Applied survival analysis

Constantin T Yiannoutsos, Ph.D.

January 24, 2016

Contents of today's lecture

Parametric survival analysis

So far, we have focused primarily on nonparametric and semi-parametric approaches to survival analysis, with heavy emphasis on the Cox proportional hazards model:

$$\lambda(t, \mathbf{Z}) = \lambda_0(t) \exp(\beta \mathbf{Z})$$

We used the following estimating approach:

- We estimated $\lambda_0(t)$ nonparametrically, using the Kaplan-Meier estimator, or using the Kalbfleisch/Prentice estimator under the PH assumption
- We estimated β by assuming a linear model between the log HR and covariates, under the PH model

Both estimates were based on maximum likelihood theory.

Reading: for parametric models see Collett.

Reasons for considering a parametric approach

There are several reasons why we should consider some alternative approaches based on parametric models:

- The assumption of proportional hazards might not be appropriate (based on major departures)
- If a parametric model actually holds, then we would probably gain efficiency
- We may want to handle non-standard situations like
  - interval censoring
  - incorporating population mortality
- We may want to make some connections with other familiar approaches (e.g. use of the Poisson likelihood)
- We may want to obtain some estimates for use in designing a future survival study.

A simple start: Exponential Regression

- **Observed data:** $(X_i, \delta_i, \mathbf{Z}_i)$ for individual i,
  $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, ..., Z_{ip})$ represents a set of p covariates.
- **Right censoring:** Assume that $X_i = \min(T_i, U_i)$
- **Survival distribution:** Assume T_i follows an exponential distribution with a parameter λ that depends on $\mathbf{Z}_i$, say $\lambda_i = \Psi(\mathbf{Z}_i)$. Then we can write:

$$T_i \sim \mathit{exponential}(\Psi(\mathbf{Z}_i))$$

Review

First, let's review some facts about the exponential distribution (from our first survival lecture):

$$
\begin{aligned}
f(t) &= \lambda e^{-\lambda t} \qquad\qquad \text{for } t \geq 0 \\
S(t) &= P(T \geq t) = \int_t^\infty f(u)du = e^{-\lambda t} \\
F(t) &= P(T < t) = 1 - e^{-\lambda t} \\
\lambda(t) &= \frac{f(t)}{S(t)} = \lambda \qquad \text{constant hazard!} \\
\Lambda(t) &= \int_0^t \lambda(u)\, du = \int_0^t \lambda\, du = \lambda t
\end{aligned}
$$

Modeling the hazard in exponential regression

Now, we say that λ is a constant *over time t*, but we want to let it depend on the covariate values, so we are setting

$$\lambda_i = \Psi(\mathbf{Z}_i)$$

The hazard rate would therefore be the same for any two individuals with the same covariate values.
Although there are many possible choices for Ψ, one simple and natural choice is:

$$\Psi(\mathbf{Z}_i) = \exp[\beta_0 + Z_{i1}\beta_1 + Z_{i2}\beta_2 + ... + Z_{ip}\beta_p]$$

WHY?

- ensures a positive hazard
- for an individual with $\mathbf{Z} = \mathbf{0}$, the hazard is e^{β_0}.

The model is called **exponential regression** because of the natural generalization from regular linear regression

Exponential regression for the 2-sample case

- Assume we have only a single covariate $\mathbf{Z} = Z$,
  i.e., $(p = 1)$.
  Hazard Rate:
  $$\Psi(\mathbf{Z}_i) = \exp(\beta_0 + Z_i\beta_1)$$
- Define: $\begin{array}{ll} Z_i = 0 & \text{if individual } i \text{ is in group 0} \\ Z_i = 1 & \text{if individual } i \text{ is in group 1} \end{array}$
- **What is the hazard for group 0?**
- **What is the hazard for group 1?**
- **What is the hazard ratio of group 1 to group 0?**
- **What is the interpretation of β_1?**

Likelihood for Exponential Model

Under the assumption of right censored data, each person has one of two possible contributions to the likelihood:

(a) they have an **event** at X_i ($\delta_i = 1$) $\Rightarrow$ contribution is

$$L_i = \underbrace{S(X_i)}_{\text{survive to } X_i} \cdot \underbrace{\lambda(X_i)}_{\text{fail at } X_i} = e^{-\lambda X_i} \lambda$$

(b) they are **censored** at X_i ($\delta_i = 0$) $\Rightarrow$ contribution is

$$L_i = \underbrace{S(X_i)}_{\text{survive to } X_i} = e^{-\lambda X_i}$$

The likelihood for the exponential model (cont'd)

The **likelihood** is the product over all of the individuals:

$$
\begin{aligned}
\mathcal{L} &= \prod_i L_i \\
&= \prod_i \underbrace{\left(\lambda e^{-\lambda X_i}\right)^{\delta_i}}_{\text{events}} \; \underbrace{\left(e^{-\lambda X_i}\right)^{(1-\delta_i)}}_{\text{censorings}} \\
&= \prod_i \lambda^{\delta_i} \; \left(e^{-\lambda X_i}\right)
\end{aligned}
$$

Maximum Likelihood for Exponential

How do we use the likelihood?

- first take the log
- then take the partial derivative with respect to $\boldsymbol{\beta}$
- then set to zero and solve for $\widehat{\boldsymbol{\beta}}$
- this gives us the **maximum likelihood estimators**

Likelihood equations

The log-likelihood is:

$$\begin{aligned}\log \mathcal{L} &= \log\left[\prod_i \lambda^{\delta_i} \quad \left(e^{-\lambda X_i}\right)\right] \\ &= \sum_i [\delta_i \log(\lambda) - \lambda X_i] \\ &= \sum_i [\delta_i \log(\lambda)] - \sum_i \lambda X_i\end{aligned}$$

For the case of exponential regression, we now substitute the hazard $\lambda = \Psi(\mathbf{Z}_i)$ in the above log-likelihood:

$$\log \mathcal{L} = \sum_i [\delta_i \log(\Psi(\mathbf{Z}_i))] - \sum_i \Psi(\mathbf{Z}_i) X_i \tag{1}$$

General Form of Log-likelihood for Right Censored Data

In general, whenever we have right censored data, the likelihood and corresponding log likelihood will have the following forms:

$$\begin{aligned}\mathcal{L} &= \prod_i \left[\lambda_i(X_i)\right]^{\delta_i} S_i(X_i) \\ \log \mathcal{L} &= \sum_i \left[\delta_i \log\left(\lambda_i(X_i)\right)\right] - \sum_i \Lambda_i(X_i)\end{aligned}$$

where

- $\lambda_i(X_i)$ is the hazard for the individual i who fails at X_i
- $\Lambda_i(X_i)$ is the cumulative hazard for an individual at their failure <u>or</u> censoring time

For example, see the derivation of the likelihood for a Cox model on p.11-18 of Lecture 4 notes. We started with the likelihood above, then substituted the specific forms for $\lambda(X_i)$ under the PH assumption.

Consider our model for the hazard rate:

$$\lambda = \Psi(\mathbf{Z}_i) = \exp[\beta_0 + Z_{i1}\beta_1 + Z_{i2}\beta_2 + ... + Z_{ip}\beta_p]$$

We can write this using vector notation, as follows:

$$\begin{aligned} \text{Let} \quad \mathbf{Z}_i &= (1, Z_{i1}, ...Z_{ip})^T \\ \text{and} \quad \boldsymbol{\beta} &= (\beta_0, \beta_1, ...\beta_p) \end{aligned}$$

(Since β_0 is the intercept (i.e., the log hazard rate for the baseline group), we put a "1" as the first term in the vector $\mathbf{Z}_i$.) Then, we can write the hazard as:

$$\Psi(\mathbf{Z}_i) = \exp[\boldsymbol{\beta}\mathbf{Z}_i]$$

Now we can substitute $\Psi(\mathbf{Z}_i) = \exp[\boldsymbol{\beta}\mathbf{Z}_i]$ in the log-likelihood shown in (1):

$$\log \mathcal{L} = \sum_{i=1}^{n} \delta_i(\boldsymbol{\beta}\mathbf{Z}_i) - \sum_{i=1}^{n} X_i \exp(\boldsymbol{\beta}\mathbf{Z}_i)$$

Score Equations

Taking the derivative with respect to β_0, the score equation is:

$$\frac{\partial \log \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n} [\delta_i - X_i \exp(\boldsymbol{\beta} \mathbf{Z}_i)]$$

For β_k, $k = 1, ...p$, the equations are:

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \beta_k} &= \sum_{i=1}^{n} [\delta_i Z_{ik} - X_i Z_{ik} \exp(\boldsymbol{\beta} \mathbf{Z}_i)] \\ &= \sum_{i=1}^{n} Z_{ik} [\delta_i - X_i \exp(\boldsymbol{\beta} \mathbf{Z}_i)] \end{aligned}$$

To find the MLE's, we set the above equations to 0 and solve (simultaneously). The equations above imply that the MLE's are obtained by setting the weighted number of failures ($\sum_i Z_{ik} \delta_i$) equal to the weighted cumulative hazard ($\sum_i Z_{ik} \Lambda(X_i)$).

Variance of the MLE

To find the variance of the MLE's, we need to take the second derivatives:

$$-\frac{\partial^2 \log \mathcal{L}}{\partial \beta_k \partial \beta_j} = \sum_{i=1}^{n} Z_{ik} Z_{ij} X_i \exp(\boldsymbol{\beta} \mathbf{Z}_i)$$

Some algebra (see Cox and Oakes section 6.2) reveals that

$$Var(\widehat{\boldsymbol{\beta}}) = I(\boldsymbol{\beta})^{-1} = \left[\mathbf{Z}(\mathbf{I} - \Pi)\mathbf{Z}^T\right]^{-1}$$

where

- $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ is a $(p+1) \times n$ matrix
  (p covariates plus the "1" for the intercept β_0)
- $\Pi = diag(\pi_1, \ldots, \pi_n)$ (this means that Π is a diagonal matrix, with the terms $\pi_1, \ldots, \pi_n$ on the diagonal)
- π_i is the probability that the i-th person is censored, so $(1 - \pi_i)$ is the probability that they failed.
- **Note:** The information $I(\boldsymbol{\beta})$ (inverse of the variance) is proportional to the number of failures, not the sample size. This will be important when we talk about study design.

The Single Sample Problem ($Z_i = 1$ for everyone)

First, what is the MLE of β_0?

We set $\frac{\partial \log \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n}[\delta_i - X_i \exp(\beta_0 Z_i)]$ equal to 0 and solve:

$$\begin{aligned}
\Rightarrow \quad \sum_{i=1}^{n} \delta_i &= \sum_{i=1}^{n} [X_i \exp(\beta_0)] \\
d &= \exp(\beta_0) \sum_{i=1}^{n} X_i \\
\exp(\widehat{\beta}_0) &= \frac{d}{\sum_{i=1}^{n} X_i} \\
\hat{\lambda} &= \frac{d}{t}
\end{aligned}$$

where d is the total number of deaths (or events), and $t = \sum X_i$ is the total person-time contributed by all individuals.

MLE estmate for β

If d/t is the MLE for λ, what does this imply about the MLE of β_0?

Using the previous formula $Var(\hat{\boldsymbol{\beta}}) = \left[\mathbf{Z}(\mathbf{I} - \Pi)\mathbf{Z}^T\right]^{-1}$,
what is the variance of $\widehat{\beta}_0$?:

With some matrix algebra, you can show that it is:

$$Var(\widehat{\beta}_0) = \frac{1}{\sum_{i=1}^{n}(1-\pi_i)} = \frac{1}{d}$$

What about $\hat{\lambda} = e^{\hat{\beta}_0}$?

By the delta method,

$$
\begin{aligned}
Var(\hat{\lambda}) &= \hat{\lambda}^2 \ Var(\widehat{\beta}_0) \\
&= \ ?
\end{aligned}
$$

The Two-Sample Problem:

	Z_i	Subjects	Events	Follow-up
Group 0:	$Z_i = 0$	n_0	d_0	$t_0 = \sum_{i=1}^{n_0} X_i$
Group 1:	$Z_i = 1$	n_1	d_1	$t_1 = \sum_{i=1}^{n_1} X_i$

The log-likelihood

$$\log \mathcal{L} = \sum_{i=1}^{n} \delta_i(\beta_0 + \beta_1 Z_i) - \sum_{i=1}^{n} X_i \exp(\beta_0 + \beta_1 Z_i)$$

so
$$\frac{\partial \log \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n} [\delta_i - X_i \exp(\beta_0 + \beta_1 Z_i)]$$
$$= (d_0 + d_1) - (t_0 e^{\beta_0} + t_1 e^{\beta_0 + \beta_1})$$

$$\frac{\partial \log \mathcal{L}}{\partial \beta_1} = \sum_{i=1}^{n} Z_i [\delta_i - X_i \exp(\beta_0 + \beta_1 Z_i)]$$
$$= d_1 - t_1 e^{\beta_0 + \beta_1}$$

This implies:

$$
\begin{aligned}
\hat{\lambda}_1 &= e^{\hat{\beta}_0+\hat{\beta}_1} = ? \\
\hat{\lambda}_0 &= e^{\hat{\beta}_0} = ? \\
\hat{\beta}_0 &= ? \\
\hat{\beta}_1 &= ?
\end{aligned}
$$

The maximum likelihood estimates (MLE's) of the hazard rates under the exponential model are the number of events divided by the person-years of follow-up!

(this result will be relied on heavily when we discuss study design)

Regression: Means and Medians

Mean Survival Time

For the exponential distribution, $E(T) = 1/\lambda$.

- **Control Group:**

$$\overline{T}_0 \;=\; 1/\hat{\lambda}_0 \;=\; 1/\exp(\hat{\beta}_0)$$

- **Treatment Group:**

$$\overline{T}_1 \;=\; 1/\hat{\lambda}_1 \;=\; 1/\exp(\hat{\beta}_0 + \hat{\beta}_1)$$

Means and medians (cont'd)

Median Survival Time

This is the value M at which $S(t) = e^{-\lambda t} = 0.5$, so $M = \text{median} = \frac{-\log(0.5)}{\lambda}$

- **Control Group:**

$$\hat{M}_0 = \frac{-\log(0.5)}{\hat{\lambda}_0} = \frac{-\log(0.5)}{\exp(\hat{\beta}_0)}$$

- **Treatment Group:**

$$\hat{M}_1 = \frac{-\log(0.5)}{\hat{\lambda}_1} = \frac{-\log(0.5)}{\exp(\hat{\beta}_0 + \hat{\beta}_1)}$$

Exponential Regression: Variance Estimates and Test Statistics

We can also calculate the variances of the MLE's as simple functions of the number of failures:

$$
\begin{aligned}
var(\hat{\beta}_0) &= \frac{1}{d_0} \\
var(\hat{\beta}_1) &= \frac{1}{d_0} + \frac{1}{d_1}
\end{aligned}
$$

Inference

So our test statistics are formed as:

For testing $H_o : \beta_0 = 0$:

$$\begin{aligned}\chi^2_w &= \frac{\left(\hat{\beta}_0\right)^2}{var(\hat{\beta}_0)} \\ &= \frac{[\log(d_0/t_0)]^2}{1/d_0}\end{aligned}$$

For testing $H_o : \beta_1 = 0$:

$$\begin{aligned}\chi^2_w &= \frac{\left(\hat{\beta}_1\right)^2}{var(\hat{\beta}_1)} \\ &= \frac{\left[\log(\frac{d_1/t_1}{d_0/t_0})\right]^2}{\frac{1}{d_0}+\frac{1}{d_1}}\end{aligned}$$

How would we form confidence intervals for the hazard ratio?

The Likelihood Ratio test statistic

This is an alternative to the Wald test. It is based on 2 times the log of the ratio of the likelihoods under the null and alternative. We reject H_0 if 2 log(LR) $> \chi^2_{1,0.05}$, where

$$LR = \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} = \frac{\mathcal{L}(\widehat{\lambda}_0, \widehat{\lambda}_1)}{\mathcal{L}(\widehat{\lambda})}$$

The Likelihood Ratio test statistic (cont'd)

For a sample of n independent exponential random variables with parameter λ, the Likelihood is:

$$\begin{aligned} L &= \prod_{i=1}^{n}[\lambda^{\delta_i}\exp(-\lambda x_i)] \\ &= \lambda^d \exp(-\lambda \sum x_i) \\ &= \lambda^d \exp(-\lambda n\bar{x}) \end{aligned}$$

where d is the number of deaths or failures. The log-likelihood is

$$\ell = d\log(\lambda) - \lambda n\bar{x}$$

and the MLE is

$$\widehat{\lambda} = d/(n\bar{x})$$

2-Sample Case: LR test calculations

Data:

Group 0: d_0 failures among the n_0 females
mean failure time is $\bar{x}_0 = (\sum_i^{n_0} X_i)/n_0$

Group 1: d_1 failures among the n_1 males
mean failure time is $\bar{x}_1 = (\sum_i^{n_1} X_i)/n_1$

Under the **alternative hypothesis**:

$$\begin{aligned}
\mathcal{L} &= \lambda_1^{d_1} \exp(-\lambda_1 n_1 \bar{x}_1) \times \lambda_0^{d_0} \exp(-\lambda_0 n_0 \bar{x}_0) \\
\log(\mathcal{L}) &= d_1 \log(\lambda_1) - \lambda_1 n_1 \bar{x}_1 + d_0 \log(\lambda_0) - \lambda_0 n_0 \bar{x}_0
\end{aligned}$$

The MLE's are:

$$\begin{aligned}
\widehat{\lambda}_1 &= d_1/(n_1 \bar{x}_1) && \text{for males} \\
\widehat{\lambda}_0 &= d_0/(n_0 \bar{x}_0) && \text{for females}
\end{aligned}$$

MLEs under the null hypothesis

$$\begin{aligned}\mathcal{L} &= \lambda^{d_1+d_0}\exp[-\lambda(n_1\bar{x}_1 + n_0\bar{x}_0)] \\ \log(\mathcal{L}) &= (d_1+d_0)\log(\lambda) - \lambda[n_1\bar{x}_1 + n_0\bar{x}_0]\end{aligned}$$

The corresponding MLE is

$$\widehat{\lambda} = (d_1+d_0)/[n_1\bar{x}_1 + n_0\bar{x}_0]$$

Constructing the LR test

A likelihood ratio test can be constructed by taking twice the difference of the log-likelihoods under the alternative and the null hypotheses:

$$-2\left[(d_0+d_1)\log\left(\frac{d_0+d_1}{t_0+t_1}\right)-d_1\log[d_1/t_1]-d_0\log[d_0/t_0]\right]$$

Nursing home example

For the females:

- $n_0 = 1173$
- $d_0 = 902$
- $t_0 = 310754$
- $\bar{x}_0 = 265$

For the males:

- $n_1 = 418$
- $d_1 = 367$
- $t_1 = 75457$
- $\bar{x}_1 = 181$

Plugging these values in, we get a LR test statistic of 64.20.

Hand Calculations using events and follow-up:

By adding up "LOS" for males to get t_1 and for females to get t_0, I obtained:

- $d_0 = 902$ (females)
  $d_1 = 367$ (males)
- $t_0 = 310754$ (female follow-up)
  $t_1 = 75457$ (male follow-up)

- This yields an estimated log HR:

$$\hat{\beta}_1 = \log\left[\frac{d1/t1}{d0/t0}\right] = \log\left[\frac{367/75457}{902/310754}\right] = \log(1.6756) = 0.5162$$

Constructing the Wald test

In the above calculations, the estimated standard error is:

$$\sqrt{var(\hat{\beta}_1)} \;=\; \sqrt{\frac{1}{d_1}+\frac{1}{d_0}} = \sqrt{\frac{1}{902}+\frac{1}{367}} \;=\; 0.06192$$

So the Wald test becomes:

$$\chi^2_W \;=\; \frac{\hat{\beta}_1^2}{var(\hat{\beta}_1)} = \frac{(0.51619)^2}{0.061915} \;=\; 69.51$$

We can also calculate $\hat{\beta}_0 = log(d_0/t_0) = -5.842$,
along with its standard error $se(\hat{\beta}_0) = \sqrt{(1/d0)} = 0.0333$

Exponential Regression in R

```
Call:
survreg(formula = Surv(losyr, fail) ~ gender, data = nurshome,
    dist = "exp")
              Value Std. Error     z        p
(Intercept) -0.0578     0.0333 -1.73 8.28e-02
gender      -0.5162     0.0619 -8.34 7.62e-17

Scale fixed at 1

Exponential distribution
Loglik(model)= -1006.3   Loglik(intercept only)= -1038.4
Chisq= 64.2 on 1 degrees of freedom, p= 1.1e-15
Number of Newton-Raphson Iterations: 5
n= 1591
```

Since $Z = 8.337$, the chi-square test is $Z^2 = 69.51$.

The Weibull regression model

At the beginning of the course, we saw that the survivorship function for a Weibull random variable is:

$$S(t) \quad = \quad \exp[-\lambda(t^{\kappa})]$$

and the hazard function is:

$$\lambda(t) \quad = \quad \kappa\,\lambda\,t^{(\kappa-1)}$$

The Weibull regression model assumes that for someone with covariates $\mathbf{Z}_i$, the survivorship function is

$$S(t;\mathbf{Z}_i) = \exp[-\Psi(\mathbf{Z}_i)(t^{\kappa})]$$

where $\Psi(\mathbf{Z}_i)$ is defined as in exponential regression to be:

$$\Psi(\mathbf{Z}_i) = \exp[\beta_0 + Z_{i1}\beta_1 + Z_{i2}\beta_2 + ...Z_{ip}\beta_p]$$

For the 2-sample problem, we have:

$$\Psi(\mathbf{Z}_i) = \exp[\beta_0 + Z_{i1}\beta_1]$$

Weibull MLEs for the 2-sample problem:

Log-likelihood:

$$\log \mathcal{L} = \sum_{i=1}^{n} \delta_i \log \left[\kappa \exp(\beta_0 + \beta_1 Z_i) X_i^{\kappa-1}\right] - \sum_{i=1}^{n} X_i^{\kappa} \exp(\beta_0 + \beta_1 Z_i)$$

$$\Rightarrow \quad \exp(\hat{\beta}_0) = d_0/t_0\kappa \qquad \exp(\hat{\beta}_0 + \hat{\beta}_1) = d_1/t_1\kappa$$

where

$$t_{j\kappa} = \sum_{i=1}^{n_j} X_i^{\hat{\kappa}} \text{ among } n_j \text{ subjects}$$

$$\hat{\lambda}_0(t) = \hat{\kappa}\ \exp(\hat{\beta}_0)\ t^{\hat{\kappa}-1} \quad \hat{\lambda}_1(t) = \hat{\kappa}\ \exp(\hat{\beta}_0 + \hat{\beta}_1)\ t^{\hat{\kappa}-1}$$

$$\begin{aligned} \widehat{HR} &= \hat{\lambda}_1(t)/\hat{\lambda}_0(t) = \exp(\hat{\beta}_1) \\ &= \exp\left(\frac{d_1/t_1\kappa}{d_0/t_0\kappa}\right) \end{aligned}$$

Weibull Regression: Means and Medians

Mean Survival Time

For the Weibull distribution, $E(T) = \lambda^{(-1/\kappa)}\Gamma[(1/\kappa)+1]$.

- **Control Group:**

$$\overline{T}_0 = \hat{\lambda}_0^{(-1/\hat{\kappa})}\,\Gamma[(1/\hat{\kappa})+1]$$

- **Treatment Group:**

$$\overline{T}_1 = \hat{\lambda}_1^{(-1/\hat{\kappa})}\,\Gamma[(1/\hat{\kappa})+1]$$

Median Survival Time

For the Weibull distribution, $M = \text{median} = \left[\frac{-\log(0.5)}{\lambda}\right]^{1/\kappa}$

- **Control Group:**

$$\hat{M}_0 = \left[\frac{-\log(0.5)}{\hat{\lambda}_0}\right]^{1/\hat{\kappa}}$$

- **Treatment Group:**

$$\hat{M}_1 = \left[\frac{-\log(0.5)}{\hat{\lambda}_1}\right]^{1/\hat{\kappa}}$$

where $\hat{\lambda}_0 = \exp(\hat{\beta}_0)$ and $\hat{\lambda}_1 = \exp(\hat{\beta}_0 + \hat{\beta}_1)$.

The Gamma function

Note: the symbol Γ is the “gamma” function. If x is an integer, then

$$\Gamma(x) = (x-1)!$$

In cases where x is not an integer, this function has to be evaluated numerically. In homework and labs, I will supply this value to you.

The Weibull regression model is very easy to fit:

- In STATA: Just specify `dist(weibull)` instead of `dist(exp)` within the `streg` command
- In SAS: use model option `dist=weibull` within the `proc lifereg` procedure
- In R: we use the `survreg` command with the `dist="exp"` option.

Note: to get more information on these modeling procedures, use the online help facilities.

Fitting the Weibull model in R

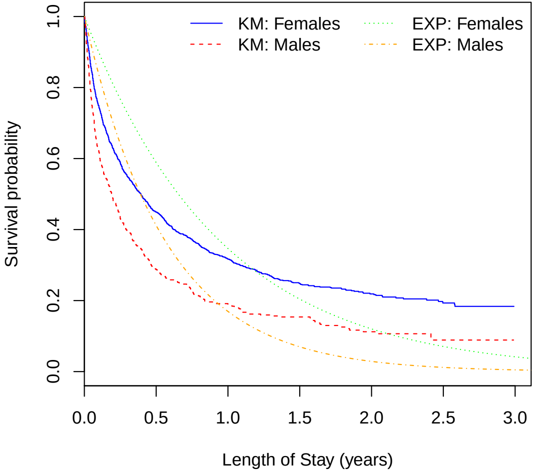
```
Call:
survreg(formula = Surv(losyr, fail) ~ gender, data = nurshome,
    dist = "weibull")
             Value Std. Error     z        p
(Intercept) -0.143     0.0542 -2.65 8.13e-03
gender      -0.673     0.1011 -6.66 2.67e-11
Log(scale)   0.487     0.0232 20.99 8.94e-98

Scale= 1.63

Weibull distribution
Loglik(model)= -731.1   Loglik(intercept only)= -751.9
Chisq= 41.73 on 1 degrees of freedom, p= 1e-10
Number of Newton-Raphson Iterations: 5
n= 1591
```

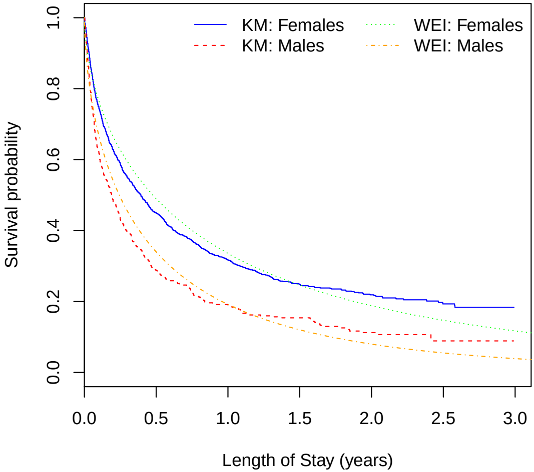
Comparison of Exponential with Kaplan-Meier

We can see how well the Exponential model fits by comparing the survival estimates for males and females under the exponential model, i.e., $P(T \geq t) = e^{(-\hat{\lambda}_z t)}$, to the Kaplan-Meier survival estimates:



Comparison of Weibull with Kaplan-Meier

We can see how well the Weibull model fits by comparing the survival estimates, $P(T \geq t) = e^{(-\hat{\lambda}_z t^{\hat{\kappa}})}$, to the Kaplan-Meier survival estimates.



Other useful plots for evaluating fit

- $-\log(\hat{S}(t))$ vs t
- $\log[-\log(\hat{S}(t))]$ vs $\log(t)$

Why are these useful?
If T is exponential, then $S(t) = \exp(-\lambda t))$

$$\begin{array}{lrcl} \text{so} & \log(S(t)) & = & -\lambda t \\ \text{and} & \Lambda(t) & = & \lambda\, t \end{array}$$

a straight line in t with slope λ and intercept=0

If T is Weibull, then $S(t) = \exp(-(\lambda t)^{\kappa})$

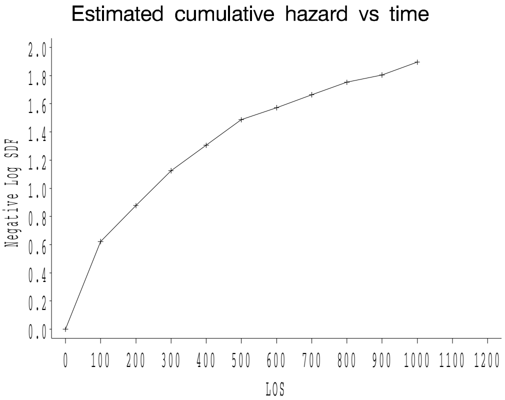
$$\begin{array}{lrcl} \text{so} & \log(S(t)) & = & -\lambda t^{\kappa} \\ \text{then} & \Lambda(t) & = & \lambda t^{\kappa} \\ \text{and} & \log(-\log(S(t))) & = & \log(\lambda) + \kappa * \log(t) \end{array}$$

a straight line in $\log(t)$ with slope κ and intercept $\log(\lambda)$.

Goodness of fit plots

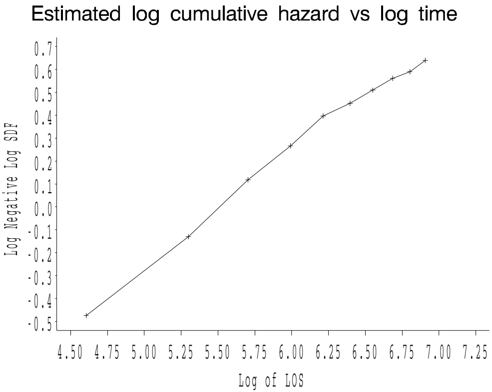
So we can calculate our estimated $\Lambda(t)$ and plot it versus t, and if it seems to form a straight line, then the exponential distribution is probably appropriate for our dataset.

Plots for nursing home data: $\hat{\Lambda}(t)$ vs t



Log-log plot in the Weibull analysis

Or we can plot $\log \hat{\Lambda}(t)$ versus $\log(t)$, and if it seems to form a straight line, then the Weibull distribution is probably appropriate for our dataset.
Plots for nursing home data: $\log[-log(\hat{S}(t))]$ **vs** $\log(t)$



Comparison of methods for the two-sample problem

Data:

	Z_i	Subjects	Events	Follow-up
Group 0:	$Z_i = 0$	n_0	d_0	$t_0 = \sum_{i=1}^{n_0} X_i$
Group 1:	$Z_i = 1$	n_1	d_1	$t_1 = \sum_{i=1}^{n_1} X_i$

In General:

$$\lambda_z(t) \quad = \quad \lambda(t, Z = z) \qquad \text{for } z = 0 \text{ or } 1.$$

The hazard rate depends on the value of the covariate Z. In this case, we are assuming that we only have a single covariate, and it is binary ($Z = 1$ or $Z = 0$)

Reading from Collett

Reference (Collett):

Section(s)	**Description**
4.1.1, 4.1.2	Exponential properties
4.1.3	Weibull properties
4.3.1, 4.4.2	Exponential ML estimation
4.3.2	Weibull ML estimation
4.5	General Weibull regression
4.6	Model selection - Weibull regression
4.7	Weibull/AFT model connection
Ch.6	AFT - Other parametric models

Models

Exponential Regression:

$$
\begin{aligned}
\lambda_z(t) &= \exp(\beta_0 + \beta_1 Z) \\
\Rightarrow \lambda_0 &= \exp(\beta_0) \\
\lambda_1 &= \exp(\beta_0 + \beta_1) \\
HR &= \exp(\beta_1)
\end{aligned}
$$

Weibull Regression:

$$
\begin{aligned}
\lambda_z(t) &= \kappa \ \exp(\beta_0 + \beta_1 Z) \ t^{\kappa-1} \\
\Rightarrow \lambda_0 &= \kappa \ \exp(\beta_0) \ t^{\kappa-1} \\
\lambda_1 &= \kappa \ \exp(\beta_0 + \beta_1) \ t^{\kappa-1} \\
HR &= \exp(\beta_1)
\end{aligned}
$$

Models (cont'd)

Proportional Hazards Model:

$$\begin{aligned} \lambda_z(t) &= \lambda_0(t)\exp(\beta_1) \\ \Rightarrow \lambda_0 &= \lambda_0(t) \\ \lambda_1 &= \lambda_0(t)\exp(\beta_1) \\ HR &= \exp(\beta_1) \end{aligned}$$

KM?

Remarks

We make the following remarks:

- Exponential model is a special case of the Weibull model with $\kappa = 1$ (note: Collett uses γ instead of κ)
- Exponential and Weibull models are both special cases of the Cox PH model. **How can you show this?**
- If either the exponential model or the Weibull model is valid, then these models will tend to be more efficient than PH (smaller s.e.'s of estimates). This is because they assume a particular form for $\lambda_0(t)$, rather than estimating it at every death time.

Exponential regression

For the Exponential model, the hazards are constant over time, given the value of the covariate Z_i:

$$
\begin{aligned}
Z_i = 0 \Rightarrow \quad \hat{\lambda}_0 &= \exp(\hat{\beta}_0) \\
Z_i = 1 \Rightarrow \quad \hat{\lambda}_0 &= \exp(\hat{\beta}_0 + \hat{\beta}_1)
\end{aligned}
$$

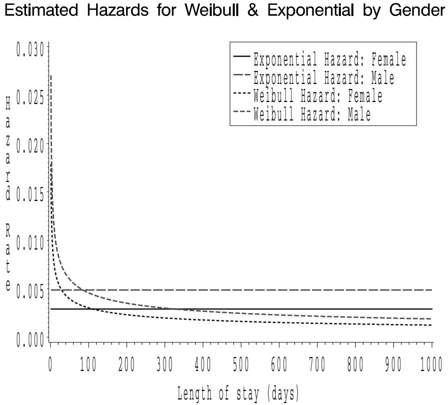
For the Weibull model, we have to estimate the hazard as a function of time, given the estimates of β_0, β_1 and κ:

$$
\begin{aligned}
Z_i = 0 \Rightarrow \quad \hat{\lambda}_0(t) &= \hat{\kappa}\ \exp(\hat{\beta}_0)\ t^{\hat{\kappa}-1} \\
Z_i = 1 \Rightarrow \quad \hat{\lambda}_1(t) &= \hat{\kappa}\ \exp(\hat{\beta}_0 + \hat{\beta}_1)\ t^{\hat{\kappa}-1}
\end{aligned}
$$

However, the ratio of the hazards is still just $\exp(\hat{\beta}_1)$, since the other terms cancel out.

Estimated hazards for the nursing home data

Here's what the estimated hazards look like for the nursing home data:



Proportional Hazards Model

To get the MLE's for this model, we have to maximize the Cox partial likelihood iteratively. There are not closed form expressions like above.

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \left[\frac{e^{\boldsymbol{\beta}\mathbf{Z}_i}}{\sum_{\ell \in \mathcal{R}(X_i)} e^{\boldsymbol{\beta}\mathbf{Z}_\ell}} \right]^{\delta_i} \\
&= \prod_{i=1}^{n} \left[\frac{e^{\beta_0 + \beta_1 Z_i}}{\sum_{\ell \in \mathcal{R}(X_i)} e^{\beta_0 + \beta_1 Z_\ell}} \right]^{\delta_i}
\end{aligned}
$$

Comparison with Proportional Hazards Model

```
Call:
coxph(formula = Surv(losyr, fail) ~ gender, data = nurshome)

  n= 1591, number of events= 1269

         coef exp(coef) se(coef)     z Pr(>|z|)
gender 0.3958    1.4855   0.0621 6.373 1.85e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

       exp(coef) exp(-coef) lower .95 upper .95
gender     1.486     0.6732     1.315     1.678

Concordance= 0.541  (se = 0.006 )
Rsquare= 0.024   (max possible= 1 )
Likelihood ratio test= 38.29  on 1 df,   p=6.11e-10
Wald test            = 40.62  on 1 df,   p=1.852e-10
Score (logrank) test = 41.14  on 1 df,   p=1.415e-10
```

For the PH model, $\hat{\beta}_1 = 0.394$ and $\widehat{HR} = e^{0.394} = 1.483$.

Comparison with the Logrank test

```
Call:
survdiff(formula = Surv(losyr, fail) ~ gender, data = nurshome)

            N Observed Expected (O-E)^2/E (O-E)^2/V
gender=0 1173      902      995      8.76      41.1
gender=1  418      367      274     31.88      41.1

 Chisq= 41.1  on 1 degrees of freedom, p= 1.46e-10
```

Comparison with the Wilcoxon test

Note that this is fit by adding `rho=1` in R:

```
Call:
survdiff(formula = Surv(losyr, fail) ~ gender, data = nurshome,
    rho = 1)

            N Observed Expected (O-E)^2/E (O-E)^2/V
gender=0 1173      529      592      6.66      41.8
gender=1  418      236      173     22.73      41.8

 Chisq= 41.8  on 1 degrees of freedom, p= 9.94e-11
```

Comparison of HRs and test statistics for effect of `gender`

Model/Method	λ_0	λ_1	HR	log(HR)	se(log HR)	Wald Statistic
Exponential	0.0029	0.0049	1.676	0.5162	0.0619	69.507
Weibull						
$t = 50$	0.0040	0.0060	1.513	0.4138	0.0636	42.381
$t = 100$	0.0030	0.0046	1.513			
$t = 500$	0.0016	0.0025	1.513			
Logrank						41.085
Wilcoxon						41.468
Cox PH						
Ties=Breslow			1.483	0.3944	0.0621	40.327
Ties=Discrete			1.487	0.3969	0.0623	40.565
Ties=Efron			1.486	0.3958	0.0621	40.616
Ties=Exact			1.486	0.3958	0.0621	40.617
Score (Discrete)						41.085

Comparison of Mean and Median Survival Times by Gender

	Mean Survival		Median Survival	
Model/Method	Female	Male	Female	Male
Exponential	344.5	205.6	238.8	142.5
Weibull	461.6	235.4	174.2	88.8
Kaplan-Meier	318.6	200.7	144	70
Cox PH (Kalbfleisch/Prentice)			131	72