Survival and Hazard Functions

Let $T$ be life time with p.d.f. $f(t)$ and c.d.f. $F(t)$. The survival function $S(t)$ and hazard function $\lambda(t)$ are given by

\[ S(t) = P(T > t) = 1 - F(t), \]
\[ \lambda(t) = P(t \leq T < t + dt \mid T \geq t)/dt = f(t)/S(t) = -d\log S(t)/dt. \]

In terms of the hazard function, $S(t)$ can be written as

\[ S(t) = \exp\{-\int_0^t \lambda(u)du\} = \exp\{-\Lambda(t)\}, \]

where $\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t)$ is the cumulative hazard.

For $T$ exponential with $f(t) = e^{-t/\beta}/\beta$, $S(t) = e^{-t/\beta}$, $\Lambda(t) = t/\beta$, and $\lambda(t) = 1/\beta$.

Weibull and Related Distributions

The Weibull distribution is the more popular parametric family used in survival analysis. Set $\Lambda(t) = (t/\beta)^{\alpha}$. It follows that

\[ \lambda(t) = (\alpha/\beta)(t/\beta)^{\alpha-1}, \]
\[ S(t) = e^{-(t/\beta)^\alpha}, \]
\[ f(t) = (\alpha/\beta)(t/\beta)^{\alpha-1}e^{-(t/\beta)^\alpha}. \]

When $\alpha = 1$, one has the exponential distribution.

For $T$ Weibull, $\tilde{T} = \log T$ follows the extreme value distribution with $F(\tilde{t}) = 1 - \exp\{-e^{\alpha(\tilde{t} - \mu)}\}$ where $\mu = \log \beta$, and $f(\tilde{t}) = \alpha e^{\alpha(\tilde{t} - \mu)} \exp\{-e^{\alpha(\tilde{t} - \mu)}\}$. It is a location-scale family.

For $T$ Weibull, $1/T$ follows the Gompertz distribution.
Log-Normal and Log-Logistic Distributions

For $T$ log-normal, or $\tilde{T} = \log T \sim N(\mu, \sigma^2)$, one has

$$S(t) = 1 - \Phi((\tilde{t} - \mu)/\sigma),$$
$$\lambda(t) = \frac{1}{\sigma t} \frac{\phi((\tilde{t} - \mu)/\sigma)}{1 - \Phi((\tilde{t} - \mu)/\sigma)},$$

where $\phi(x)$ and $\Phi(x)$ are the p.d.f. and the c.d.f. of $N(0,1)$.

For $T$ log-logistic, or $\tilde{T} = \log T$ logistic, one has

$$S(t) = \frac{1}{1 + e^{\alpha(\tilde{t} - \mu)}},$$
$$\lambda(t) = \frac{\alpha}{t} \frac{e^{\alpha(\tilde{t} - \mu)}}{1 + e^{\alpha(\tilde{t} - \mu)}}.$$

It is a location-scale family very similar to normal. Roughly, $\alpha^{-1} = 1$ in logistic corresponds to $\sigma = 1.7$ in normal.

Censoring and Truncation

When $T$ is not observed but only $T \geq C$ is known, one has right-censored data. When $C$ is independent of $T$, it is called independent-censoring. When $P(C = c) = 1$, one has the type I censoring. When $C = T_{(r)}$, one has the Type II censoring. Type I is independent but type II is not.

Life time data may also come left-censored or interval-censored.

A subject may enter the “risk set” at time $V \geq 0$, resulting in left-truncation.

Typical life time data are observed with right-censoring and possibly also with left-truncation, $(V, X, \delta)$, where $X = \min(T, C)$ and $\delta = I_{[T=X]}$. Some covariate $Z$ may also be observed.
Censoring and Truncation: Example

Consider the survival information of 97 male members in the Channing House retirement community, Palo Alto, California. $V$ is the age of the person at admission to the community, $C$ is the age at the end of the study or at the person’s withdrawal from the community, and $T$ is the age at death, all in months.

```r
ind<-sample(97,20); delta<-chan$Death[ind]
V<-chan$Entry[ind]; X<-chan$Exit[ind];
plot(V,1:20,pch="|",xlim=c(720,1200))
points(X,1:20,pch=c("o","x")[delta+1])
for(i in 1:20) lines(c(V[i],X[i]),c(i,i),lty=2)
# risk set size
```

```r
t<-seq(720,1200,len=100)
at.risk<-outer(chan$En,t,"<=")&outer(chan$Ex,t,">")
plot(t,apply(at.risk,2,sum),type="l")
```

Likelihood for Censored Data

For observed $T$, $P(t \leq T < t + dt) = f(t)dt = \lambda(t)S(t)dt$, and for $T \geq C$, $P(T \geq C) = S(C)$, so a right-censored datum $(X, \delta)$ has a likelihood proportional to

$$(\lambda(x)S(x))^\delta S(x)^{1-\delta} = \lambda(x)^\delta S(x).$$

With left-truncation at $V$, the survival function is rescaled to be $S(t)/S(v)$, so a left-truncated and right-censored datum $(V, X, \delta)$ has a likelihood proportional to $\lambda(x)^\delta S(x)/S(v)$.

Consider $(V_i, X_i, \delta_i)$ from an exponential distribution with $\lambda(t) = \lambda$ and $S(t) = e^{-\lambda t}$. The log likelihood is seen to be

$$\log \lambda \sum \delta_i - \lambda \sum (X_i - V_i).$$

The MLE of $\lambda$ is $\hat{\lambda} = \sum \delta_i / \sum (X_i - V_i)$ with s.e. $\hat{\lambda}/\sqrt{\sum \delta_i}$. 

C. Gu 
Spring 2017
Kaplan-Meier Estimate of $S(t)$

Let $t_1 < t_2 < \cdots < t_k$ be observed failure times, with $d_j$ items failing at $t_j$ and $m_j$ items right-censored in $[t_j, t_{j+1})$. The MLE of $S(t)$ is of form $\hat{S}(t_j) = \prod_{l=1}^{j-1} (1 - \hat{\lambda}_l)$ and $\hat{S}(t_j^-) = \prod_{l=1}^{j-1} (1 - \lambda_l)$ that maximizes $L = \prod_{j=1}^{k} \{(S(t_j^-) - S(t_j))^{d_j} S(t_j)^{m_j}\}$. One has

$$L = \prod_{j=1}^{k} \{\lambda_j^{d_j} \prod_{l=1}^{j-1} (1 - \lambda_l)^{d_j} \prod_{l=1}^{j} (1 - \lambda_l)^{m_j}\} = \prod_{j=1}^{k} \lambda_j^{d_j} (1 - \lambda_j)^{n_j - d_j},$$

where $n_j = (d_j + m_j) + \cdots + (d_k + m_k)$ is the risk set size at $t_j$; $L$ is maximized at $\hat{\lambda}_j = d_j/n_j$. So $\hat{S}(t) = \prod_{j:t_j \leq t} (n_j - d_j)/n_j$.

When $m_j = 0$, $\hat{S}(t) = n^{-1} \sum_{i=1}^{n} I_{[T_i > t]}$, reducing to the standard empirical distribution.

Asymptotic Variance of Kaplan-Meier

Note that $\hat{\lambda}_j = d_j/n_j$ is an MLE with $\hat{\sigma}^2(\hat{\lambda}_j) = \hat{\lambda}_j(1 - \hat{\lambda}_j)/n_j$. By the $\delta$-method, $\hat{\sigma}^2[\log(1 - \hat{\lambda}_j)] = \hat{\lambda}_j/(1 - \hat{\lambda}_j) n_j = d_i/n_j (n_j - d_j)$. Thus, $\hat{\sigma}^2[\log \hat{S}(t)] = \sum_{j:t_j \leq t} d_i/n_j (n_j - d_j)$. Again by $\delta$-method,

$$\hat{\sigma}^2[\hat{S}(t)] = \hat{S}(t)^2 \sum_{j:t_j \leq t} d_i/n_j (n_j - d_j),$$

which is known as Greenwood’s formula.

The Kaplan-Meier estimate can be calculated using `survfit` in the R package `survival`.

```r
library(survival); data(leukemia)
kml <- survfit(Surv(time,status)~x, leukemia)
plot(kml); plot(kml[1,])
```

CIs may be on the scale of $S(t)$ (conf.type="plain"), $\Lambda(t)$ (conf.type="log", default), or log $\Lambda(t)$ (conf.type="log-log").
Logrank Test for Equality of Survivals

Consider a $2 \times 2$ table \[ \begin{array}{cc} d_0 & d_1 \\ n_0 - d_0 & n_1 - d_1 \end{array} \] with fixed column totals $n_0$ and $n_1$. To test the homogeneity of the columns, one may use a hypergeometric model conditioning on the row totals $d = d_0 + d_1$ and $n - d = n_0 + n_1 - d$, with $o = d_1$ having mean $e = n_1 d/n$ and variance \[ w = \frac{n_0 n_1}{n-1} \frac{d(n-d)}{n^2}. \] Note that \[ d_1 - n_1 d/n = -(d_0 - n_0 d/n) = (d_1(n_0 - d_0) - d_0(n_1 - d_1))/n. \]

At the observed failure times $t_j$ of the merged sample, one has $o_j$, $e_j$, and $w_j$. Calculating $O = \sum_j o_j$, $E = \sum_j e_j$, and $W = \sum_j w_j$, one may use the logrank statistic $(O - E)^2 / W$ to test equality of survival, which is approximately $\chi^2_1$ under the null.

The test is known as the Mantel-Haenszel test for the assessment of “treatment effect” over different groups of “patients.” Its validity for comparison of survivals is through conditioning arguments.

Logrank Test: Example and Extension

logrank test can be conducted using `survdiff` in `survival`. Mantel-Haenszel test can be done using `mantelhaen.test`.

```r
survdiff(Surv(time,status) ~ x, data=leukemia)
source("tbl.leukemia.R"); tbl.leuk
mantelhaen.test(tbl.leuk[,,-18], cor=FALSE)
```

For more than 2 samples, consider a $2 \times (p + 1)$ table \[ \begin{array}{ccc} d_0 & \ldots & d_p \\ n_0 - d_0 & \ldots & n_p - d_p \end{array} \] with fixed column totals $n_i$. To test the homogeneity of columns, one may use a hypergeometric model with $o_i = d_i$, $e_i = n_i d/n$, $w_{ii} = \text{var}(o_i) = \frac{n_i(n-n_i)}{n-1} \frac{d(n-d)}{n^2}$, $w_{il} = \text{cov}(o_i, o_l) = -\frac{n_i n_l}{n-1} \frac{d(n-d)}{n^2}$, $i = 1, \ldots, p$. Summing over observed failure times $t_j$, one gets vectors $O$ and $E$ and a matrix $W$. The logrank statistic $(O - E)^T W^{-1} (O - E)$ is approximately $\chi^2_p$ under homogeneity.
Weighted and Stratified Logrank Tests

When putting things together over $t_j$, one may add weights to form $O - E = \sum_j g_j (o_j - e_j)$ and $W = \sum_j g_j^2 w_j$. The choice of $g_j = n_j$, where $n_j$ is the risk set size at $t_j$, leads to generalizations of Wilcoxon or Kruskal-Wallis. The choice of $g_j = \{S(t_j^-)\}^\rho$ forms the $G^\rho$ family implemented in `survdiff`.

\[ \text{survdiff(Surv(time,status)~x,data=leukemia,rho=1)} \]

Similar to the original Mantel-Haenszel test, $O - E$ and $W$ can also be aggregated over different strata to accommodate samples from inhomogeneous populations.

\[ \text{data(lung); table(lung$pat.karno); table(lung$inst)} \]
\[ \text{survdiff(Surv(time,status)~pat.karno+strata(inst),lung)} \]

One-Sample Test: PH Alternatives

To compare a cohort to a “standard” population, one needs to test $H_0 : \Lambda(t) = \Lambda_0(t)$ for a known $\Lambda_0(t)$. With “proportional hazard” alternatives of the form $\Lambda(t) = e^\theta \Lambda_0(t)$, one may test for $\theta = 0$.

For $T$ with a cumulative hazard $e^\theta \Lambda_0(t)$, consider a transformation of the time axis, $\tilde{t} = \Lambda_0(t)$. One has

\[ P(\tilde{T} > t) = P(\Lambda_0(T) > t) = P(T > \Lambda_0^{-1}(t)) = e^{-\Lambda(\Lambda_0^{-1}(t))} = e^{-\theta t}. \]

Observing $(X_i, \delta_i)$, the log likelihood is $\theta \sum_i \delta_i - e^\theta \sum_i \Lambda_0(X_i)$, yielding $\hat{\theta} = \log(d/\tilde{e})$, where $d = \sum_i \delta_i$ and $\tilde{e} = \sum_i \Lambda_0(X_i)$, with s.e. $1/\sqrt{d}$ (or $1/\sqrt{\tilde{e}}$). The test statistic is $Z = \sqrt{d}\hat{\theta}$ or $Z^2$.

The null can be “localized” to each individual, $H_0 : \Lambda_i(t) = \Lambda_{0i}(t)$, with alternatives of the form $\Lambda_i(t) = e^\theta \Lambda_{0i}(t)$ for a common $\theta$. The test is of the same form but with $\tilde{e} = \sum_i \Lambda_{0i}(X_i)$. 

C. Gu

Spring 2017
One-Sample Test: Example, Logrank

The log likelihood is the same as a constant Poisson model with $\delta_i$ as the responses and $\log \Lambda_0(X_i)$ as the offset.

```r
lam0 <- rep(c(4,3),c(11,12))
cum.hzd <- leuk$time/exp(lam0)
summary(glm(status~offset(log(cum.hzd)),poisson,leuk))
d<-sum(leuk$status); e<-sum(cum.sum); log(d/e)*sqrt(d)
```

One-sample logrank tests can be derived with $o_i = \delta_i$, $e_i = \Lambda_0(X_i)$, and $w_i = e_i$; censored times also contribute information here, and one may form $O = \sum_i g_i o_i$, $E = \sum_i g_i e_i$, and $W = \sum_i g_i^2 w_i$. When $g_i = 1$, the test is of form $Z^2 = (O - E)^2/E$, where $Z = (d - \bar{e})/\sqrt{\bar{e}}$.

```r
survdiff(Surv(time,status)~offset(exp(-cum.hzd)),leuk)
```

Parametric Models: Location-Scale Families

Parametric regression models are based on location-scale families, $f(x|\mu,\sigma) = f((x - \mu)/\sigma) = f(u)$. Linear models can be specified for the location, $\mu = \beta^T z$.

The basic families are normal with $f(u) \propto \exp\{-u^2/2\}$, extreme value with $f(u) = we^{-w}$ where $w = e^u$, and logistic with $f(u) = w/(1 + w)^2$ where $w = e^u$. A log transform of time reduces log-normal, Weibull, and log-logistic to the 3 basic families.

For right-censored data, let $u_i = (t(x_i) - \beta^T z_i)/\sigma$ where $t(x_i)$ are transformed times, one has the log likelihood

$$L(\beta) = \sum_i \{\delta_i \log f(u_i) + (1 - \delta_i) \log \int_{u_i}^{\infty} f(s)ds\},$$

or in terms of $\lambda(t; z)$, $S(t; z)$,

$$L(\beta) = \sum_i \{\delta_i \log \lambda(x_i; z_i) + \log S(x_i; z_i)\}.$$
Parametric Models: Examples

We use the leukemia data to illustrate the equivalence.

\[
survreg(Surv(log(time),status)~x,leuk,dist="extreme")
\]
\[
survreg(Surv(time,status)~x,leuk,dist="weibull")
\]
\[
survreg(Surv(log(time),status)~x,leuk,dist="gaussian")
\]
\[
survreg(Surv(time,status)~x,leuk,dist="loggaussian")
\]
\[
survreg(Surv(time,status)~x,leuk,dist="logistic")
\]
\[
survreg(Surv(log(time),status)~x,leuk,dist="loglogistic")
\]

There are several types of residuals available through resid; also available are DFBETAS with type="dfbetas".

Several types of fitted values are available through predict, including type="quantile".

\[
survreg\] also takes left- and interval-censored data.

---

Parametric Survival Curve

Estimated survival from parametric model.

\[
surv.wei <- survreg(Surv(time,status)^1,leuk,\]
dist="wei",sub=x="Maintained")
\]
\[
surv.exp <- update(surv.wei,dist="exp"); summary(surv.exp)
\]
\[
plot(survfit(Surv(time,status)^x,leuk)[1,,conf=TRUE])
\]
\[
points(leuk$time,exp(-exp((log(leuk$time)-surv.exp$coef)))
\]
\[
points(leuk$time,exp(-exp((log(leuk$time)-(surv.exp$coef-1.96*.378)))),col=2)
\]

# quantiles
\[
quan<-predict(surv.exp,type="quan",p=seq(05,.95,len=19))[1,]
\]
\[
1-exp(-exp((log(quan)-surv.exp$coef)))
\]

One could also get the standard errors for the quantiles, useful for the CIs of quantiles, but not of the survival curve.
Parametric Models: Miscellaneous

For Weibull, \( \Lambda(t) = e^{(\log t - \mu)/\sigma} \), or
\[
\log \Lambda(t) = (\log t - \mu)/\sigma,
\]
so a log-log plot of \( t - \Lambda(t) \) provides a diagnostic for the appropriateness of the Weibull model.

\[
\text{plot(survfit(Surv(time,status)~x,leuk)[1,],fun="cloglog")}
\]

Weibull, log-normal, and log-logistic regression are examples of accelerated failure models, as the covariate effectively rescales the time axis: \( \Lambda(t; z) = \Lambda(t^*; 0) \), where \( t^* = \exp\{-\beta^T z\} t \). Note that
\[
P(T \leq t) = F((\log(t) - \mu)/\sigma) = F(\log(te^{-\mu})/\sigma),
\]
where \( F(u) \) is the c.d.f. of \( f(x; 0, 1) \), the “standard” in the location-scale family.

Cox (Relative Risk) Model

With covariate \( z_i(t) \), assume a proportional hazards model
\[
\lambda(t; z_i) = \lambda_0(t)r(t; z_i) = \lambda_0(t)r_i(t),
\]
where \( \lambda_0(t) \) is a base hazard and \( r_i(t) = r(t; z_i) = \exp\{\beta^T z_i(t)\} \) is the relative risk. Treating \( \lambda_0(t) \) as a nuisance parameter, one may estimate \( \beta \) by maximizing the partial likelihood
\[
L(\beta) = \prod_j \frac{r_{i(j)}(t_j)}{\sum_{l \in R(t_j)} r_l(t_j)} = \prod_j \frac{\exp\{\beta^T z_{i(j)}(t_j)\}}{\sum_{l \in R(t_j)} \exp\{\beta^T z_l(t_j)\}},
\]
where the \( i(j) \)th item fails at \( t_j \) and \( R(t_j) \) is the risk set at \( t_j^- \).

Conditional on the history up to \( t_j^- \) and the fact that one item fails at \( t_j \), the term \( r_{i(j)}(t_j)/\sum_{l \in R(t_j)} r_l(t_j) \) is proportional to the likelihood of a multinomial model.
Cox Model: Example, Ties

Cox model can be fitted using `coxph` in `survival`, with static $z_i$.

```r
data(ovarian); coxph(Surv(futime,fustat)~age,ovarian)
cox.fit <- coxph(Surv(futime,fustat)~age+ecog.ps+strata(rx),ovarian)
```

The `strata` term allows separate base hazards for different groups, each having their own risk sets.

With ties, say $x_1 = x_2 = t_j$, $\delta_1 = \delta_2 = 1$, the term $r_1(t_j)r_2(t_j)$ goes to the numerator of $L(\beta)$. For the denominator,

- `method="efron"`: $(r_1 + r_2 + \cdots)(.5r_1 + .5r_2 + \cdots)$.
- `method="breslow"`: $(r_1 + r_2 + \cdots)(r_1 + r_2 + \cdots)$.
- `method="exact"`: $\left\{ \frac{1}{(r_1+r_2+\cdots)(r_2+\cdots)} + \frac{1}{(r_2+r_1+\cdots)(r_1+\cdots)} \right\}^{-1}$.

Estimated Survival with Cox Model: Static $z_i$

With $\lambda(t; z_i) = \lambda_0(t)r(z_i) = \lambda_0(t)r_i$, $S(t; z_i) = e^{-\Lambda_0(t)r_i} = S_0(t)^{r_i}$.

$$L = \prod_{j=1}^{k} \left\{ \prod_{l \in D_j} [S(t_j^-; z_l) - S(t_j; z_l)] \prod_{l \in C_j} S(t_j; z_l) \right\},$$

$$= \prod_{j=1}^{k} \left\{ [S_0(t_j^-) - S_0(t_j)]^{\hat{d}_j} S_0(t_j)^{\hat{n}_j} \right\},$$

where $D_j$ contains items failing at $t_j$, $C_j$ contains items censored over $[t_j, t_{j+1})$, $\hat{d}_j = \sum_{l \in D_j} r_l$, and $\hat{n}_j = \sum_{l \in C_j} r_l$. The MLE of $S_0(t)$ is of form $\hat{S}_0(t_j^-) = \prod_{l=1}^{j-1} (1 - \hat{\lambda}_l)$ and $\hat{S}_0(t_j) = \prod_{l=1}^{j} (1 - \hat{\lambda}_l)$, with $\hat{\lambda}_j = \hat{d}_j/\hat{n}_j$ for $\hat{n}_j = \sum_{l \in R(t_j)} r_l$. This is a weighted version of Kaplan-Meier.

```r
plot(survfit(cox.fit,new=cox.fit$mean))
plot(survfit(cox.fit,new=list(age=60,ecog.ps=2)))
```
Parametric PH Model

With the Weibull model discussed earlier, one has

\[ \Lambda(t; z) = \exp\left\{ (\log t - \tilde{\beta}^T \tilde{z})/\sigma \right\} = \left( t/e^{\tilde{\beta}_0} \right)^{1/\sigma} \exp\left\{ -\beta^T z/\sigma \right\}, \]

where \( \tilde{\beta}^T \tilde{z} = \tilde{\beta}_0 + \beta^T z \). It is a PH model with \( \Lambda_0(t) = \left( t/e^{\tilde{\beta}_0} \right)^{1/\sigma} \) and \( r(t; z) = \exp\left\{ -\beta^T z/\sigma \right\} \). Note that

\[ \log \Lambda_0(t) = (\log t - \tilde{\beta}_0)/\sigma, \]

so a log \( t \)-log \( \hat{\Lambda}_0(t) \) plot from a Cox fit can be used as a diagnostic for the adequacy of a Weibull model.

\[
\text{cox.fit; plot(survfit(cox.fit),fun="cloglog")}
\]
\[
\text{wei.fit <- survreg(Surv(futime,fustat)~age+ecog.ps +strata(rx),data=ovarian,dist="weibull")}
\]
\[
\text{summary(wei.fit); summary(cox.fit)}
\]

Note that \( \hat{\beta}'s \) from Cox fit and Weibull fit differ in sign and scale.

Survival Analysis: Summary

- Life time distributions are characterized by survival and hazard. Failure times are typically censored, and often also truncated. (Slides 1-6)
- Kaplan-Meier extends empirical distribution to censored data. Logrank tests extend Wilcoxon and Kruskal-Wallis to censored data. One-sample test is related to Poisson model. (Slides 7-13)
- Parametric accelerated failure time models are among generalized linear models for life time data. (Slides 14-17)
- Cox model with parametric relative risk can be fitted via partial likelihood. The base survival can be estimated by weighted Kaplan-Meier. (Slides 18-21)