Survival Analysis (cont.)

Survival Analysis

Dependent variable of interest: *time* from some initial observation until the occurrence of an event.

Examples:

- Time to failure of a light bulb
- Time to death from initial diagnosis of a disease
- Time to relapse

Properties of Survival Times

- The distribution is skewed
- Survival times of some units may not be known (i.e., *censored*).
  This may be due to loss to follow-up, drop-outs, or the event of interest not occurring when the experiment ended.

Basic assumption: Censoring mechanism is independent of failure times (i.e., cenosring is non-informative).

Characterization of the Distribution of Survival Times

Let T be survival time, with p.d.f. $f(t)$.

The *survival function* $S(t)$ is defined to be

$$S(t) = Pr[T > t]$$

The *hazard function*

The probability of failure at a specified time, given that there has been survival until that time point. Also known as *instantaneous failure rate*.

$$\lambda(t) = \lim_{u \to 0} \frac{\Pr\{t < T \leq t + u \,|\, T > t\}}{u} = \frac{f(t)}{S(t)}$$

Relationship between Survival and hazard functions:

The cumulative hazard function

$$\Lambda(t) \equiv \int_0^t \lambda(v)dv = -\log S(t)$$

$$S(t) = \exp[-\Lambda(t)]$$

Note: The hazard function or the cumulative hazard function NOT a probability. They are a measure of risk. The greater the value of the cumulative hazard function, the greater the risk of failure by time t.

Exponential distribution

$$\Lambda(t) = \lambda t$$

$$S(t) = \exp(-\Lambda(t)) = \exp(-\lambda t)$$

$$\Pr\{T > t_0 + t \mid T > t_0\} = \Pr\{T > t\}$$

NB: hazard is constant for exponential; memoryless property

Weibull distribution

$$\lambda(t) = \alpha\gamma t^{\gamma-1}$$

$$\Lambda(t) = \alpha t^{\gamma}$$

$$S(t) = \exp(-\Lambda(t)) = \exp(-\alpha t^{\gamma})$$

NB: γ is shape parameter, and $\alpha = \lambda^{\gamma}$

NB: When γ > 1, hazard increases with increasing time t.

Objectives of Survival Analysis

- Estimation of survival curves
- Estimation of parameters of survival curves and/or hazard functions
- Comparing survival curves
- Model building and diagnostics

Kaplan-Meier Estimator

Approach uses the actual times, rather than grouping them into intervals.

Denote the ranked times of failure and censoring by $t_{(1)} < \cdots < t_{(n)}$.

Let n_j be the number alive just before time $t_{(j)}$, and d_j the number that died at time $t_{(j)}$. Put

$$s_j = Pr[T > t_{(j)} \mid T > t_{(j-1)}]$$

Then a reasonable estimate of of $1 - s_j$ is

$$1 - \hat{s}_j = \frac{d_j}{n_j}$$

Hence

$$\hat{S}(t) = \Pi_{t_{(j)} \leq t} \hat{s}_j$$

where the multiplication is over the uncensored failure times $t_{(j)}$.

Kaplan-Meier Estimator

Nelson estimator

The Nelson estimtor of the integrated hazard function is given by

$$\hat{\Lambda}_N(t) = \sum_{uncensored: t_{(j)} \leq t} \left(\frac{d_j}{n_j} \right).$$

Fleming-Harrington Estimator

$$\hat{S}_{FH}(t) = e^{-\hat{\Lambda}_N(t)}$$

Variance Estimation

Greenwood's formula

$$Var\hat{\Lambda}_N(t) = \sum_{uncensored: t_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)}$$

$$Var\hat{S}(t) = \hat{S}^2(t) Var\hat{\Lambda}_N(t)$$

Approximate confidence intervals based on the original scales:

$$\hat{S}(t) \pm Z_{\alpha/2} \times SE(\hat{S}(t))$$

may give values outside the acceptable range [0,1].

The cumulative hazard or log S scale. Guarantees positive values.

$$exp\{logS \pm Z_{\alpha/2} SE(\hat{\Lambda})\}$$

Log hazard scale. Guarantees values in [0,1].

$$exp\{exp\{log(-logS) \pm Z_{\alpha/2} SE(\hat{\Lambda}_N)\}\}$$

S-Plus commands to fit a Kaplan-Meier and plot it

```
fit <- survfit(Surv(time, status) ~ group, data=dataset)
plot(fit, lty=2:3)
```

```
> data(ovarian)
> ovarian[1:5,]
  futime   fustat     age        resid.ds rx ecog.ps
1     59      1        72.3315        2       1       1
2    115      1        74.4932        2      1       1
3    156      1        66.4658        2      1       2
4    421      0        53.3644        2      2       1
5    431      1        50.3397        2      1       1
......
```

```
fit <- survfit(formula = Surv(futime, fustat) ~ rx, data = ovarian)
```

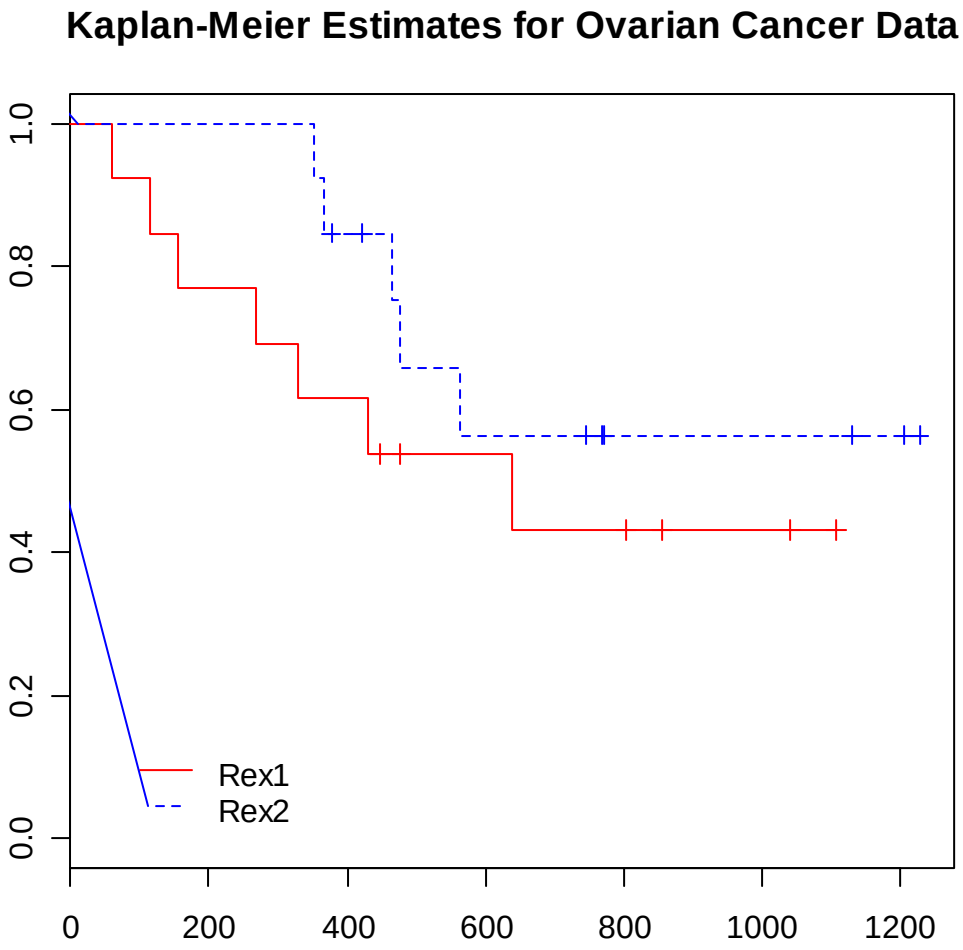
> **summary(fit)**

rx=1

time	n.risk	n.event	survival	std.err	95% CI	
59	13	1	0.923	0.0739	0.789	1.000
115	12	1	0.846	0.1001	0.671	1.000
156	11	1	0.769	0.1169	0.571	1.000
268	10	1	0.692	0.1280	0.482	0.995
329	9	1	0.615	0.1349	0.400	0.946
431	8	1	0.538	0.1383	0.326	0.891
638	5	1	0.431	0.1467	0.221	0.840

rx=2

time	n.risk	n.event	survival	std.err	95% CI	
353	13	1	0.923	0.0739	0.789	1.000
365	12	1	0.846	0.1001	0.671	1.000
464	9	1	0.752	0.1256	0.542	1.000
475	8	1	0.658	0.1407	0.433	1.000
563	7	1	0.564	0.1488	0.336	0.946


Kaplan-Meier Estimates for Ovarian Cancer Data
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200
Rex1
Rex2

Comparing Survival Curves

Two Samples

Suppose we are interested in comparing the median survival times corresponding to two distributions.

Let $X_1, \cdots, X_n$, and $Y_1, \cdots Y_m$ be independent samples of survival times.

2. *Mantel-Haenszel or log-rank test*

Consider a sequence of r 2×2 tables.

	Number Dead	Number Alive	
Group 1	a_k	b_k	n_{k1}
Group 2	c_k	d_k	n_{k2}
	m_{k1}	m_{k2}	n_k

The hypothesis of interest may be formulated as

$$H_0 : p_{k1} = p_{k2}, k = 1, \cdots, r,$$

where p_{k1} and p_{k2} are stratum specific death probabilities for Group 1 and 2, respectively.

When the strata are independent, the Mantel-Haenszel test is given by

$$T_{MH} = \frac{\Sigma_{k=1}^{r}(a_k - E_k)}{\sqrt{\Sigma_{k=1}^{r} V_k}}$$

where $E_k = E[A_k \mid H_0]$ and $V_k = Var[A_k \mid H_0]$.

T_{MH} is approximately standard normal under H_o, when the strata are independent.

Now assume that the strata correspond to uncensored (failure) time points. The test thus obtained is called the log-rank test.

It is noted that:

- *Although the strata are not independent, under regularity conditions, asymptotic normality still holds.*
- *Unlike the Gehan test, the long-rank test is not affected by unequal censoring patterns.*
- *This test is known to be most powerful under proportional hazards alternatives.*

The test may be generalized to Tarone-Ware and Fleming-Harrington family of tests. Define

$$T = \frac{\Sigma_{k=1}^{r} W_k(a_k - E_k)}{\sqrt{\Sigma_{k=1}^{r} W_k^2 V_k}}$$

- $W_k = 1$: *The log-rank test*
- $W_k = n_k$: *Gehan test*
- $W_k = \sqrt{n_k}$: *Tarone-Ware class of tests. Intermediate between Gehan and log-rank.*
- $W_k = \hat{S}^{\rho}(t_k)$. *When* $\rho = 1$, *we have Peto-Peto, while for* $\rho \epsilon (0, 1)$, *we get Fleming-Harrington tests.*

- S-Plus implementation:

survdiff(Surv(time, status) ~ group, data=dataset, rho=0)

> survdiff(Surv(futime, fustat) ~ rx, rho=0, data=ovarian)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
rx=1	13	7	5.23	0.596	1.06
rx=2	13	5	6.77	0.461	1.06

Chisq= 1.1 on 1 degrees of freedom, p= 0.303

Regression Approaches in Survival Analysis

In the presence of covariates, the standard linear model formulation is not appropriate for survival times, due to censoring and the skewed nature of the distributions.

Proportional Hazards Models

A class of models where the hazard function $\lambda(t;x)$ can be written as:

$$\lambda(t;x) = \lambda_o(t)\, g(x)$$

Where g is a non-negative function.

Assumption:

$$\lambda(t;x)/\lambda_o(t) = g(x)$$

i.e., hazard ratio does not depend on time!

Cox Proportional Hazards Model

Let $\mathbf{x} = (\mathbf{x_1}, \cdots, \mathbf{x_b})'$ be a vector of covariates.
Let T be failure time.

Define the hazard function for a given subject

$$\lambda(t; \mathbf{x}) = \lambda_{\mathbf{o}}(\mathbf{t})\mathbf{e}^{\beta' \mathbf{x}}$$

Hazard Ratio = $\lambda(t;x)/\lambda_o(t)$

= $\exp(\beta'x)$

i.e., hazard ratio is constant over time!

In the above formulation,

- β is a vector of coefficients, and does not include the intercept, which is absorbed in $\lambda_o(t)$.
- A remarkable advantage of this formulation is that the baseline hazard $\lambda_o(t)$ does not need to be specified explicitely.
- The exponential guarantees positive λ for any β.

Estimation

Let $t_1, \cdots, t_n$ be survival times for n individuals, and $\mathbf{x_1}, \cdots, \mathbf{x_n}$ be the corresponding covariates. Let

$$Y_i(t) = \begin{cases} 1 & \text{i'th person is alive at time t} \\ 0 & \text{otherwise} \end{cases}$$

Suppose a death is observed at time t. The conditional probability it is subject j is

$$L_j(\beta) = \frac{\lambda_o(t) e^{\beta' \mathbf{x_j}}}{\Sigma_k^n Y_k(t) \lambda_o(t) e^{\beta' \mathbf{x_k}}}$$

which gives

$$L_j(\beta) = \frac{e^{\beta' \mathbf{x_j}}}{\Sigma_k^n Y_k(t) e^{\beta' \mathbf{x_k}}}$$

The *partial likelihood* is defined as

$$L(\beta) = \Pi_j^n L_j(\beta)$$

One would then find $\hat{\beta}$ such that $logL(\hat{\beta})$ is maximum.

Remarks

- *Estimation does not depend on* $\lambda_o(t)$
- *Model depends only on the ranks of survival times. Therefore, it is a nonparametric procedure.*
- *The estimators may be sensitive to outliers in the covariates.*
- *For tied observations, appropriate adjustment is necessary.*

Inference

Consider the problem of testing the null hypothesis: $H_o : \beta = \beta_o$*:*

1. *Likelihood Ratio Test (Wilks LR test)*

$$T_{LR} = -2ln\frac{L(\beta_o)}{L(\hat{\beta})}$$

which has an approximate χ^2_p *distribution under the null.*

2. *Wald Statistic*

$$T_W = (\hat{\beta} - \beta_o)' \hat{\Sigma}_{\hat{\beta}}^{-1} (\hat{\beta} - \beta_o)$$

where $\hat{\Sigma}$ *is the variance covariance matrix of* $\hat{\beta}$ *and is given by*

$$\hat{\Sigma}_{\hat{\beta}} = \mathbf{I^{-1}}(\beta)$$

and

$$I(\beta) = -\frac{\partial^2}{\partial \beta^2} lnL(\beta)$$

Under H_o, T_W *has an appriximate* χ_p^2 *distribution. For individual hypotheses*

$$T_W = \frac{\hat{\beta}_k - \beta_{k,o}}{SE(\hat{\beta}_k)}$$

3. Rao's Score Test

$$T_S = U'(\beta_\circ) I(\beta_\circ) U(\beta_\circ)$$

$$where\ U(\beta) = \frac{\partial}{\partial \beta} ln L(\beta)$$

```
>      fitcox <- coxph( Surv(futime,fustat)~rx,data=ovarian)

> summary(fitcox)
     coef exp(coef) se(coef)     z    p
rx -0.596     0.551    0.587 -1.02 0.31

   exp(coef) exp(-coef) lower .95 upper .95
rx     0.551       1.82     0.174      1.74


Likelihood ratio test        = 1.05  on 1 df,   p=0.305
Wald test            = 1.03  on 1 df,           p=0.310
Score (logrank) test = 1.06  on 1 df,           p=0.303
```

Remarks

- *In terms of giving correct p-values, the LR test is most reliable.*
- *Since Wald's test can be used for individual, as opposed to overall testing, it may be advantageous.*
- *The score test is less cumbersome computationally, since it does not involve* $\hat{\beta}$.

Proportional Hazards Assumption

Suppose x_1 is an indicator variable, taking the values 0 or 1, corresponding to treatments 1 and 2, respectively. Then the ratio of hazards, under proportional hazards assumption, is

$$\frac{h(t;\beta_1, x_1 = 1)}{h(t;\beta_1, x_1 = 0)} = e^{\beta_1}$$

which is independent of t. Now, consider the hypotheses

H_o : Ratio of hazards does not depend on t

vs

H_1 : Ratio of hazards depends on t

This may be tested by defining the auxiliary variable

$$x_2(t) = \begin{cases} log\ t - \bar{t}, & if\ treatment\ 1 \\ 0, & otherwise \end{cases}$$

The log transformation avoids numerical instability, while centering at the mean survival time is intended to improve interpretability. Then the hazard ratio becomes

$$e^{\beta_1+\beta_2(lnt-\bar{t})} = t^{\beta^2} e^{\beta_1-\beta_2\bar{t}}$$

Thus, the hypotheses of interest: $H_o : \beta_2 = 0$ *vs.* $\beta_2 \neq 0$. *If* H_o *is rejected,* $\beta_2 < 0$ *indicates decreasing hazard rate over time, while* $\beta_2 > 0$ *implies increasing hazard rate.*

Residual Analysis

1. *Martingale Residulas*

Martingale residulas may be used to determine appropriate functional forms of the covariates.

- For each covariate x_j, compute Martingale residuals excluding that covariate from the model
- Plot the residuals thus obtained vs. x_j
- The scatter plot may suggest the appropriate functional form for x_j

2. DFBETAS: Influential point identification

   Recall that the proportional hazards model may be sensitive to extreme values in $\mathbf{x}$.

3. Schoenfeld Residuals: Proportional hazards assumptions

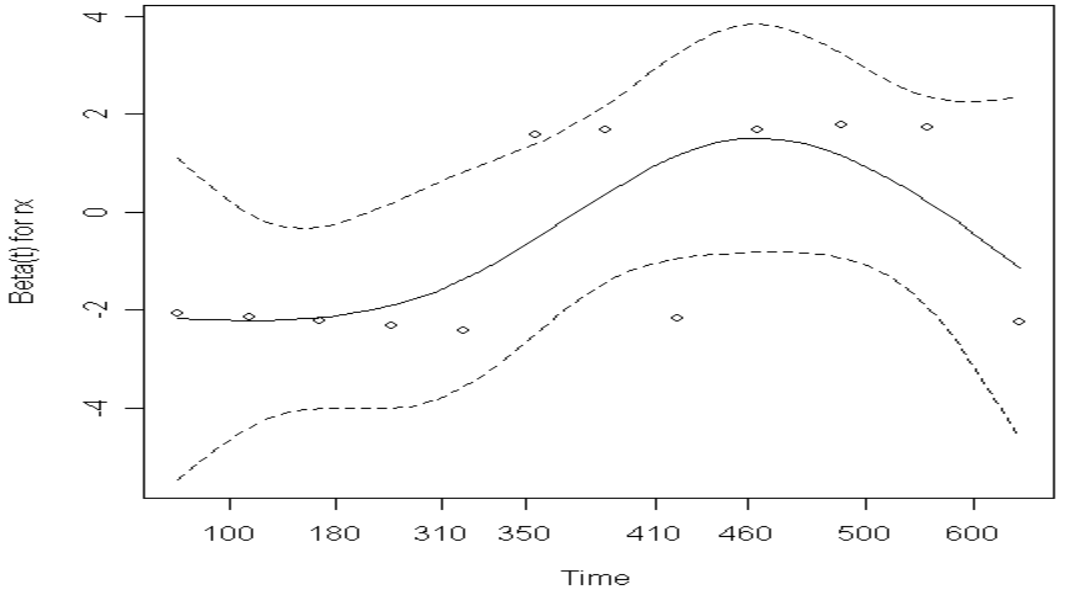
Under the proportional hazards assumption, a plot of the Schoenfeld residuals vs. time should be a random walk in the plane.

Test the proportional hazards assumption of a Cox regression.

```
prop <-cox.zph(fitcox)

plot(prop)
```

Plot of estimate of time-dependent coefficient 'beta(t)'.



If proportional hazards assumption true, 'beta(t)' horizontal line. .

test for slope=0

```
  > prop
                rho     chisq       p
    rx          0.494   2.64        0.104
```

Problem Set 11

Consider the *colon data in the R* package "survival". It gives adjuvant chemotherapy data for colon cancer. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals; 5-FU is a moderately toxic (as these things go) chemotherapy agent. There are two records per person, one for recurrence (etype=1) and one for death (etype=2). Other important variables include:

rx:	Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU
sex:	1=male
age:	in years
time:	days until event or censoring
status:	censoring status

For the following, consider survival to be “Days until Death”, i.e., etype=2.

1. Using a Cox proportional hazards model, estimate the hazard rate for Levamisole relative to 5-FU**.**
2. Interpret the estimated value for: **exp(coef),** and give a 95% confidence interval for the hazard rate .
3. Assess the validity of the proportional hazards assumption in (1) above.
4. Using a Cox proportional hazards model, estimate the hazard rate for Levamisole relative to 5-FU, adjusting for **Age and Sex.**
5. Reading Assignment
   - Martingale Residuals
   - Schoenfeld Residuals
   - Stratified Cox Proportional Hazards Model

```
> library(splines)
> library(survival)
> data(ovarian)
......
Some useful R functions:
> help(package="survival")
Surv             Package a survival variable
coxph            Proportional Hazards Regression
lines.survfit    Add lines to a survival plot
plot.survfit     Plot method for survfit.
print.survfit    Short summary of a survival curve
survfit(formula, data, weights, subset, na.action,
        newdata, individual=F, conf.int=.95, se.fit=T,
            type=c("kaplan-meier","fleming-harrington", "fh2"),
            error=c("greenwood","tsiatis"),
            conf.type=c("log","log-log","plain","none"),
            conf.lower=c("usual", "peto", "modified"))
    'print.survfit', 'plot.survfit', 'lines.survfit',
    'summary.survfit', 'survfit.object' 'coxph', 'Surv', 'strata'.
fit <- survfit(formula = Surv(futime, fustat) ~ rx, data = ovarian)
```