Chapters 5 & 6

PubH 7450

©Wei Pan

Email: weip@biostat.umn.edu

Http: www.biostat.umn.edu/∼weip

§5.4 Life-Table Methods

- Goal: to estimate $S(t)$, $h(t)$,...
- When: 1) n is large; 2) grouped data
- Data:

  1) $I_j = [a_{j-1}, a_j)$, $j = 1, 2, ..., k+1$, $a_0 = 0$, $a_{k+1} = \infty$;
  2) y'_j =#(risk set at a_{j-1});
  3) W_j =#(censored in I_j), e.g. loss-to-followup;
  4) d_j =#(events in I_j).
- Estimates:

  Assuming that censoring is uniform inside I_j,

$$y_j = \#(\text{risk set in } I_j) = y'_j - W_j/2.$$

$$\hat{S}(a_j) = \hat{S}(a_{j-1}) \left(1 - \frac{d_j}{y_j}\right) = \prod_{i=1}^{j} \left(1 - \frac{d_i}{y_i}\right).$$

At the middle point a_{mj} of interval I_j,

$$\hat{f}(a_{mj}) = \frac{\hat{S}(a_{j-1}) - \hat{S}(a_j)}{a_j - a_{j-1}}.$$

$$\hat{h}(a_{mj}) = \frac{\hat{f}(a_{mj})}{\hat{S}(a_{mj})} = \frac{2\hat{f}(a_{mj})}{\hat{S}(a_{j-1}) + \hat{S}(a_j)},$$

or, by the interpretation of $h()$ (as ...),

$$\hat{h}(a_{mj}) = \frac{d_j}{(a_j - a_{j-1})(y_j - d_j/2)}.$$

Conditional probability of having an event in J_j is $\hat{q}_j = d_j/y_j$, thus (as discussed before?) $\hat{S}(a_j) = \hat{S}(a_{j-1})(1 - \hat{q}_j)$.

Greenwood's (1926) formula:

$$\widehat{Var}(\hat{S}(a_{j-1})) = \hat{S}(a_{j-1})^2 \sum_{i=1}^{j-1} \frac{d_i}{y_i(y_i - d_i)}.$$

- mrl, mdrl
  mrl(x) is the mean of $X - x$ with the conditional distribution $(X|X \geq x)$, i.e. $S(t)/S(x) \Longrightarrow mrl(x) = \int_x^\infty S(t)dt/S(x)$.
  mdrl(x) is the median of $X - x$ with the conditional distribution of $(X|X \geq x)$:
  $\Longrightarrow mdrl(x) =$ [median with $S(t)/S(x)$] $- x$.

- Q: mrl(0) = mrl(2) + 2?
  1) No,
  2) Yes,

- Example 5.4; SAS handout.

§5.2 Arbitrarily Censored and Truncated Data

- Goal: to estimate $S(t)$ for X *nonparametrically*.
- Given data: possibly right-censored, left-censored (and thus doubly-censored), and interval-censored; possibly left-, right- and even interval-truncated.
- Approach: NPMLE
  1) Write down NP likelihood L, then numerically maximize it.
  2) Via self-consistency or expectation-maximization (EM) algorithm; an extension of that for right-censored data.
  Why 2)?
  Why not 2)?
- References:
  Turnbull (1974, JASA): doubly-censored data.
  Turnbull (1976, JRSS-B): interval-censored and truncated data.
- First, consider only interval-censored data: $(L_i, R_i]$, i=1,...,n.

- Ordering L_i's and R_i's to get distinct $\tau_0 < \tau_1 < ... < \tau_m$.
  Let $p_j = Pr(\tau_{j-1} < X \le \tau_j)$, $j = 1, ..., m$
  known or unknown?

- $\alpha_{ij} = I\{(\tau_{j-1}, \tau_j] \subseteq (L_i, R_i]\} = I(\tau_{j-1} \ge L_i, \tau_j \le R_i)$.
  known or not?

- $I_{ij} = I\{X_i \in (\tau_{j-1}, \tau_j]\}$.
  known or unknown?
  If not, how to estimate it?

- Use its ...

  $$E(I_{ij}|Data) = Pr\{X_i \in (\tau_{j-1}, \tau_j] | X_i \in (L_i, R_i]\} = \frac{\alpha_{ij} p_j}{\sum_{k=1}^{m} \alpha_{ik} p_k}.$$

- Then,
  $d_j = \sum_{i=1}^{n} E(I_{ij})$, UPDATE: $p_j =$

Or, $y_j = \sum_{k=j}^{m} d_k$, then use the K-M estimator:

$$\hat{S}(\tau_i) = \prod_{j \le i} \left(1 - \frac{d_j}{y_j}\right) .$$

$p_j = \hat{S}(\tau_j) - \hat{S}(\tau_{j-1})$.

- iterate until convergence (i.e. not much change of p_j's.
- How to choose an initial estimate $\hat{S}$?
  Any $\hat{S}$?
  My recommendation:
  1)
  2)
- A potential problem:
- A toy example: observe $(0, 2]$ and $(1, 3]$, $n = 2$.
  $L =$
  To max $L \Longrightarrow$

- Candidate non-zero probability mass intervals: $(u_i, v_i]$'s.
  u_1: starting from 0, find the largest L_i without jumping over any R_i;
  u_2: jumping over consecutive R_i's at the next right of u_1 until encounter an L_i; keep going, find the largest L_i without jumping over another R_i;
  ...
  v_i: the smallest R_j that is larger than u_i.
- How to handle exact event times in the above self-consistency algorithm?
- How to handle left-censoring?
  $L_i =$
- How to handle right-censoring?
  $R_i =$

- R package `Icens`
- Example: Example 5.2 in R; Table 5.4.

Interval	Survival Probability
0–4	1.000
5–6	0.954
7	0.920
8–11	0.832
12–24	0.761
25–33	0.668
34–38	0.586
40–48	0.467
≥48	0.000

τ	Initial $S(t)$	Estimated Number of Deaths d	Estimated Number at Risk Y	Updated $S(t)$	Change
0	1.000	0.000	46.000	1.000	0.000
4	0.979	0.842	46.000	0.982	−0.002
5	0.955	1.151	45.158	0.957	−0.002
6	0.934	0.852	44.007	0.938	−0.005
7	0.905	1.475	43.156	0.906	−0.001
8	0.874	1.742	41.680	0.868	0.006
10	0.848	1.286	39.938	0.840	0.008
11	0.829	0.709	38.653	0.825	0.004
12	0.807	1.171	37.944	0.799	0.008
14	0.789	0.854	36.773	0.781	0.008
15	0.775	0.531	35.919	0.769	0.006
16	0.767	0.162	35.388	0.766	0.001
17	0.762	0.063	35.226	0.764	−0.002
18	0.748	0.528	35.163	0.753	−0.005
19	0.732	0.589	34.635	0.740	−0.009
22	0.713	0.775	34.045	0.723	−0.011
24	0.692	0.860	33.270	0.705	−0.012
25	0.669	1.050	32.410	0.682	−0.012
26	0.652	0.505	31.360	0.671	−0.019
27	0.637	0.346	30.856	0.663	−0.026
32	0.615	0.817	30.510	0.646	−0.031
33	0.590	0.928	29.693	0.625	−0.035
34	0.564	1.056	28.765	0.602	−0.039
35	0.542	0.606	27.709	0.589	−0.047
36	0.523	0.437	27.103	0.580	−0.057
37	0.488	1.142	26.666	0.555	−0.066
38	0.439	1.997	25.524	0.512	−0.073
40	0.385	2.295	23.527	0.462	−0.077
44	0.328	2.358	21.233	0.410	−0.082
45	0.284	1.329	18.874	0.381	−0.097
46	0.229	1.850	17.545	0.341	−0.112
48	0.000	15.695	15.695	0.000	0.000

- A special case: doubly-censored data

  The algorithm on p.141 seems incorrect, e.g. y_i is not defined.

- A similar and simpler algorithm

  1) define τ_j's as before; start with initial p_j's;

  2) **excluding** left-censored data, define d_j and y_j;

  3) for each left censored obs i,

  $E(I_{ij}) = Pr\{X_i \in (\tau_{j-1}, \tau_j] | X_i \in (0, R_i]\} = \frac{\alpha_{ij} p_j}{\sum_{k=1}^m \alpha_{ik} p_k}$;

  o/w, $E(I_{ij}) = 0$;

  4) $d'_j = d_j + \sum_{i=1}^n E(I_{ij})$;

  5) $y'_j = y_j + \sum_{k=j}^m \sum_{i=1}^n E(I_{ik})$.

  6) plug-into the K-M estimator for new estimates p_j's;

  7) repeat 1)-6) until convergence.

- R package `dblcens`

- Truncated data: conditional on $X_i \in B_i$, observe X_i...
  $\beta_{ij} = I\{(\tau_{j-1}, \tau_j] \subseteq B_i\}$.
- $J_{ij} = \#$ of **unobserved** X_j''s that would fall in $(\tau_{j-1}, \tau_j]$ if a random sample were taken (i.e. no truncation), given $X_i \in B_i$.

$$E(J_{ij}) = \frac{(1-\beta_{ij})p_j}{\sum_{k=1}^m \beta_{ik} p_k}.$$

- Modify
  $d_j = \sum_{i=1}^n [E(I_{ij}) + E(J_{ij})]$.
  Then plug-into $y_j = \sum_{k=j}^m d_k$ and K-M estimator.
- Main idea: if you only know the winning games of the Vikings, is it possible to estimate the number of their losing games? How about not and then allowing the possibility of tied games.
- An estimation problem:
  An alternative:
  Reference: Pan and Chappell (1998, *Lifetime Data Analysis*).

- (not required) NPMLE may not be consistent for left-truncated and interval-censored data;
  Reference: Pan and Chappell (1999, *Lifetime Data Analysis*).
- Special cases: only left-censored, or right-truncated data.
  transform into a problem of right-censoring, or left-truncation!
  how?

§6.2 Estimating the hazard function

- Goal: to estimate $h(t)$ for X *nonparametrically*.
- Given data: right-censored data.
  Note: for arbitrarily censored and truncated data, get the NPMLE $\hat{S}(t)$, then $\hat{H}(t)$; then the following idea applies.
- N-A estimator:
  $\tilde{H}(t) = \sum_{t_i \le t} d_i/y_i$
  $Var[\tilde{H}(t)] = \ldots$
- Method 1: crude
  $\tilde{h}(t_i) = \tilde{H}(t_i) - \tilde{H}(t_{i-1}) = d_i/y_i$.
  plot: $\tilde{H}(t)$, $\tilde{h}(t)$.
  discrete; not continuous.
- Method 2:

$\tilde{\tilde{H}}(t)$: piece-wise linear; based on $\tilde{H}(t)$.

$$\tilde{\tilde{h}}(t) = \frac{d\tilde{\tilde{H}}(t)}{dt} \quad = \quad \frac{d_1}{y_1}/(t_1 - 0) \text{ if } t \in [0, t_1);$$
$$= \quad \frac{d_2}{y_2}/(t_2 - t_1) \text{ if } t \in [t_1, t_2);$$
$$= \quad \ldots$$

plot: $\tilde{\tilde{H}}(t)$, $\tilde{\tilde{h}}(t)$.

piece-wise constant; discontinuous

- Method 3: "smooth" Kernel estimates
- Idea: assuming $h(t)$ is smooth, then $\hat{h}(t^*)$ could be a .... of $h(t)$ with $t \in N(t^*)$, a neightborhood of t^*:

$$\hat{h}(t^*) = \frac{\sum_{t_k \in N(t^*)} w_k \tilde{h}(t_k)}{\sum_{t_k \in N(t^*)} w_k}.$$

1) $N(t^*) = (t^* - b, t^* + b)$ is determined by b, called bandwidth;

2) Weights w_k are determined by a kernel function:

i) Uniform kernel; equal weights:

$K(x) = 1/2$ for $x \in (-1, 1)$; =0 o/w.

ii) Epanechnikov kernel:

$K(x) = 0.75(1 - x^2)$ for $x \in (-1, 1)$; =0 o/w.

iii) others: biweight; Gaussian (i.e. pdf of $N(0, 1)$).

- Kernel smoother:

$$\hat{h}(t) = \frac{1}{b} \sum_{i=1}^{D} K(\frac{t - t_i}{b}) \frac{d_i}{y_i}.$$

$$\widehat{Var}[\hat{h}(t)] = \frac{1}{b^2} \sum_{i=1}^{D} K(\frac{t - t_i}{b})^2 \frac{d_i}{y_i^2}.$$

- A Kernel estimate depends on the choice of b and $K()$, especially on b.

Small/large $b \Longrightarrow$...

bias/variance trade-off:

How to choose? based on subject-matter knowledge, or some predictive performance measurement, e.g. mean integrated squared error (MISE) via cross-validation (CV); see p.172.

- Examples: Figs 6.2-6.4
- Example: R; R package `muhaz`
- Note: a kernel density estimate

$$\hat{f}(t) = \frac{1}{b} \sum_{i=1}^{D} K(\frac{t - t_i}{b}) d\hat{F}(t_i).$$

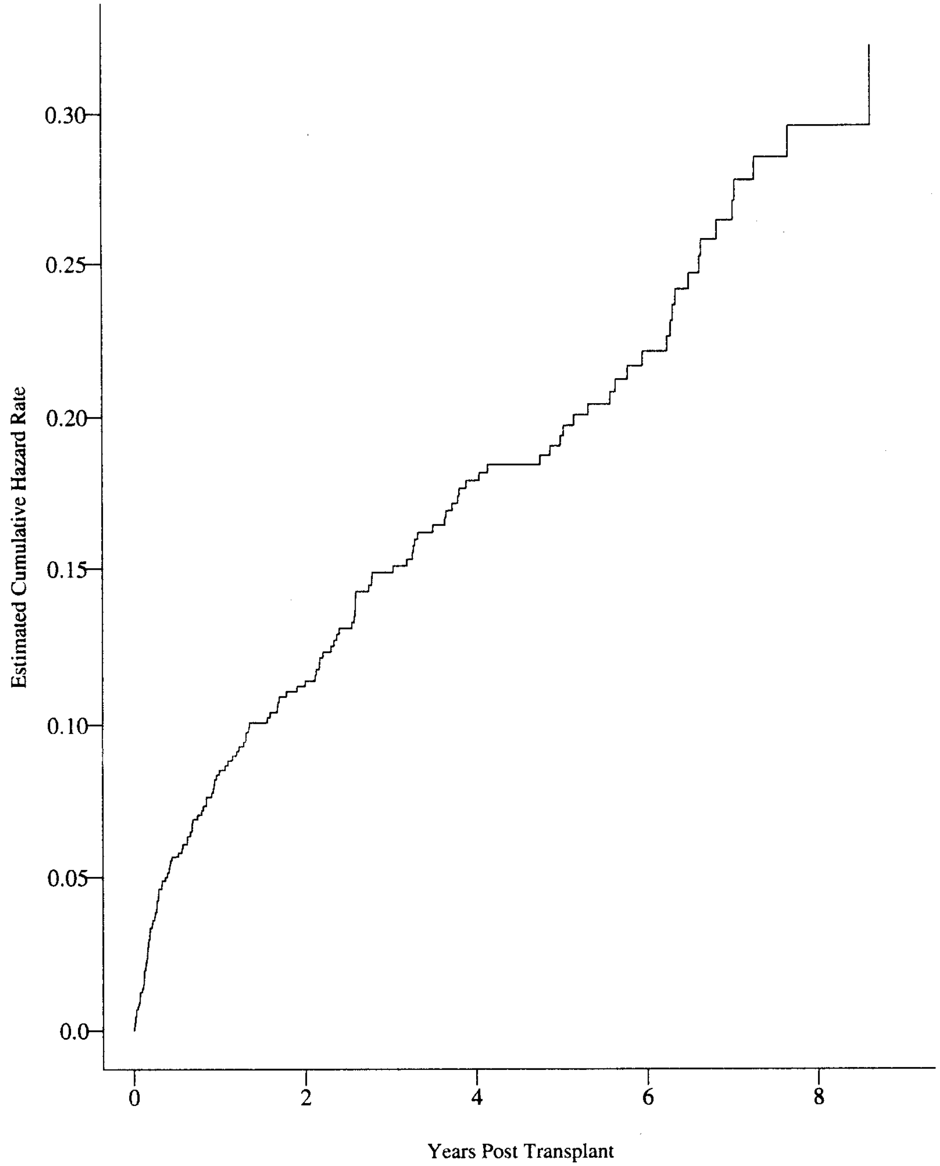

Figure 6.2 *Estimated cumulative hazard rate for kidney transplant patients*

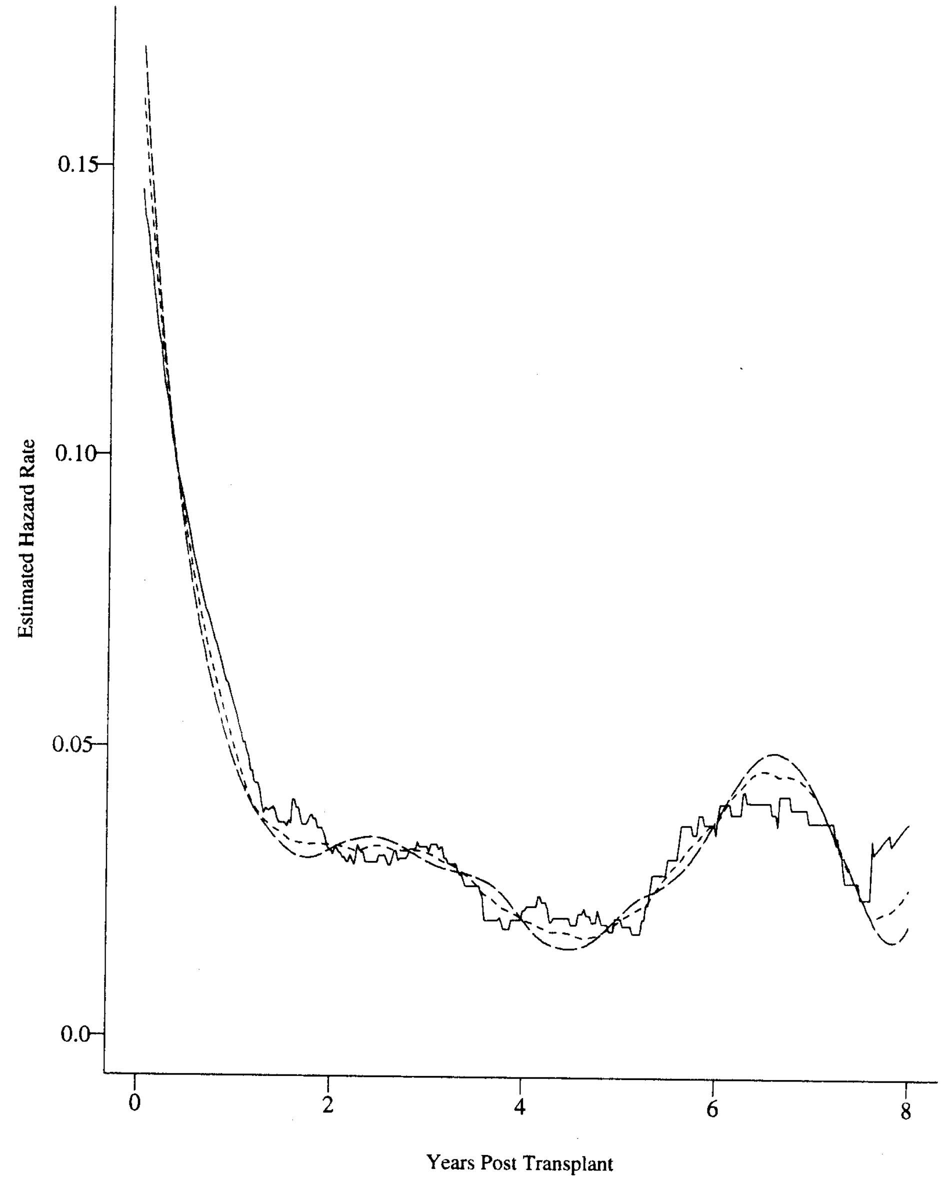

Figure 6.3 *Effects of changing the kernel on the smoothed hazard rate estimates for kidney transplant patients using a bandwidth of 1 year. Uniform kernel (———); Epanechnikov kernel (------) Biweight kernel (— — —)*

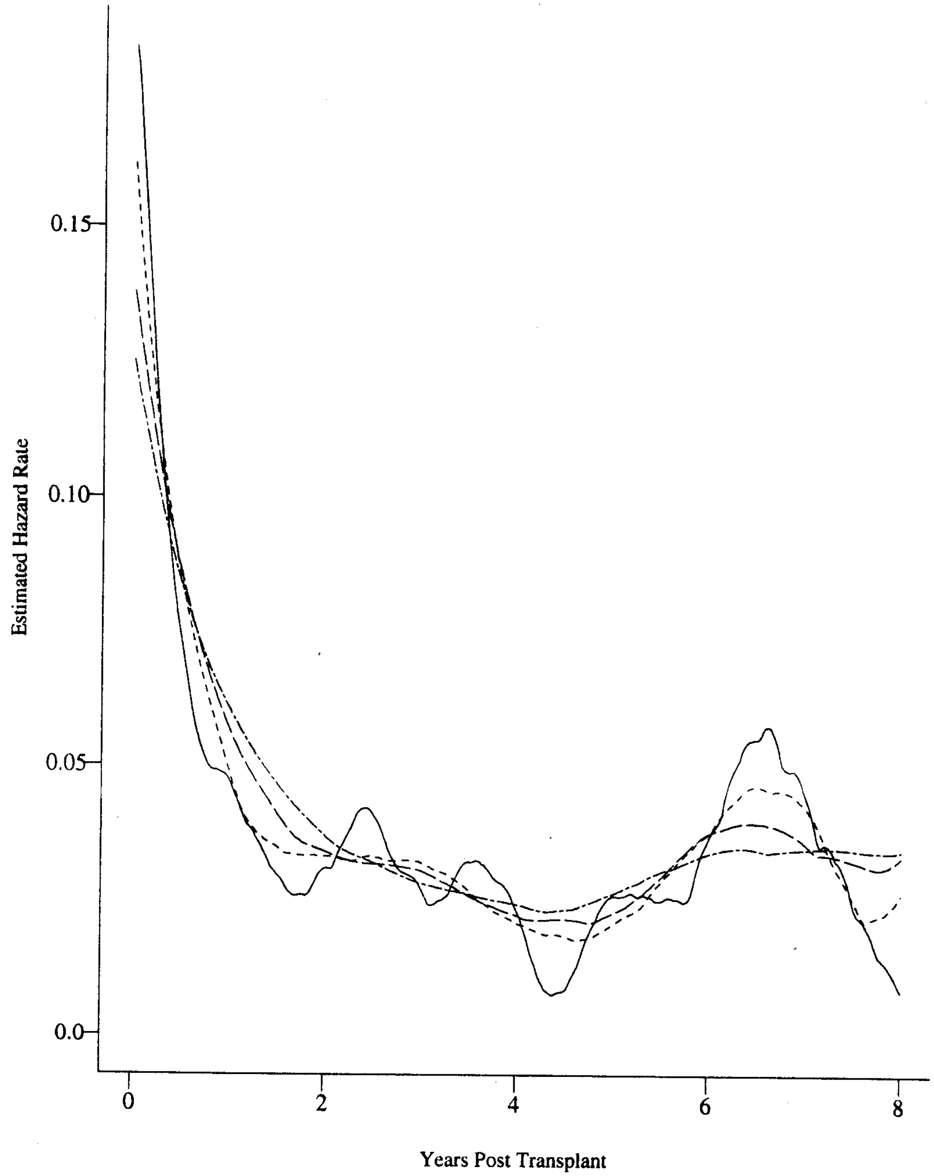

Figure 6.4 *Effects of changing the bandwidth on the smoothed hazard rate estimates for kidney transplant patients using the Epanechnikov kernel. bandwidth = 0.5 years (———) bandwidth = 1.0 years (-------) bandwidth = 1.5 years (— — —) bandwidth = 2.0 years (– · – · –)*

16-3

§6.3 Estimation of excess mortality

- Given data: (T_i, δ_i), $i = 1, ..., n$.
- Goal: to compare the mortality risk of a group of subjects (observed) to that of a standard/reference population.
- Example 6.3: compare 26 psychiatric patients in Iowa with the Iowa population.
- Recall: standardized mortality ratio (SMR)
  $SMR = \frac{\text{\# obs'ed deaths}}{\text{\# exp'ed deaths}}$, a constant.
- Now generalize SMR to time-dependent cases:
  $\beta(t) = \frac{h(t)}{\theta(t)}$, relative (excess) mortality.
  $h(t)$: hazard of the (sub)population of interest;
  $\theta(t)$: hazard of the reference population.
- From (T_i, δ_i), $i = 1, ..., n \Longrightarrow t_i$, d_i, y_i.

- $$\hat{\beta}(t_i) = \frac{d_i}{\#\text{ Exp'ed given } y_i} = \frac{d_i}{\theta(t_i) y_i}.$$

- Cumulative relative excess mortality:
  $\hat{B}(t) = \sum_{t_i \le t} \hat{\beta}(t_i)$.
  $\widehat{Var}[\hat{B}(t)] = \sum_{t_i \le t} \frac{d_i}{\theta(t_i)^2 y_i^2}$.

- can be generalized to a heterogeneous population:
  $Q(t_i) = \#$ Exp'ed given $y_i = \sum_j \theta_j(t_i) y_{ij}$,
  $\sum_j y_{ij} = y_i$; j: subpopulation j.

- Example 6.3.
  Table 6.2.
  $h_F(t) = \log S_F(t+1) - \log S_F(t)$
  $h_M(t) = \log S_M(t+1) - \log S_M(t)$
  Table 1.7 on p. 16.

Gender	Age at admission	Time/status
F	51	1
F	58	1
F	55	2
F	28	22
M	21	30+
......		

$t_i = 1$: $d_i = 2$,

$Q(t_i) = \sum_j \theta_j y_j =$

$h_F(52) + h_F(59) + h_F(56) + h_F(29) + h_M(22) + ...$

$t_i = 2$: $d_i = 1$,

$Q(t_i) = \sum_j \theta_j y_j = h_F(57) + h_F(30) + h_M(23) + ...$

$\hat{B}(t) = \sum_{t_i \le t} \frac{d_i}{Q(t_i)}$.

$\widehat{Var}[\hat{B}(t)] = \sum_{t_i \le t} \frac{d_i}{Q(t_i)^2}$.

Table 6.3 and Fig 6.8.

Age	Survival Function	Hazard Rate	Age	Survival Function	Hazard Rate
Females					
18–19	0.97372	0.00057	48–49	0.93827	0.00352
19–20	0.97317	0.00056	49–50	0.93497	0.00381
20–21	0.97263	0.00055	50–51	0.93141	0.00414
21–22	0.97210	0.00054	51–52	0.92756	0.00448
22–23	0.97158	0.00054	52–53	0.92341	0.00481
23–24	0.97106	0.00056	53–54	0.91898	0.00509
24–25	0.97052	0.00059	54–55	0.91431	0.00536
25–26	0.96995	0.00062	55–56	0.90942	0.00565
26–27	0.96935	0.00065	56–57	0.90430	0.00600
27–28	0.96872	0.00069	57–58	0.89889	0.00653
28–29	0.96805	0.00072	58–59	0.89304	0.00724
29–30	0.96735	0.00075	59–60	0.88660	0.00812
30–31	0.96662	0.00079	60–61	0.87943	0.00912
31–32	0.96586	0.00084	61–62	0.87145	0.01020
32–33	0.96505	0.00088	62–63	0.86261	0.01132
33–34	0.96420	0.00095	63–64	0.85290	0.01251
34–35	0.96328	0.00103	64–65	0.84230	0.01376
35–36	0.96229	0.00110	65–66	0.83079	0.01515
36–37	0.96123	0.00121	66–67	0.81830	0.01671
37–38	0.96007	0.00130	67–68	0.80474	0.01846
38–39	0.95882	0.00140	68–69	0.79002	0.02040
39–40	0.95748	0.00152	69–70	0.77407	0.02259
40–41	0.95603	0.00162	70–71	0.75678	0.02494
41–42	0.95448	0.00176	71–72	0.73814	0.02754
42–43	0.95280	0.00193	72–73	0.71809	0.03067
43–44	0.95096	0.00216	73–74	0.69640	0.03446
44–45	0.94891	0.00240	74–75	0.67281	0.03890
45–46	0.94664	0.00268	75–76	0.64714	0.04376
46–47	0.94411	0.00296	76–77	0.61943	0.04902
47–48	0.94132	0.00325	77–78	0.58980	0.05499

TABLE 6.3

Computation of Cumulative Relative Mortality for 26 Psychiatric Patients

t_i	d_i	$Q(t_i)$	$\hat{B}(t)$	$\hat{V}[\hat{B}(t)]$	$\sqrt{\hat{V}[\hat{B}(t)]}$
1	2	0.05932	33.72	568.44	23.84
2	1	0.04964	53.86	974.20	31.21
11	1	0.08524	65.59	1111.84	33.34
14	1	0.10278	75.32	1206.51	34.73
22	2	0.19232	85.72	1260.58	35.50
24	1	0.19571	90.83	1286.69	35.87
25	1	0.18990	96.10	1314.42	36.25
26	1	0.18447	101.52	1343.81	36.66
28	1	0.19428	106.67	1370.30	37.02
32	1	0.18562	112.05	1399.32	37.41
35	1	0.16755	118.02	1434.94	37.88
40	1	0.04902	138.42	1851.16	43.03

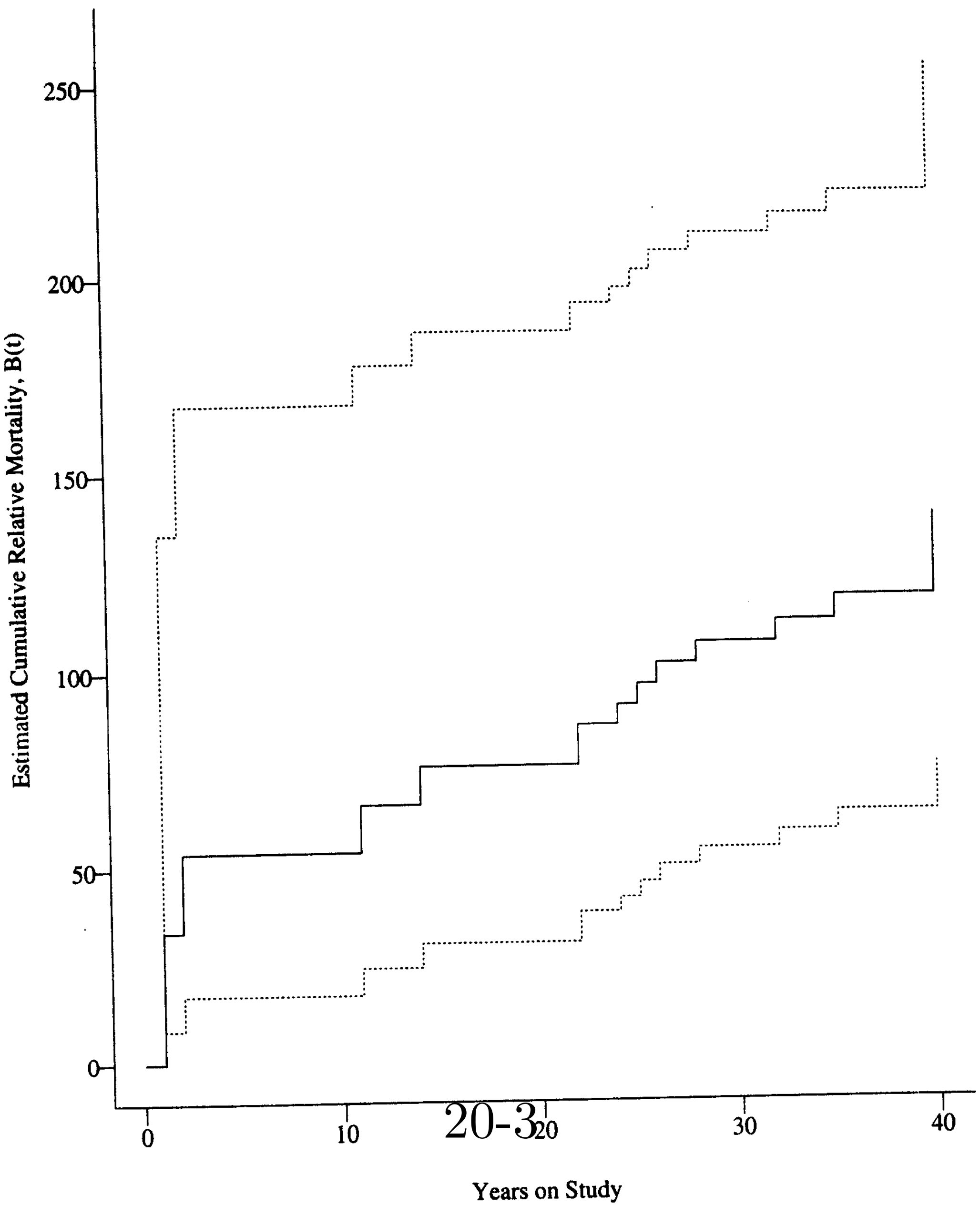

Figure 6.8 *Estimated cumulative relative mortality (solid line) and 95% pointwise confidence interval (dashed line) for Iowa psychiatric patients*