Survival Analysis: Parametric Analysis

Samiran Sinha
Texas A&M University
sinha@stat.tamu.edu

October 3, 2021

Types of Censoring

- Left censoring: In a study if it is known that the failure time of a subject is below a known threshold (left censoring time), then we call the subject is left censored. For that subject we know that the event has happened some time before the left censoring time. In this scenario, the exact failure time is observed if it is larger than the left censoring time.
- Example (Reference: Klein and Moeschberger, 2003): In a study to determine the distribution of the time until first marijuana use among high school boys in California, the question was asked, "When did you you first use marijuana?" One of the responses was "I have used it but can not recall just when the first time was." A boy who chose this response is indicating that the event had occurred prior to the boy's age at interview but the exact age at which he started using marijuana is unknown. This is an example of a left-censored event time. [Copied from the book]

Types of Censoring

- Interval censoring: In a study if it is known that the failure time of a subject is below a known threshold (left censoring time), then we call the subject is left censored. If the failure time is above a known threshold we call the subject right censored. In the case of interval censoring, we observe the exact failure time if the event happens after the left censoring time and before the right censoring time. For the interval censored subject, although we do not know the exact failure time, we know if the subject is left or right censored.

Types of Censoring

For example[1] consider the acquired immune deficiency syndrome (AIDS) trials, where the interest is in times to AIDS for human immunodeficiency virus (HIV) infected subjects. The onset of AIDS is determined by blood testing (CD4 T cell counts fall below 200 per cubic millimeter of blood or AIDs defining complications) that is performed obviously only periodically but not continuously. Consequently, only interval-censored data may be available for AIDS diagnosis times. A similar case is for studies on HIV infection times. The HIV infection time of a HIV positive person is usually determined by a retrospective analysis of the person's medical history. Therefore, we are only able to obtain an interval given by the last HIV negative test date and the first HIV positive test date for the HIV infection time.

[1] Zhang Z, Sun J. Interval censoring. *Stat Methods Med Res*. 2010;19(1):53–70.

Types of censoring

Right censoring: Type I censoring happens when the event is observed only if it occurs prior to some prespecified time [2]. These censoring times may vary from individual to individual. A typical animal study or clinical trial starts with a fixed number of animals or patients to which a treatment (or treatments) is (are) applied. Then subjects are followed up to a fixed time point, say 2 years. Any subjects that are censored between 0 to 2 years are considered randomly censored, any subjects that did experience the event of interest over (0, 2) years are failures, and the remaining subjects that didn't experience the event at the end of 2 years are administratively censored.

[2] Joffe, MM. Administrative and artificial censoring in censored regression models. *Stat in Med*, 2001; 20 (15): 2287–2304.

Types of censoring

Right censoring: Type II censoring occurs when a study continues until the failure of the first r subjects, where r is some predetermined integer ($r \leq n$). Experiments involving Type II censoring are often used in testing of equipment life. You may find numerous articles on accelerated life testing on battery life that make use of the failure time models. Here, all items are put on test at the same time, and the test is terminated when r of the n items have failed. Such an experiment may save time and money because it could take a very long time for all items to fail. It is also true that the statistical treatment of Type II censored data is simpler because the data consists of the r smallest lifetimes in a random sample of n lifetimes, so that the theory of order statistics is directly applicable to determining the likelihood and any inferential technique employed.

Censoring assumptions

- **Independent censoring:** Within any subgroup, subjects censored at time t representative of all subjects in that subgroup who are still at risk at time t.
  - Censoring is random within any subgroup of interest
  - Censoring time is independent of failure time
- **Non-informative censoring:** The distribution of survival times (T) provides no information about the distribution of censorship times (C), and vice verse.

Truncation

- Truncation is a different concept than censoring. We do not have any information on the truncated subjects.
- On the other hand, if a subject is censored that means we get to observe a partial information from that subject.
- Suppose that in a clinical study on HIV therapy, subjects with age more than 20 years and less than 60 years are recruited. That means those subjects whose age is outside the range $(20, 60)$ will not be included in the study and we will not have any information on those subjects. Thus, we will not have any information (not even partial) on the subjects outside the truncation limit.

Example (parametric):

- T : time-to-event
- Let V be the observed time, Δ be the censoring indicator (0 if right-censored, 1 if event is obseved).
- $Y = \log(T)$, and suppose that $Y \sim f_{\boldsymbol{\theta}}$, and $C \sim g$.
- Assume that censoring is independent of the time-to-event T.
- For an uncensored observation $(\log(V_i) = \log(v_i), \Delta_i = 1)$, the likelihood contribution is

$$\begin{aligned}\mathcal{L}_i &= \text{pr}\{\log(V_i) = \log(v_i), \Delta_i = 1\} \\ &= \text{pr}\{Y_i = \log(T_i) = \log(v_i), \log(C_i) \geq \log(v_i)\} \\ &= f_{\boldsymbol{\theta}}(\log(v_i))\text{pr}\{\log(C_i) \geq \log(v_i)\}\end{aligned}$$

- For a censored observation $(\log(V_i) = \log(v_i), \Delta_i = 0)$, the likelihood contribution is

$$\begin{aligned}\mathcal{L}_i &= \text{pr}\{\log(V_i) = \log(v_i), \Delta_i = 0\} \\ &= \text{pr}\{\log(C_i) = \log(v_i), Y_i = \log(T_i) \geq \log(v_i)\} \\ &= g(v_i)\text{pr}\{Y_i \geq \log(v_i)\} \\ &= g(v_i)\int_{\log(v_i)}^{\infty} f_{\boldsymbol{\theta}}(u)du\end{aligned}$$

Example (parametric):

- Suppose that we have a data on (V, Δ) on two subjects, and they are $(5, 1)$ and $(1, 0)$.
- The likelihood for the first subject is

$$\begin{aligned}\mathcal{L}_1 &= \text{pr}\{Y_i = \log(T_i) = \log(5), \log(C_i) \geq \log(5)\} \\ &= f_{\boldsymbol{\theta}}(\log(5))\text{pr}\{\log(C_i) \geq \log(5)\}.\end{aligned}$$

- The likelihood for the second subject is

$$\begin{aligned}\mathcal{L}_2 &= \text{pr}\{\log(C_i) = \log(1), Y_i = \log(T_i) \geq \log(v_i)\} \\ &= g(1) \int_{\log(1)}^{\infty} f_{\boldsymbol{\theta}}(u)du.\end{aligned}$$

Example (parametric) continued:

The likelihood for an observed sample of n subjects is thus

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{n}\left\{ f_{\boldsymbol{\theta}}(\log(v_i))\mathrm{pr}\{\log(C_i) \geq \log(v_i)\}\right\}^{\Delta_i} \\ &\quad \times \left\{ g(v_i)\int_{\log(v_i)}^{\infty} f_{\boldsymbol{\theta}}(u)du\right\}^{1-\Delta_i} \\ &= \prod_{i=1}^{n} f_{\boldsymbol{\theta}}^{\Delta_i}(\log(v_i))\left\{\int_{\log(v_i)}^{\infty} f_{\theta}(u)du\right\}^{1-\Delta_i} \\ &\quad \times \prod_{i=1}^{n}\left\{\int_{v_i}^{\infty} g(u)du\right\}^{\Delta_i} g^{1-\Delta_i}(v_i)\end{aligned}$$

Example (parametric) continued:

The log-likelihood (part involving the parameters of the distribution of T) is

$$\ell = \sum_{i=1}^{n} \left[\Delta_i \log\{f_{\boldsymbol{\theta}}(\log(v_i))\} + (1-\Delta_i)\log\left\{ \int_{\log(v_i)}^{\infty} f_{\boldsymbol{\theta}}(u)du \right\} \right]$$

Example (parametric) continued:

- Suppose that $f_{\boldsymbol{\theta}} \sim N(\mu, \sigma^2)$. We can estimate $\boldsymbol{\theta}' = (\mu, \sigma^2)$ by numerically maximizing the log-likelihood.
- Note: Next, we can then estimate the survival function as $1 - F_{\widehat{\boldsymbol{\theta}}}$.
- Next we shall discuss some special cases.

AFT model

- Suppose that along with (V_i, Δ_i) we observe a set of covariates (explanatory variables), X_i.
- We are interested in the following linear model

$$Y_i = \log(T_i) = \beta_0 + X_i^T \beta_1 + U_i,$$

where the random noise U_i has a known distribution. This model is known as accelerated failure time (AFT) model. If the distribution of U is assumed to be a known family of distributions, we call it parametric AFT otherwise it is a non-parametric AFT model. Here we shall focus on the parametric AFT model.

- A special case is when U_i is assumed to follow $\text{Normal}(0, \sigma)$ distribution (the normal family of distributions with mean zero and unknown scale σ).
- When U_i follows Normal$(0, \sigma)$, the distribution of T_i is called lognormal and the mean of $\log(T)$ is $\beta_0 + X_i^T \beta_1$ and scale σ.
- The log-likelihood when $U_i \sim \text{Normal}(0, \sigma)$ is

$$\ell = \sum_{i=1}^{n} \Bigg(-\Delta_i \left[\frac{\{\log(V_i) - \beta_0 - X_i^T \beta_1\}^2}{2\sigma^2} + 0.5 \log(\sigma^2) \right]$$

$$+(1-\Delta_i) \log \left\{ \int_{\log(v_i)}^{\infty} \frac{\exp\{-(u - \beta_0 - X_i^T \beta_1)^2 / 2\sigma^2\}}{\sqrt{2\pi}\sigma} du \right\} \Bigg)$$

AFT model fitting

- In a model fitting, we want to estimate the unknown model parameters from the data.
- Next, we want to estimate the uncertainties, and then statistical inference and prediction.

Log-normal model fitting to a simulated data example

Code

```
library(survival)
 set.seed(10)
 n=500
 myx=runif(n, -1, 1)
 myy=1+0.5*myx+0.25*rnorm(n, 0, 1)
 mytime=exp(myy)
 mycen=runif(n, 3, 15)
 myv=apply(cbind(mytime, mycen), 1, min)
 mydel=as.numeric(mytime<mycen)
 mydata=data.frame(myv, mydel, myx)
 names(mydata)<-c("time", "mystatus", "mycov")
 head(mydata)
 out1=survreg(Surv(time, mystatus)~mycov,data=mydata, dist="lognormal")
 #### Alternative method to obtain the exact same result
 out2=survreg(Surv(log(time), mystatus)~mycov,data=mydata,
 dist="gaussian")
```

Simulated data example

Code

```
summary(out1)

Call:
survreg(formula = Surv(time, mystatus) ~ mycov, data = mydata,
    dist = "lognormal")
              Value Std. Error     z      p
(Intercept)  1.0018     0.0119  84.0 <2e-16
mycov        0.4870     0.0209  23.3 <2e-16
Log(scale)  -1.3284     0.0323 -41.1 <2e-16

Scale= 0.265

Log Normal distribution
Loglik(model)= -529.2   Loglik(intercept only)= -714.6
Chisq= 370.83 on 1 degrees of freedom, p= 1.2e-82
Number of Newton-Raphson Iterations: 5
n= 500
```

Interpretations

- The estimate of β_0 is 1.0018 and the standard error is 0.0119. The 95% CI is $(1.0018 \pm 1.96 \times 0.0119) = (0.978, 1.025)$.
- The estimate of β_1 is 0.487 and the standard error is 0.0209. The 95% CI is $(.487 \pm 1.96 \times 0.0209) = (0.446, 0.528)$.
- The estimate of σ is $exp(-1.3284) = 0.265$, and the 95% CI is $\exp(-1.328 \pm 1.96 \times 0.0323) = (0.249, 0.282)$.
- The Newton-Raphson method was used to maximize the log-likelihood ℓ in 3 slides back.

Estimation of survival probability

- Suppose that we are interested in survival probability when $X = X_0$ (a given value). That means we want to estimate

$$S(t|X_0) = \text{pr}(T > t|X_0) = \text{pr}(Y > \log(t)|X_0)$$
$$= 1 - \Phi\left(\frac{\log(t) - \beta_0 - X_0^T\beta_1}{\sigma}\right),$$

  where Φ is the CDF of Normal$(0, 1)$ distribution.
- The estimator of this survival probability is

$$\widehat{S}(t|X_0) = 1 - \Phi\left(\frac{\log(t) - \widehat{\beta}_0 - X_0^T\widehat{\beta}_1}{\widehat{\sigma}}\right)$$

- How do we compute the standard error of the 95% CI for $S(t|X_0)$? First, we shall use the delta method to compute the standard error of $\{\log(t) - \widehat{\beta}_0 - X_0^T\widehat{\beta}_1\}/\widehat{\sigma}$, say it is se^*.
- Then the 95% CI for $S(t|X_0)$ is

$$1 - \Phi\left(\frac{\log(t) - \widehat{\beta}_0 - X_0^T\widehat{\beta}_1}{\widehat{\sigma}} \pm 1.96se^*\right)$$

Simulated data example

Suppose that we are interested in making inference on $S(2|X = 0.5) = \text{pr}(T > 2|X = 0.5)$.

Code

```
1-pnorm((log(2)-1.0018-0.5*0.487)/exp(-1.3284)) # estimate of S(2|X=0.5)

library(msm) # you need this package for the delta method
sestar= deltamethod(~(log(2)-x1-0.5*x2)/exp(x3), c(1.0018,
0.4870, -1.3284), out1$var)
# 95% CI

ul=1-pnorm((log(2)-1.0018-0.5*0.487)/exp(-1.3284)
-1.96*sestar) # upper limit
ll=1-pnorm((log(2)-1.0018-0.5*0.487)/exp(-1.3284)
+1.96*sestar) # lower limit

print(c(ll, ul))
```

Another AFT model

- Suppose that our next model is

$$Y_i = \log(T_i) = X_i^T \beta_1 + U_i,$$

where U_i is the random noise. Suppose that $\exp(U_i)$ follows the exponential distribution with mean λ_0. That means $\text{pr}\{\exp(U_i) > r\} = \exp(-r/\lambda_0)$.

- Let us figure out the survival function of T_i given covariate X_i.

$$\begin{aligned}
S(t|X_i) = \text{pr}(T_i > t|X_i) =& \text{pr}\{\log(T_i) > \log(t)|X_i\} \\
=& \text{pr}\{X_i^T \beta + U_i > \log(t)\} \\
=& \text{pr}\{U_i > \log(t) - X_i^T \beta\} \\
=& \text{pr}\{U_i > \log(t) - X_i^T \beta\} \\
=& \text{pr}[\exp(U_i) > \exp\{\log(t) - X_i^T \beta\}] \\
=& \text{pr}\{\exp(U_i) > t \exp(-X_i^T \beta)\} \\
=& \exp\{-t \exp(-X_i^T \beta)/\lambda_0\}
\end{aligned}$$

Another AFT model

- The density function of T conditional on the covariate is

$$f(t|X_i) = -\frac{dS(t|X_i)}{dt} = -\frac{d}{dt}\exp\{-t\exp(-X_i^T\beta)/\lambda_0\}$$
$$= \exp\left\{-\frac{t\exp(-X_i^T\beta)}{\lambda_0}\right\}\frac{\exp(-X_i^T\beta)}{\lambda_0}$$

- The hazard function of T conditional on the covariate is

$$\lambda(t|X_i) = \frac{f(t|X_i)}{S(t|X_i)} = \frac{\exp(-X_i^T\beta)}{\lambda_0}$$

Simulated data example for the exponential model

Code

```
library(survival)
set.seed(10)
n=500
myx=runif(n, -1, 1)
myy=0.5*myx+log(rexp(n, 0.8)) #
### rexp(n, 0.8) generates data from Exponential distribution with
### mean lambda_0=1/0.8=1.25
 mytime=exp(myy)
mycen=runif(n, 2, 15)
myv=apply(cbind(mytime, mycen), 1, min)
mydel=as.numeric(mytime<mycen)
mydata=data.frame(myv, mydel, myx)
names(mydata)<-c("time", "mystatus", "mycov")
head(mydata)
out1=survreg(Surv(time, mystatus)~mycov,data=mydata, dist="exponential")
```

Simulated data example for the exponential model

Code

```
summary(out1)
Call:
survreg(formula = Surv(time, mystatus) ~ mycov, data = mydata,
    dist = "exponential")
             Value Std. Error    z        p
(Intercept) 0.241     0.0453 5.32 1.05e-07
mycov       0.508     0.0799 6.36 2.00e-10

Scale fixed at 1

Exponential distribution
Loglik(model)= -600.8   Loglik(intercept only)= -620.9
Chisq= 40.39 on 1 degrees of freedom, p= 2.1e-10
Number of Newton-Raphson Iterations: 4
n= 500
```

Interpretations of the output

- The estimate of β_1 is 0.508 with the standard error 0.0799.
- The estimate of $\eta_0 = \log(\lambda_0)$ is 0.241 with the standard error 0.0453.
- The estimator of $S(t|X_*)$, the survival function at a given time t for covariate $X = X_*$ is $\exp\{-t\exp(-X_*^T\widehat{\beta})/\exp(\widehat{\eta}_0)\}$, where $\widehat{\beta}$ and $\widehat{\eta}_0$ are the estimators of β and η_0 respectively.
- The 95% CI for $S(t|X_*)$ is

$$\exp\left\{-t\exp(-\widehat{\eta}_0 - X_*^T\widehat{\beta} + 1.96se_*)\right\}, \exp\left\{-t\exp(-\widehat{\eta}_0 - X_*^T\widehat{\beta} - 1.96se_*)\right\}$$

- Here se_* is the standard error of $-\widehat{\eta}_0 - X_*^T\widehat{\beta}$.

Simulated data example

Suppose that we are interested in making inference on $S(1.8|X = 0.5)$.

Code

```
exp(-1.8*exp(-0.241-0.5*0.508)) # estimate of S(1.8|X=0.5)
0.3379
library(msm) # you need this package for the delta method
sestar= deltamethod(~(-x1-0.5*x2), c(0.241, 0.508), out1$var)
# 95% CI

ul=exp(-1.8*exp(-0.241-0.5*0.508-1.96*sestar)) # upper limit
ll=exp(-1.8*exp(-0.241-0.5*0.508+1.96*sestar))# lower limit

print(c(ll, ul))
[1] 0.2901000 0.3780109
```

Mean for the Exponential AFT model

- We know when $Y_i = \log(T_i) = X_i^T \beta_1 + U_i$ where the random noise $\exp(U_i)$ follows the exponential distribution with mean $\lambda_0 = \exp(\eta_0)$, the survival function of T_i given covariate X_i is

$$S(t|X_i) = \exp\{-t \exp(-X_i^T \beta)/\lambda_0\}.$$

- The mean of T_i conditional on covariate X_i is

$$\begin{aligned}\mu(X_i) = \int_0^\infty S(t|X_i)dt &= \int_0^t \exp\{-t \exp(-X_i^T \beta)/\lambda_0\}dt \\ &= \frac{1}{\exp(-X_i^T \beta)/\lambda_0} \\ &= \exp(X_i^T \beta)\lambda_0 \\ &= \exp(\eta_0 + X_i^T \beta).\end{aligned}$$

Inference on the mean

- Inference on the mean conditional on covariate $X = X_*$
- The estimator of $\mu(X_*)$ is $\exp(\widehat{\eta}_0 + X_*^T \widehat{\beta})$.
- The 95% CI for $\mu(X_*)$ is

$$\left\{\exp(\widehat{\eta}_0 + X_*^T \widehat{\beta} - 1.96 se_*),\ \ \exp(\widehat{\eta}_0 + X_*^T \widehat{\beta} + 1.96 se_*)\right\},$$

where se_* is the standard error of $\widehat{\eta}_0 + X_*^T \widehat{\beta}$.

Inference on the mean

Suppose that we are interested in making inference on $\mu(X = 0.5)$.

Code

```
exp(0.241+0.5*0.508) # estimate of mu(0.5)
1.640
library(msm) # you need this package for the delta method
sestar= deltamethod(~(x1+0.5*x2), c(0.241, 0.508), out1$var)
# 95% CI

ll=exp(0.241+0.5*0.508-1.96*sestar) # lower limit
ul=exp(0.241+0.5*0.508+1.96*sestar)# upper limit

print(c(ll, ul))

[1] 1.454511 1.850268
```

Median for the Exponential AFT model

- We know when $Y_i = \log(T_i) = X_i^T \beta_1 + U_i$ where the random noise $\exp(U_i)$ follows the exponential distribution with mean $\lambda_0 = \exp(\eta_0)$, the survival function of T_i given covariate X_i is

$$S(t|X_i) = \exp\{-t \exp(-X_i^T \beta)/\lambda_0\}$$

- The median of T_i conditional on covariate X_i is

$$\begin{aligned} m(X_i) = \inf\{t : S(t|X_i) \leq 0.5\} &= \inf[t : \exp\{-t \exp(-X_i^T \beta)/\lambda_0\} \leq 0.5] \\ &= \inf[t : -t \exp(-X_i^T \beta)/\lambda_0 \leq \log(0.5)] \\ &= \inf[t : t \exp(-X_i^T \beta)/\lambda_0 \geq -\log(0.5)] \\ &= \inf[t : t \geq -\log(0.5)\lambda_0 \exp(X_i^T \beta)] \\ &= \inf[t : t \geq -\log(0.5) \exp(\eta_0 + X_i^T \beta)] \\ &= -\log(0.5) \exp(\eta_0 + X_i^T \beta) \end{aligned}$$

Inference on the median

- Inference on the median conditional on covariate $X = X_*$
- The estimator of $m(X_*)$ is $-\log(0.5)\exp(\widehat{\eta}_0 + X_*^T\widehat{\beta})$.
- The 95% CI for $m(X_*)$ is

$$\left\{-\log(0.5)\exp(\widehat{\eta}_0 + X_*^T\widehat{\beta} - 1.96se_*),\ \ -\log(0.5)\exp(\widehat{\eta}_0 + X_*^T\widehat{\beta} + 1.96se_*)\right\},$$

where se_* is the standard error of $\widehat{\eta}_0 + X_*^T\widehat{\beta}$.

Inference on the median

Suppose that we are interested in making inference on $m(X = 0.5)$.

Code

```
-log(0.5)*exp(0.241+0.5*0.508) # estimate of m(0.5)
1.137

library(msm) # you need this package for the delta method
sestar= deltamethod(~(x1+0.5*x2), c(0.241, 0.508), out1$var)
# 95% CI

ll=-log(0.5)*exp(0.241+0.5*0.508-1.96*sestar) # lower limit
ul=-log(0.5)*exp(0.241+0.5*0.508+1.96*sestar)# upper limit

print(c(ll, ul))
[1] 1.008190 1.282508
```

Another AFT model: Weibull

- Suppose the model is
$$Y_i = \log(T_i) = X_i^T \beta + U_i,$$
where $\epsilon_i = \exp(U_i)$ follows the Weibull distribution with the scale b and shape a.
- The density function of Weibull distribution with the scale b and shape a is
$$f(\epsilon_i) = \frac{a}{b}\left(\frac{\epsilon_i}{b}\right)^{a-1} \exp\left\{-\left(\frac{\epsilon_i}{b}\right)^a\right\}$$
and the survival function is
$$\text{pr}(\epsilon_i > r) = \exp\left\{-\left(\frac{r}{b}\right)^a\right\}$$

Another AFT model: Weibull

- Under the Weibull model for $\exp(U_i)$, let us figure out the survival function of T_i given the covariate X_i.

$$\begin{aligned}
\text{pr}(T_i > t|X_i) =& \text{pr}(Y_i = \log(T_i) > \log(t)|X_i) \\
=& \text{pr}(X_i^T\beta + U_i > \log(t)) \\
=& \text{pr}(U_i > \log(t) - X_i^T\beta) \\
=& \text{pr}[\epsilon_i = \exp(U_i) > \exp\{\log(t) - X_i^T\beta\}] \\
=& \text{pr}\{\epsilon_i > t\exp(-X_i^T\beta)\} \\
=& \exp\left[-\left\{\frac{t\exp(-X_i^T\beta)}{b}\right\}^a\right]
\end{aligned}$$

Another AFT model: Weibull

- Let us figure out the hazard function of T_i given the covariate X_i.

$$\begin{aligned}
\lambda(t|X_i) &= -\frac{d}{dt}\log\{S(t|X_i)\} \\
&= -\frac{d}{dt}\log\left(\exp\left[-\left\{\frac{t\exp(-X_i^T\beta)}{b}\right\}^a\right]\right) \\
&= \frac{d}{dt}\left\{\frac{t\exp(-X_i^T\beta)}{b}\right\}^a \\
&= at^{a-1}\left\{\frac{\exp(-X_i^T\beta)}{b}\right\}^a
\end{aligned}$$

- It is seen that the hazard function is not a constant, rather it is a polynomial function of t.

Median

- Let us figure out the median of T given when the covariate $X = X_*$.
- The median is defined as

$$m(X_*) = \inf\{t : \text{pr}(T > t|X_*) = \inf\{t : \exp\left[-\left\{\frac{t\exp(-X_*^T\beta)}{b}\right\}^a\right] \leq 0.5\}.$$

- Thus, $m(X_*) = \{-\log(0.5)\}^{1/a}\exp\{X_*^T\beta + \log(b)\}$.

Mean

- Let us figure out the mean of T given when the covariate $X = X_*$.
- The mean is (using gamma integration)

$$\begin{aligned}\mu(X_*) &= \int_0^\infty \text{pr}(T > t|X_*)dt \\ &= \int_0^\infty \exp\left[-\left\{\frac{t\exp(-X_*^T\beta)}{b}\right\}^a\right]dt \\ &= \frac{\Gamma(1/a)}{a\exp[(1-a)\{-X_*^T\beta - \log(b)\}]}.\end{aligned}$$

Simulated data example for the Weibull model

Code

```
 library(survival)
 set.seed(10)
 n=500
 myx=runif(n, -1, 1)
 myy=0.3*myx+log(rweibull(n, scale=2, shape=0.8)) #
### rexp(n, scale=2, shape=0.8) generates data from the Weibull
### distribution with shape=0.8 and scale=2. That means according to
### our notations a=0.8, b=2.
mytime=exp(myy)
mycen=runif(n, 2, 15)
myv=apply(cbind(mytime, mycen), 1, min)
mydel=as.numeric(mytime<mycen)
mydata=data.frame(myv, mydel, myx)
names(mydata)<-c("time", "mystatus", "mycov")
head(mydata)
out1=survreg(Surv(time, mystatus)~
mycov,data=mydata, dist="weibull")
```

Output of the analysis

Code

```
summary(out1)
Call:
survreg(formula = Surv(time, mystatus) ~ mycov, data = mydata,
    dist = "weibull")
             Value Std. Error    z        p
(Intercept) 0.637     0.0661 9.63 5.84e-22
mycov       0.283     0.1127 2.51 1.20e-02
Log(scale)  0.318     0.0387 8.21 2.14e-16

Scale= 1.37

Weibull distribution
Loglik(model)= -745.9   Loglik(intercept only)= -749.1
Chisq= 6.32 on 1 degrees of freedom, p= 0.012
Number of Newton-Raphson Iterations: 5
n= 500
```

Interpretations of the results

- The estimate of β is 0.283 with the standard error 0.1127.
- The estimate of the shape parameter a is the inverse of the Scale value in the output. It is important to note that the Scale value in the output is different from the scale parameter of the Weibull model. So the estimate of a is $1/1.374 = 0.727$.
- Also, note that we can obtain the estimate of a as $\exp\{-\text{Log(scale)}\} = \exp(-0.318) = 0.727$.
- The 95% for a is $[\exp\{-\text{Log(scale)} - 1.96se\}, \exp\{-\text{Log(scale)} + 1.96se\}]$, where se is the standard error corresponding to Log(scale). In this example, $se = 0.0387$, so the 95% CI for a is

$$\Big[\exp(-0.318 - 1.96 \times 0.0387), \exp(-0.318 + 1.96 \times 0.0387)\Big] = (0.674,\ 0.785)$$

Interpretations of the results

- The estimate of the Weibull scale parameter b is $\exp(\text{Intercept})$, and here it is $\exp(0.637) = 1.891$.
- The 95% CI for b is $\exp(0.637 \pm 1.960.0661) = (1.661, 2.152)$.
- The good news is that the true values of a and b are in the respective confidence intervals. Here we can verify this because we know the true values of a and b. In case of real dataset, we do not know the true values of a and b.

Inference of the survival probability for the Weibull model

- The following formula always produces a CI that is between zero and one.
- Suppose that we are interested in estimating $S(t_*|X = X_*) = \exp\left[-\left\{\frac{t_* \exp(-X_*^T\beta)}{b}\right\}^a\right]$.
- Suppose that $a = \exp(-\zeta)$, where ζ is the $\mathrm{Log(scale)}$ in the output.
- Then re-write

$$\begin{aligned} S(t_*|X = X_*) &= \exp\left(-\left[\exp\{-X_*^T\beta - \log(b) + \log(t_*)\}\right]^{\exp(-\zeta)}\right) \\ &= \exp\left(-\exp[\{-X_*^T\beta - \log(b) + \log(t_*)\}\exp(-\zeta)]\right) \end{aligned}$$

Inference of the survival probability for the Weibull model

- Suppose that *se* is the standard error of $\{-X_*^T\beta - \log(b) + \log(t_*)\}\exp(-\zeta)$.
- Then the 95% CI for $S(t_*|X = X_*)$ is
  $\exp\left(-\exp[\{-X_*^T\widehat{\beta} - \widehat{\log(b)} + log(t_*)\}\exp(-\widehat{\zeta}) \pm 1.96se]\right)$.

Inference on the survival probability

Code

```
library(survival)
 set.seed(10)
 n=500
 myx=runif(n, -1, 1)
 myy=0.3*myx+log(rweibull(n, scale=2, shape=0.8)) #
### rexp(n, scale=2, shape=0.8) generates data from the Weibull
### distribution with shape=0.8 and scale=2. That means according to
### our notations a=0.8, b=2.
mytime=exp(myy)
mycen=runif(n, 2, 15)
myv=apply(cbind(mytime, mycen), 1, min)
mydel=as.numeric(mytime<mycen)
mydata=data.frame(myv, mydel, myx)
names(mydata)<-c("time", "mystatus", "mycov")
head(mydata)
out1=survreg(Surv(time, mystatus)~
mycov,data=mydata, dist="weibull")
```

Inference on the survival probability

Suppose that we are making inference for $S(2|X = 0.5)$.

Code

```
exp( -(2*exp(-0.5*out1$coef[2]-out1$coef[1]))^(1/out1$scale)  ) #estimate
## of S(2|X=0.5)
[1] 0.3907299

library(msm)
### 95% CI
se=deltamethod(~(-0.5*x2-x1+log(2))*exp(-x3), c(0.637,
 0.2831,0.3182), out1$var)

ul=exp( -exp((-0.5*out1$coef[2]-out1$coef[1]+log(2))*
(1/out1$scale)-1.96*se))

ll=exp( -exp((-0.5*out1$coef[2]-out1$coef[1]+log(2))*
(1/out1$scale)+1.96*se))

print(as.numeric(c(ll, ul)))
[1] 0.3441040  0.4370086
```

Inference of the median conditional on the covariate value

- Suppose that we are interested in estimating the median survival time when $X = X_*$, $m(X_*) = \{-\log(0.5)\}^{1/a} \exp\{X_*^T \beta + \log(b)\}$.
- The estimator is $\widehat{m}(X_*) = \{-\log(0.5)\}^{1/\widehat{a}} \exp\{X_*^T \widehat{\beta} + \log(\widehat{b})\}$. Let $a = \exp(-\zeta)$, where ζ is the $\mathrm{Log(scale)}$ of the output. Then we can re-write

$$\begin{aligned}\widehat{m}(X_*) &= \{-\log(0.5)\}^{\exp(\widehat{\zeta})} \exp\{X_*^T \widehat{\beta} + \widehat{\log(b)}\} \\ &= \exp\left[X_*^T \widehat{\beta} + \log(\widehat{b}) + \exp(\widehat{\zeta})\log\{-\log(0.5)\}\right]\end{aligned}$$

- The 95% CI for $m(X_*)$ is $\exp\left[X_*^T \widehat{\beta} + \log(\widehat{b}) + \exp(\widehat{\zeta})\log\{-\log(0.5)\} \pm 1.96 \times se\right]$, where se is the standard error of $X_*^T \widehat{\beta} + \log(\widehat{b}) + \exp(\widehat{\zeta})\log\{-\log(0.5)\}$.

Inference on the median

Suppose that we are interested in making inference on the median survival time when $X = 0.5$.

Code

```
 as.numeric(exp(0.5*out1$coef[2]+out1$coef[1]+(1/out1$scale)*
log(-log(0.5)))) # estimate of m(0.5)
[1] 1.668574

### 95% CI
se=deltamethod(~ (0.5*x2+x1+exp(x3)*log(-log(0.5))),
c(0.637, 0.2831,0.3182), out1$var)

ll=as.numeric(exp(0.5*out1$coef[2]+out1$coef[1]+
(1/out1$scale)*log(-log(0.5))-1.96*se  ))

ul=as.numeric(exp(0.5*out1$coef[2]+out1$coef[1]+
(1/out1$scale)*log(-log(0.5))+1.96*se  ))
print(c(ll, ul))
[1] 1.391288 2.001122
```

Model comparison

- Here we know that data are generated from the Weibull model (or the exponential model), therefore we fit the Weibull model (or the exponential model) to the data.
- In real life scenario, we do not know how the *nature* generates data. In other words, for real life example, we don't know the true data generating process.
- In that case, we can fit different models to data, and choose the best model based on the AIC or BIC criteria.

Model comparison

- Consider the colon cancer data available in the survival package of R.
- See `https://stat.ethz.ch/R-manual/R-devel/library/survival/html/colon.html` for more details on the dataset.
- The response is the time (in days) from the surgery to the event (recurrence).
- For the time being we consider age, treatment, number of nodes involved, and gender as the explanatory variables.

AIC in survreg

Code

```
colondata=colon[colon$etype==1, ]
out1=survreg(Surv(time, status)~age+rx+sex+nodes, data=colondata,
 dist="lognormal")
out2=survreg(Surv(time, status)~age+rx+sex+nodes, data=colondata,
 dist="exponential")
out3=survreg(Surv(time, status)~age+rx+sex+nodes, data=colondata,
 dist="weibull")
extractAIC(out1)
[1]    7.000 7879.945
extractAIC(out2)
[1]    6.000 8042.662
extractAIC(out3)
[1]    7.000 7968.488
# Since the minimum AIC occurs for the lognormal model, we would prefer
# the lognormal model over the other two.
```

Variable selection

- When there are many potential predictors, for choosing a set of informative predictors for prediction and model fitting purpose we use the stepwise variable selection method.
- Let us consider the colon dataset. Consider only the subset where etype$=1$ (discard deaths). We discard death because from the given data we are not sure if the death happened due to colon cancer or not. Thus, subjects will experience recurrence of the disease after the surgery or censored at some time.
- Consider the time (in days) from the surgery to the event (recurrence) as the response variable.
- We shall consider age, treatment, number of nodes involved, gender, perfor (perforation of colon) as the potential explanatory variables. Also, we consider all two-factor interactions among these explanatory variables.
- Then use the stepwise regression to choose the best subset of the explanatory variables for this dataset.

step function in survreg

Code

```
library(survival)
colondata=colon[colon$etype==1, ]
# To remove rows with any missing values, we do
colondata=colondata[complete.cases(colondata), ]
outf=survreg(Surv(time, status)~sex+age+rx+nodes+perfor+
sex*age+sex*rx+sex*nodes+sex*perfor+
age*rx+age*nodes+age*perfor+
rx*nodes+rx*perfor+
nodes*perfor,
data=colondata, dist="lognormal")
out2=step(outf)
```

The final selected model using the step function

Code

```
summary(out2)
Call:
survreg(formula = Surv(time, status) ~ sex + age + rx + nodes +
    perfor + sex:age + sex:rx + sex:nodes + sex:perfor, data = colondata,
    dist = "lognormal")
                Value Std. Error     z       p
(Intercept)    7.13184    0.50733 14.06 < 2e-16
sex            0.90190    0.73328  1.23   0.219
age            0.01578    0.00789  2.00   0.046
rxLev          0.04477    0.23778  0.19   0.851
rxLev+5FU      0.42338    0.23249  1.82   0.069
nodes         -0.19994    0.02697 -7.41 1.2e-13
perfor        -1.00204    0.52586 -1.91   0.057
sex:age       -0.02121    0.01128 -1.88   0.060
sex:rxLev      0.01222    0.32204  0.04   0.970
sex:rxLev+5FU  0.64213    0.33701  1.91   0.057
sex:nodes      0.08344    0.03636  2.29   0.022
sex:perfor     1.14422    0.76143  1.50   0.133
Log(scale)     0.58914    0.03745 15.73 < 2e-16

Scale= 1.8

Log Normal distribution
Loglik(model)= -3837.3   Loglik(intercept only)= -3894.9
Chisq= 115.13 on 11 degrees of freedom, p= 1.7e-19
Number of Newton-Raphson Iterations: 3
n= 888
```

Important points on the final selected model

- The model selected by the stepwise regression method has all five explanatory variables, sex, age, treatment, nodes, perfor, and only the interaction effect of sex with the other explanatory variables.
- As seen in the output, in the final selected model (out2), not necessarily all the included variables are statistically significant. We see some variables, like sex, rxLev, sex:rxLev, and sex:perfor have large p-values. Only age, nodes, and sex:nodes are statistically significant at the 5% level.
- In this variable selection method we have used only lognormal distribution. We can also use some other distribution, such as the exponential, or Weibull distribution, and then use the stepwise method. Then compare the AIC values of the final model for each of the three distributions.

Test of hypothesis

- Suppose that we are interested in testing if gender has any effect on the time-to-recurrence.
- Based on the best fitted model (using the lognormal distribution) the null hypothesis would be
  $H_0 : \beta_{\text{gender}} = \beta_{\text{gender:age}} = \beta_{\text{sex:rxLev}} = \beta_{\text{sex:rxLev+5FU}} = \beta_{\text{sex:nodes}} = \beta_{\text{sex:perfor}} = 0$
  versus

  H_a: at least one of the coefficients mentioned in H_0 is non-zero.
- This test can be done using Wald's approach or the likelihood ratio test. Here we shall talk about the likelihood ratio test (LRT).

Test of hypothesis

- The basic idea of the LRT is that we compute the test statistic $T = -2\{\log(\mathcal{L}_{H_0}) - \log(\mathcal{L}_{H_a})\}$.
- Under H_0, T follows approximate χ^2 distribution with ρ degrees of freedom. Here ρ is the number of coefficients we test in H_0.
- Important to note that, LRT is applicable where H_0 is a special case of H_a (we call it nested model). Suppose that we are interested in checking if the Weibull model fits the data well whereas the null model is the lognormal model. You clearly see that in this case the null model is not nested within the alternative model, hence, LRT cannot be applied in this scenario. We can make our decision on which of lognormal or Weibull model to fit, based on the AIC values.

Hypothesis test for testing the effect of gender (sex) using lognormal

Code

```
out2=survreg(formula = Surv(time, status) ~ sex + age + rx +
    nodes + perfor +
 sex:age + sex:rx + sex:nodes + sex:perfor, data = colondata,
    dist = "lognormal")

out3=survreg(formula = Surv(time, status) ~  age + rx +
    nodes + perfor, data = colondata,
    dist = "lognormal")

anova(out3, out2)
                                                          Terms
1                                          age+rx+nodes+perfor
2 sex+age+rx+nodes+perfor+sex:age+sex:rx+sex:nodes+sex:perfor
  Resid. Df    -2*LL Test Df Deviance    Pr(>Chi)
1       881 7692.896      NA       NA          NA
2       875 7674.602    =  6 18.29479 0.005536211
```

Interpretation of the results

- The test statistic was 18.29, and the corresponding p-value was 0.0055. Hence we reject H_0 at the 1% level, and conclude that sex has a statistically significant effect on the time to recurrence.
- In a similar way, we can do that test for the exponential or the Weibull model.

Hypothesis test for testing the effect of sex using exponential

Code

```
out2=survreg(formula = Surv(time, status) ~ sex + age + rx +
    nodes + perfor +
 sex:age + sex:rx + sex:nodes + sex:perfor, data = colondata,
    dist = "exponential")

out3=survreg(formula = Surv(time, status) ~  age + rx +
    nodes + perfor, data = colondata,
    dist = "exponential")

anova(out3, out2)
                                                          Terms
1                                              age+rx+nodes+perfor
2 sex+age+rx+nodes+perfor+sex:age+sex:rx+sex:nodes+sex:perfor
  Resid. Df    -2*LL Test Df Deviance    Pr(>Chi)
1       882 7854.004      NA       NA           NA
2       876 7824.564    =  6 29.43968 5.021971e-05
```

Hypothesis test for testing the effect of sex using Weibull

Code

```
out2=survreg(formula = Surv(time, status) ~ sex + age + rx +
    nodes + perfor +
 sex:age + sex:rx + sex:nodes + sex:perfor, data = colondata,
    dist = "weibull")

out3=survreg(formula = Surv(time, status) ~  age + rx +
    nodes + perfor, data = colondata,
    dist = "weibull")

anova(out3, out2)
                                                          Terms
1                                          age+rx+nodes+perfor
2 sex+age+rx+nodes+perfor+sex:age+sex:rx+sex:nodes+sex:perfor
  Resid. Df    -2*LL Test Df Deviance    Pr(>Chi)
1       881 7779.324      NA       NA           NA
2       875 7754.851    =  6 24.47213 0.0004274692
```

Alternative to the anova function

Instead of using the anova function, we can directly calculate the test statistic as follows. Then we compute the p-value.

Code

```
out2=survreg(formula = Surv(time, status) ~ sex + age + rx +
    nodes + perfor +
 sex:age + sex:rx + sex:nodes + sex:perfor, data = colondata,
    dist = "weibull")

out3=survreg(formula = Surv(time, status) ~  age + rx +
    nodes + perfor, data = colondata,
    dist = "weibull")

mytest=as.numeric(-2*(logLik(out3)-logLik(out2)))
print(mytest)
[1] 24.47213
### p-value
1-pchisq(24.47213, 6)
[1] 0.0004274686
```