

# STAT 6390: Analysis of Survival Data

Textbook coverage: Chapter 8

Steven Chiou

Department of Mathematical Sciences,  
University of Texas at Dallas

# Parametric models

- Methods described previously are *non-parametric*; no distributional assumptions were made.
- Non-parametric (and semi-parametric) methods have the flexibility to accommodate a wide range of applications.
- If the assumption of a particular probability distribution for the data is valid, inferences based on such an assumption will be more precise.
- The validity of the parametric methods depends heavily on the appropriateness of the distributional assumption.
- Parametric models are often much easier to work with.

# Maximum likelihood estimation

- Suppose *actual survival times* observed for  $n$  individuals are  $\{t_1, \dots, t_n\}$ .
- If the probability density function of the random variable associated with the survival times is  $f(t)$ , the likelihood of the  $n$  observations is

$$\prod_{i=1}^n f(t_i).$$

- If a distributional assumption is made (e.g.,  $f(t) = \lambda e^{-t\lambda}$ ), the unknown parameters ( $\lambda$ ) can be estimated by maximizing the likelihood.

# Maximum likelihood estimation

- Now suppose a the survival data includes (right) censored data.
- In this case,  $n$  pairs of observations are observed  $(\tilde{t}_i, \Delta_i), i = 1, \dots, n$ , where  $\tilde{T}_i$  is the observed survival time.
- When  $\Delta_i = 0$ ,  $t_i$  is right-censored.
- The likelihood then takes the form

$$L = \prod_{i=1}^n [f_T(t_i)]^{\Delta_i} \cdot [S_T(t_i)]^{1-\Delta_i} = \prod_{i=1}^n [h_T(t_i)]^{\Delta_i} \cdot S_T(t_i). \quad (1)$$

- The last equation follows from the property  $h(t) = f(t)/S(t)$ .
- Note that the derivation of of  $L$  does not require a distributional assumption.

# Maximum likelihood estimation

- A more careful derivation of the likelihood function in (1) is to assume the censoring times to be random.
- Let  $C_i$  be the random variable associated with the censoring time.
- Let  $\tilde{T}_i$  be the observed survival time,  $\tilde{T}_i = \min(C_i, T_i)$ .
- We will consider censored and uncensored cases separately.
- For the censored observation:

$$P(\tilde{T}_i = t, \Delta_i = 0) = P(C_i = t, T_i > t).$$

- For the uncensored observation:

$$P(\tilde{T}_i = t, \Delta_i = 1) = P(T_i = t, C_i > t).$$

- The likelihood is then

$$L^* = \prod_{i=1}^n [P(T_i = t, C_i > t)]^{\Delta_i} \cdot [P(C_i = t, T_i > t)]^{1-\Delta_i}.$$

# Maximum likelihood estimation

- Under the assumption that  $C_i$  and  $T_i$  are independent,  $L^*$  becomes

$$L^* = \prod_{i=1}^n [f_T(t_i)S_C(t_i)]^{\Delta_i} \cdot [f_C(t_i)S_T(t_i)]^{1-\Delta_i}.$$

- If the interest is in the parameter estimation in  $f_T(\cdot)$ , e.g., the  $\lambda$  in the exponential assumption,  $f_C(\cdot)$  and  $S_C(\cdot)$  can be considered as constant in the maximum likelihood estimation and  $L^*$  reduces to  $L$ .
- This construction shows the relevance of the assumption of independent censoring.

# Exponential model

- If a random variable,  $T$ , follows an exponential distribution with rate  $\lambda$ , then

$$f(t) = \lambda e^{-\lambda t}, S(t) = e^{-\lambda t}, \text{ and } h(t) = \lambda.$$

- If we are willing to assumption  $\{t_1, \dots, t_n\}$  are iid samples from an exponential distribution with rate  $\lambda$ , then the likelihood  $L(\lambda)$  is

$$L(\lambda) = \prod_{i=1}^n [\lambda e^{-\lambda t_i}]^{\Delta_i} \cdot [e^{-\lambda t_i}]^{1-\Delta_i} = \prod_{i=1}^n \lambda^{\Delta_i} \cdot e^{-\lambda t_i}.$$

- The log-likelihood is

$$\log L(\lambda) = \ell(\lambda) = \log(\lambda) \left( \sum_{i=1}^n \Delta_i \right) - \lambda \sum_{i=1}^n t_i.$$

# Exponential model

- Solving for

$$\frac{d \log L(\lambda)}{d\lambda} = \ell'(\lambda) = 0 \text{ gives } \hat{\lambda} = \frac{\sum_{i=1}^n \Delta_i}{\sum_{i=1}^n t_i}.$$

- The maximum likelihood estimator (MLE),  $\hat{\lambda}$ , is the *number of deaths* divides by the total survival time (*number of person-years*).
- The MLE for the average survival time is  $1/\hat{\lambda}$ , which is the total survival time divides by the number of deaths.
- With  $\hat{\lambda}$ , other quantities like the MLE for median survival times, can be derived.



# Exponential model

- The second derivative of  $\ell(\lambda)$  gives the *information*.
- In the exponential model, we have

$$\ell''(\lambda) = -\frac{1}{\lambda^2} \sum_{i=1}^n \Delta_i.$$

- The standard MLE theory implies

$$\text{Var}(\hat{\lambda}) \approx \frac{\hat{\lambda}^2}{\sum_{i=1}^n \Delta_i}.$$

- The  $100(1 - \alpha)\%$  confidence interval can be constructed accordingly.
- The Delta method can be applied to obtain standard errors for  $g(\lambda)$ , e.g., average survival time, median survival time, etc.

# Weibull model

- The simplicity of the exponential distribution makes it attractive for some specialized applications.
- A more flexibility alternative is modeling with the Weibull distribution.
- If  $T$  follows a Weibull distribution with scale parameters  $\lambda$  and shape parameter  $\gamma$ , then

$$f(t) = \lambda \gamma t^{\gamma-1} e^{-\lambda t^\gamma}, S(t) = e^{-\lambda t^\gamma}, \text{ and } h(t) = \lambda \gamma t^{\gamma-1}.$$

- It is easy to see that when  $\gamma = 1$ , Weibull reduces to an exponential distribution with rate  $\lambda$ .

# Weibull model

- Following the similar procedure as before, the likelihood  $L(\lambda, \gamma)$  is

$$\prod_{i=1}^n \left\{ \lambda \gamma t_i^{\gamma-1} \right\}^{\Delta_i} e^{-\lambda t_i^{\gamma}}.$$

- Let  $\ell(\lambda, \gamma) = \log L(\lambda, \gamma)$ , the MLE for  $\lambda$  turns out to be

$$\hat{\lambda} = \frac{\sum_{i=1}^n \Delta_i}{\sum_{i=1}^n t_i^{\hat{\gamma}}},$$

but there is no close-form solution for  $\hat{\gamma}$ .

# Weibull model

- The MLE  $\hat{\theta} \equiv (\hat{\lambda}, \hat{\gamma})$  can be obtained directly implementing the likelihood and optimized with `optim`.
- Numerical method like the Newton-Raphson procedure can also be used.
- The basic idea of the Newton-Raphson procedure iterates

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \left( -\frac{d^2\ell(\hat{\theta}_n)}{d\theta^2} \right)^{-1} \cdot \frac{d\ell(\hat{\theta}_n)}{d\theta}.$$

- The variance-covariance matrix comes as a by-product.

# Weibull model

- Since parametric models are sensitive to the distributional assumption, it is important to have a diagnostic tool.
- A diagnostic tool for Weibull model is derived from its survival curve.
- The log-log transformation of the Weibull survival function gives

$$\log[-\log S(t)] = \log(\lambda) - \gamma \log(t).$$

- This suggest that if  $\log[-\log S(t)]$  is plotted against  $\log(t)$ , we would expect to see a straight line if the Weibull assumption is valid.
- $S(t)$  can be replaced with  $\hat{S}_{KM}(t)$  or  $\hat{S}_{NA}(t)$ .

# Weibull model

- An alternative way is to select  $\lambda$  and  $\gamma$  to match the survival data at two specified time points.
- This approach is motivated by the linear relationship between  $\log[-\log S(t)]$  and  $\log(t)$ .
- Suppose we have  $(t_1, s_1)$ , and  $(t_2, s_2)$  that are two time points on a estimated survival curve (e.g., set  $s_i = \hat{S}_{KM}(t_i)$  for  $i = 1, 2$ ).
- Then  $\hat{\lambda}$  and  $\hat{\gamma}$  can be obtained by solving the system of equation

$$\begin{cases} \log(-\log s_1) = \log(\lambda) - \gamma \log(t_1) \\ \log(-\log s_2) = \log(\lambda) - \gamma \log(t_2) \end{cases},$$

for  $\lambda$  and  $\gamma$ .

# Weibull model

- The `Weibull2` function in the **Hmisc** package can be used to produce a Weibull function that matches the two points  $(t_1, s_1)$ , and  $(t_2, s_2)$ .

- Recall that  $\hat{S}_{KM}(t)$  for the `whas100` can be obtained with the `survfit`:

```
> km <- survfit(Surv(lenfol, fstat) ~ 1, whas100)
```

- Suppose we want to find a Weibull distribution that matches the KM estimator at the 1st and the 6th year ( $t = 365$  and  $t = 2190$ ).

```
> summary(km, time = c(365, 2190))
```

```
Call: survfit(formula = Surv(lenfol, fstat) ~ 1, data = whas100)
```

time	n.risk	n.event	survival	std.err	lower	95% CI	upper	95% CI
365	80	20	0.800	0.0400		0.725		0.882
2190	15	27	0.505	0.0537		0.410		0.622

- The two points are  $(365, 0.8)$  and  $(2190, 0.505)$ .

# Weibull model

- Matching the two points with

```
> weiSurv <- Weibull12(c(365, 2190), c(.8, .505))  
> str(weiSurv)  
function (times = NULL, alpha = 0.00560289516761242, gamma = 0.624507785991088)  
> class(weiSurv)  
[1] "function"
```

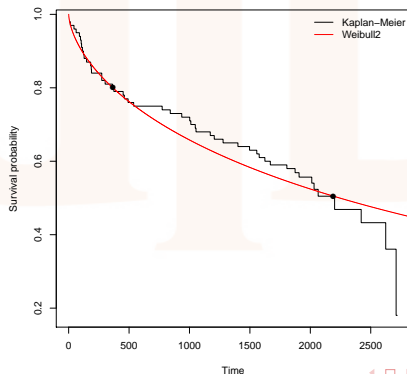
- Weibull12 returns a function.
- The parameters are  $\lambda = 0.006$  and  $\gamma = 0.625$ .
- The “alpha” used in Weibull12 is equivalent to our  $\lambda$ .



# Weibull model

- `weiSurv` returns survival probability depending on inputs.

```
> weiSurv(365)
[1] 0.8
> weiSurv(2190)
[1] 0.505
> weiSurv(0:13)
[1] 1.0000000 0.9944128 0.9913993 0.9889347 0.9867714 0.9848084 0.9829920
[8] 0.9812894 0.9796788 0.9781448 0.9766759 0.9752633 0.9739001 0.9725807
```



# Exponential regression model

- Recall that if  $T$  follows an exponential distribution with rate  $\lambda$ ,  $T$  has the hazard function  $h(t) = \lambda$ .
- If one wants to construct a regression model under the exponential assumption, it is natural to model the exponential parameter  $\lambda$ .
- Suppose a covariates vector  $X = (X_1, \dots, X_p)'$  is available for an individual.
- The hazard at time  $t$  for an individual can be written as

$$\lambda(t; x) = \lambda \cdot r(X' \beta),$$

where  $\beta = (\beta_1, \dots, \beta_p)'$  is the regression coefficient,  $\lambda$  is a constant, and  $r(\cdot)$  is a specified functional form.

# Exponential regression model

- A few choices of  $r(\cdot)$  have been proposed:
  - 1  $r(u) = u$
  - 2  $r(u) = u^{-1}$
  - 3  $r(u) = e^u$
- The first two forms suffer from the disadvantage that  $\beta$  must be restricted to guarantee  $r(X'\beta) > 0$  for all possible  $X$ .
- The third form is commonly considered and will be used here.

# Exponential regression model

- A few choices of  $r(\cdot)$  have been proposed:
  - 1  $r(u) = u$
  - 2  $r(u) = u^{-1}$
  - 3  $r(u) = e^u$
- The first two forms suffer from the disadvantage that  $\beta$  must be restricted to guarantee  $r(X'\beta) > 0$  for all possible  $X$ .
- The third form is commonly considered and will be used here, but we should keep in mind that there may be more appropriate forms in specific settings.

# Exponential regression model

- Working with model with hazard function

$$\lambda(t; \mathbf{x}) = \lambda e^{\mathbf{x}'\beta}. \quad (2)$$

- This model specifies that log failure rate is a linear function of  $X$ .
- Setting  $Y = \log(T)$ , the model (2) implies

$$Y = \alpha - \mathbf{X}'\beta + \epsilon, \quad (3)$$

where  $\alpha = -\log(\lambda)$  and  $\epsilon$  follows an extreme value distribution.

- The model (3) is a log-linear model.

# Exponential regression model

- The model (2) also implies

$$S(t; x) = \lambda t e^{x' \beta},$$

and the conditional density function of  $T$  given  $X$  is then

$$f(t; x) = \lambda e^{x' \beta} \cdot e^{\{-\lambda t e^{x' \beta}\}}.$$

- Parameters can be solved by maximizing the likelihood.

# Weibull regression model

- The similar idea can be applied to Weibull assumption.

$$\lambda(t; \mathbf{x}) = \lambda \gamma t^{\gamma-1} e^{\mathbf{x}'\beta}. \quad (4)$$

- Setting  $Y = \log(T)$ , model 4 implies

$$Y = \alpha - \mathbf{X}'\beta^* + \sigma\epsilon, \quad (5)$$

where  $\alpha = -\log(\lambda)/\gamma$ ,  $\beta^* = \beta/\gamma$ ,  $\sigma = 1/\gamma$ , and  $\epsilon$  is an extreme value distribution.

- The (Weibull) relationship suggests the effect of the covariates
  - act multiplicatively on the hazard function.
  - act additively on  $Y$ ; the general model has a log-linear models.
- The conditional density function and the survival function can be derived for likelihood estimation

# Weibull regression model

- The `survreg` function in **survival** package covers a large family of parametric models.

```
> library(survival)
> args(survreg)
function (formula, data, weights, subset, na.action, dist = "weibull",
        init = NULL, scale = 0, control, parms = NULL, model = FALSE,
        x = FALSE, y = TRUE, robust = FALSE, score = FALSE, ...)
NULL
```

- Suppose we want to fit a parametric model using the `whas100` data with covariates:
  - gender
  - age
  - gender-age interaction
  - body mass index (BMI)
- We can create a `Surv` formula as

```
> fm <- Surv(lenfol, fstat) ~ (age + gender)^2 + bmi
```



# Weibull regression model

- Exponential regression model:

```
> summary(survreg(fm, data = whas100, dist = "exp"))
```

Call:

```
survreg(formula = fm, data = whas100, dist = "exp")
```

	Value	Std. Error	z	p
(Intercept)	9.2897	1.6200	5.73	9.8e-09
age	-0.0532	0.0157	-3.39	0.0007
gender	-3.9324	1.8098	-2.17	0.0298
bmi	0.0935	0.0376	2.49	0.0128
age:gender	0.0498	0.0241	2.06	0.0394

Scale fixed at 1

Exponential distribution

Loglik(model)= -444.4    Loglik(intercept only)= -458.5

Chisq= 28.25 on 4 degrees of freedom, p= 1.1e-05

Number of Newton-Raphson Iterations: 5

n= 100

- The same result is presented in Table 8.2.

# Weibull regression model

- Weibull regression model:

```
> summary(survreg(fm, data = whas100))
```

Call:

```
survreg(formula = fm, data = whas100)
```

	Value	Std. Error	z	p
(Intercept)	9.8727	2.0470	4.82	1.4e-06
age	-0.0639	0.0206	-3.10	0.0019
gender	-4.6895	2.2848	-2.05	0.0401
bmi	0.1055	0.0465	2.27	0.0232
age:gender	0.0592	0.0304	1.94	0.0518
Log(scale)	0.2254	0.1242	1.81	0.0695

Scale= 1.25

Weibull distribution

Loglik(model)= -442.6    Loglik(intercept only)= -455.3

Chisq= 25.36 on 4 degrees of freedom, p= 4.3e-05

Number of Newton-Raphson Iterations: 5

n= 100

- The same result is presented in Table 8.5.