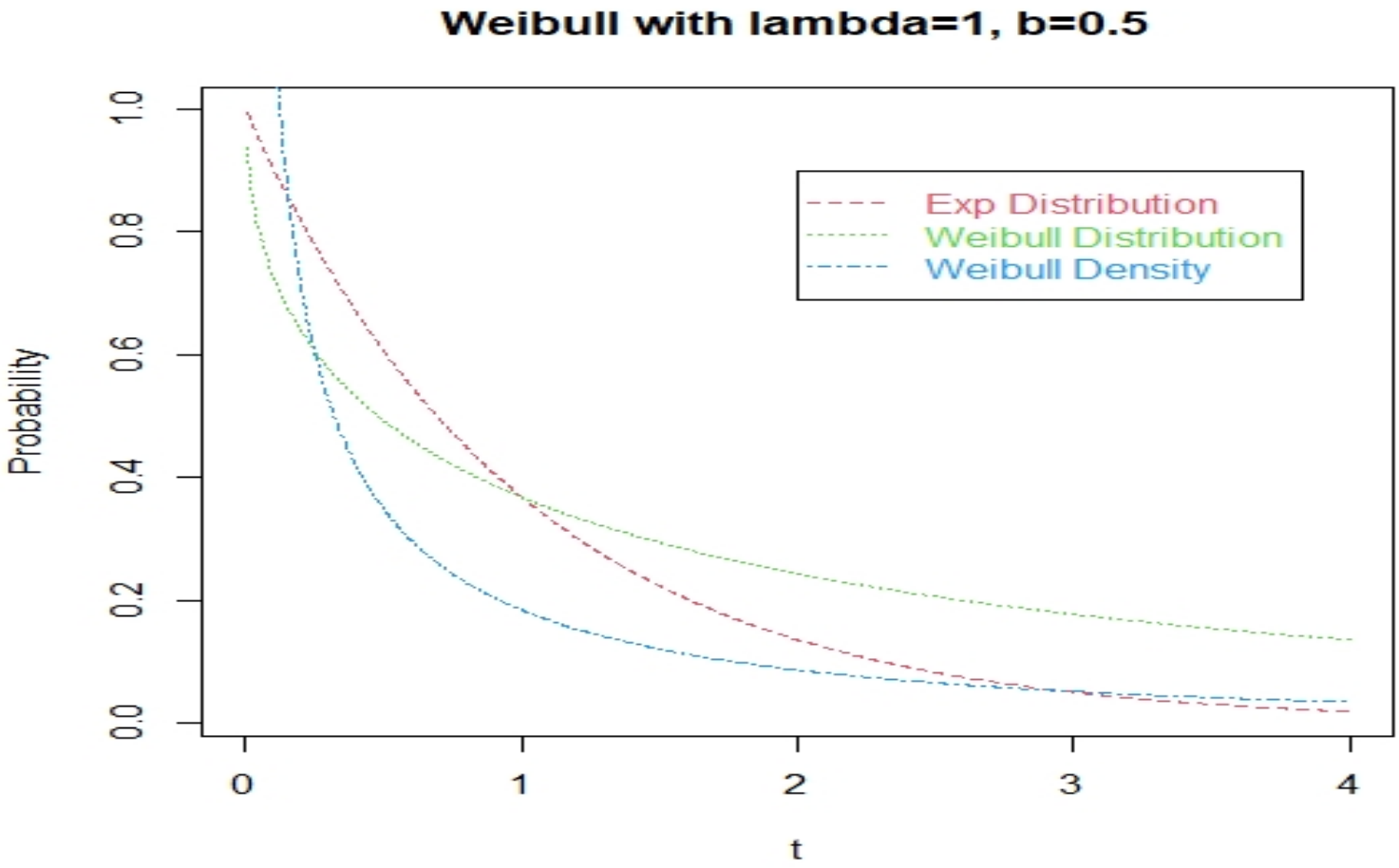

Figure: Weibull distribution compared with exponential distribution

Fit Weibull Model to Right-censored Data

1. The likelihood under the Weibull model

$$\prod_{i=1}^{n} h^{\delta_i}(X_i)S(X_i) = \prod_{i=1}^{n} b^{\delta_i}\lambda^{b\delta_i}X_i^{(b-1)\delta_i}\exp\left\{-(\lambda X_i)^b\right\}.$$

2. Log-likelihood

$$l(\lambda, b) = \sum_{i=1}^{n}\delta_i\left(\log b + b\log\lambda\right)+(b-1)\sum_{i=1}^{n}\delta_i\log X_i-\lambda^b\sum_{i=1}^{n}X_i^b$$

3. First derivatives

$$\begin{aligned}\frac{\partial l}{\partial \lambda} &= \frac{b}{\lambda}\sum_{i=1}^{n}\delta_i - b\lambda^{b-1}\sum_{i=1}^{n}X_i^b,\\ \frac{\partial l}{\partial b} &= \sum_{i=1}^{n}\delta_i\left(\frac{1}{b}+\log\lambda\right)+\sum_{i=1}^{n}\delta_i\log X_i - \lambda^b\sum_{i=1}^{n}X_i^b\log(\lambda X_i).\end{aligned}$$

Fit Weibull Model to Right-censored Data (continuing)

1. Parameter estimator for λ satisfying

$$\hat{\lambda} = \left[\frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} X_i^b} \right]^{1/b}.$$

2. Parameter estimator for b satisfying

$$\sum_{i=1}^{n} \delta_i \left(\frac{1}{b} + \log X_i \right) - \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} X_i^b} \sum_{i=1}^{n} X_i^b \log X_i = 0.$$

The solution for b needs iterative approximation.

Variance Estimation

1. Second derivatives of the log-likelihood

$$\begin{aligned}
\frac{\partial^2 l}{\partial \lambda^2} &= -\frac{b}{\lambda^2}\sum_{i=1}^{n}\delta_i - b(b-1)\lambda^{b-2}\sum_{i=1}^{n}X_i^b, \\
\frac{\partial^2 l}{\partial b^2} &= -\sum_{i=1}^{n}\delta_i/b^2 - \lambda^b\sum_{i=1}^{n}X_i^b\left\{\log(\lambda X_i)\right\}^2, \\
\frac{\partial^2 l}{\partial \lambda \partial b} &= \frac{1}{\lambda}\sum_{i=1}^{n}\delta_i - \lambda^{b-1}\sum_{i=1}^{n}X_i^b - b\lambda^{b-1}\sum_{i=1}^{n}X_i^b\log(\lambda X_i).
\end{aligned}$$

2. When the maximum likelihood estimators for λ and b are obtained, the variance for $(\hat{\lambda}, \hat{b})$ is

$$\begin{pmatrix} -\frac{\partial^2 l}{\partial \lambda^2}(\hat{\lambda},\hat{b}) & -\frac{\partial^2 l}{\partial \lambda \partial b}(\hat{\lambda},\hat{b}) \\ -\frac{\partial^2 l}{\partial \lambda \partial b}(\hat{\lambda},\hat{b}) & -\frac{\partial^2 l}{\partial b^2}(\hat{\lambda},\hat{b}) \end{pmatrix}^{-1}.$$

3. In general, we need software to fit such models to survival data.

Extension of Weibull model

1. The Weibull regression model can be further generalized to

$$\log T = \beta_0 + \beta Z + \sigma\epsilon,$$

   where ϵ has survival function $S(u)$.
2. When the distribution for ϵ are

$$\begin{aligned} \text{extreme-value}: &\quad S_\epsilon(u) = \exp\{-\exp(u)\}, \\ \text{logistic}: &\quad S_\epsilon(u) = \{1+\exp(u)\}^{-1}, \\ \text{normal}: &\quad S_\epsilon(u) = 1-\Phi(u), \end{aligned}$$

   the distribution for T are respectively Weibull, log-logistic, and log-normal, where Φ denotes standard normal distribution.
3. Such models are collectively called accelerated failure time model.

The Name: Accelerated failure time model

1. In the accelerated failure time (AFT) model, let T_0 denote the survival time when $Z = 0$, i.e.,

$$\log T_0 = \beta_0 + \sigma\epsilon.$$

2. Let T_z denote the survival time when $Z = z$, then

$$\log T_z = \beta_0 + \beta Z + \sigma\epsilon = \log T_0 + \beta Z.$$

3. That means

$$T_z = T_0 \exp(\beta Z).$$

   The survival time with covariate value $Z = z$ for a subject is shorten/prolonged by a factor of $\exp(\beta Z)$ when compared with the same subject with a covariate value $Z = 0$

Parameter interpretation for the aAccelerated failure time model

1. Note that
$$T = \exp(\epsilon)e^{\beta_0+\beta Z}.$$
It follows that
$$E(T \mid Z) = e^{\beta_0+\beta Z}E\{\exp(\epsilon)\}.$$
2. For two covariate values z and $z+1$,
$$\frac{E(T \mid Z = z+1)}{E(T \mid Z = z)} = e^{\beta}.$$

Fit AFT model

1. Accelerated failure time model

$$\log T = \beta_0 + \beta Z + \sigma\epsilon,$$

   where ϵ has survival function $S(u)$.

2. The survival function for a subject with covariate Z,

$$\begin{aligned} S_T(t|Z) &= P(T > t|Z) \\ &= P(\log T - \beta_0 - \beta Z > \log t - \beta_0 - \beta Z|Z) \\ &= P(\epsilon > \frac{1}{\sigma}(\log t - \beta_0 - \beta Z)|Z) \\ &= S_\epsilon\left\{\frac{1}{\sigma}(\log t - \beta_0 - \beta Z)\right\}. \end{aligned}$$

3. The hazard density function is

$$h(t|Z) = \frac{-S'_\epsilon\left\{\frac{1}{\sigma}(\log t - \beta_0 - \beta Z)\right\}}{S_\epsilon\left\{\frac{1}{\sigma}(\log t - \beta_0 - \beta Z)\right\}}\frac{1}{\sigma t}.$$

Fit AFT model (continuing)

1. The likelihood for the observed data (X_i, δ_i, Z_i), $i = 1, \cdots, n$.

$$\prod_{i=1}^{n} h^{\delta_i}(X_i|Z_i) S_T(X_i|Z_i)$$

2. SAS Proc lifereg is designed to fit AFT models.

Fit Weibull Model in SAS using PROC LIFEREG

1. We can use SAS PROC LIFEREG to fit the model.
2. SAS proc lifereg statement.

   Proc lifereg;
   Model time*censor(censoring indicators)=
   /d=Weibull;
   run;

   The distribution option refers to the survival time distribution.
3. Other distribution options include: d=log-normal, d=log-normal etc.

Parametrization of AFT model in SAS

In AFT model, $Y = \log T$, $\mu = \beta_0 + \beta Z$, and $Y = \mu + \sigma\epsilon$.

1. When ϵ follows normal distribution

$$Y = \log T \sim N(\mu, \sigma), \quad S_Y(y) = 1 - \Phi\left(\frac{y-\mu}{\sigma}\right).$$

T follows the log-normal distribution.

2. When ϵ follows logistic

$$Y = \log T \sim \textit{logistic}(\mu, \sigma), \quad S_Y(y) = \left\{1 + \exp\left(\frac{y-\mu}{\sigma}\right)\right\}^{-1}.$$

T follows the log-logistic distribution.

3. When ϵ follows extreme-value distribution,

$$Y = \log T \sim \text{Extreme-value}, \quad S_Y(y) = \exp\left\{-\exp\left(\frac{y-\mu}{\sigma}\right)\right\}$$

T follows the log-extreme=value distribution, which is Weibull distribution.

Regression Models Fit to the Myeloma Data

Show the different parametric model fits to the same dataset using SAS proc lifereg.

1. If extreme-valuel distribution is assumed, *BUN* is significant at 5% error rate.
2. If log-normal distribution is assumed, *BUN* and *HB* are significant at 5% error rate.
3. If log-logistic distribution is assumed, *BUN*, *HB*, and *BJ* are significant at 5% error rate.

Question: Which model should we use?

Model Checking

1. Check the regression form: we can enlarge the model by adding high order terms to the regression and test for the necessity for including those terms.
2. Check the error distribution: checking the residual distribution is complicated by the presence of censoring.
3. Model comparison.

Model checking is challenging. We will study this problem more carefully later.

Weibull Parametrization in SAS PROC LIFEREG

1. Parametrization of Weibull distribution in SAS Proc lifereg.

$$S_T(t) = \exp\left\{-\exp\left(\frac{\log t - \mu}{\sigma}\right)\right\}.$$

2. Our parametrization

$$S_T(t) = \exp\left\{-(\lambda t)^b\right\}$$

3. Relationship in parameters

$$\lambda = \exp(-\mu), \qquad b = 1/\sigma.$$

Weibull Model Fit in SAS Using PROC LIFEREG (continuing)

1. Model fit in SAS

$$\hat{\mu} = 3.4348, \qquad \hat{\sigma} = 1.0336.$$

2. Weibull model parameters

$$\begin{aligned} \hat{\lambda} &= \exp(-\hat{\mu}) = \exp(-3.4348) = 0.0332, \\ \hat{b} &= 1/\hat{\sigma} = 1/1.0336 = 0.9675. \end{aligned}$$

3. Recall the exponential fit has

$$\hat{\lambda} = 0.0321, \qquad b = 1.$$

There is only a very small difference.

Survival Function Estimates Based on Weibull Model

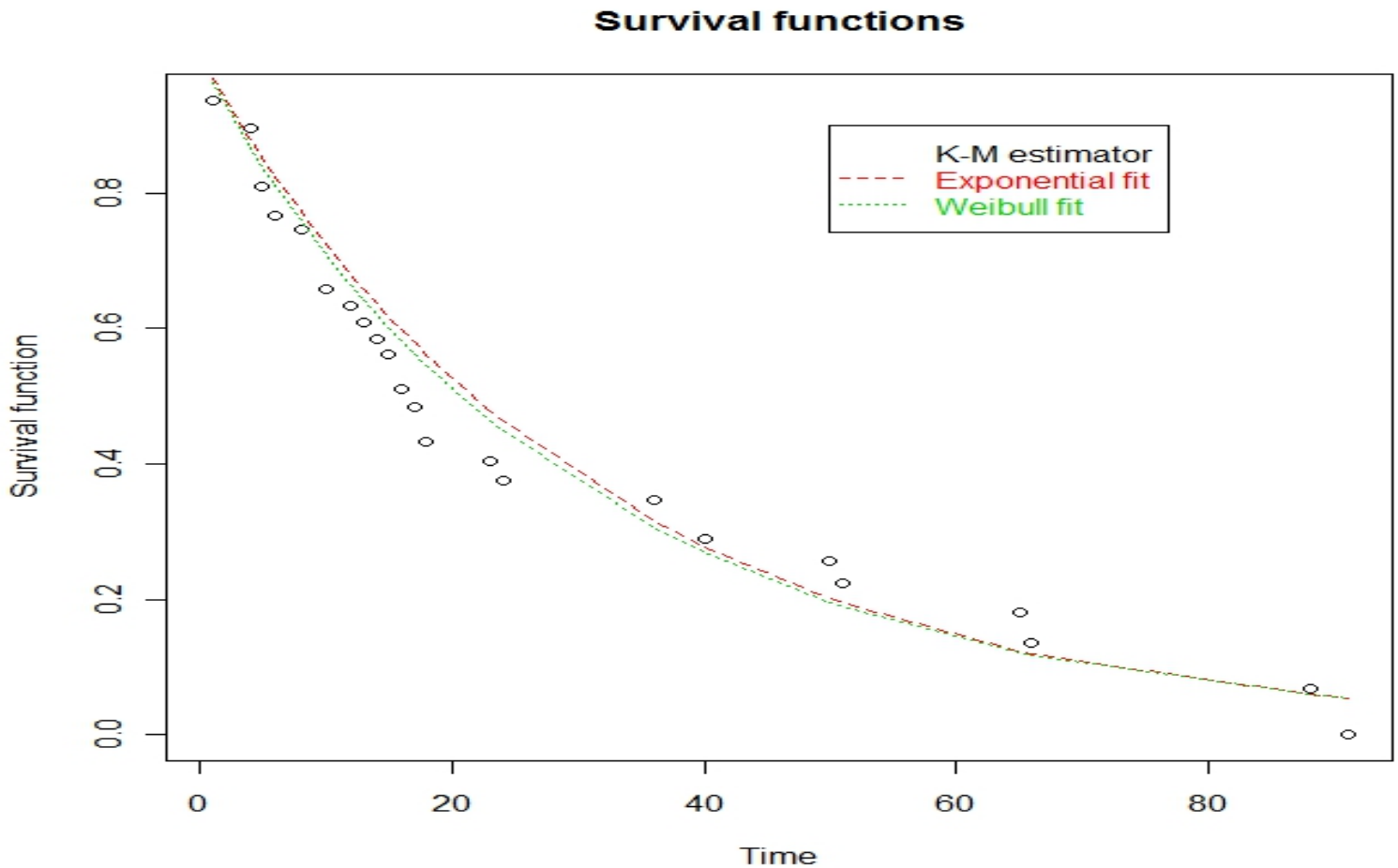

Figure: Survival function estimate for the data on multiple myeloma.