B.3 Survival regression models and testing for survival distributions

1. (a) In a Weibull model, the survival function is $S(x) = e^{-(\rho x)^\alpha}$. Thus $\log(-\log S(x)) = \alpha \log \rho + \alpha \log x$, and if we plot $\log(-\log \hat{S}(x))$ against $\log x$ we should see something close to a straight line. Since the exponential model is a submodel of the Weibull (with $\alpha = 1$), we can apply the likelihood ratio test. If $\ell(\rho, \alpha)$ is the log likelihood, we have under the null model (that the data were sampled from an exponential distribution)

$$2\big(\sup_{(\rho,\alpha)} \ell(\rho, \alpha) - \sup_{\lambda} \ell(\lambda, 1)\big) \sim \chi^2_1.$$

For the log-logistic model, we expect the plot of $\log(\frac{1}{\hat{S}(x)} - 1)$ against $\log x$ to be approximately linear.

(b) Let S_1 and S_2 be the survival curves for the two populations, and S_0 the baseline survival. Under the accelerated lifetime model, $S_i(x) = S_i(\rho_i x)$ for some positive constants ρ_1, ρ_2. Then if we plot $S_i(x)$ against $\log x$, we see that whatever value S_0 takes at ordinate $\log x$, S_i will take the same value at an interval of $\log \rho_i$. (The same will be true of any function of S_i.) Thus, the graphs corresponding to $\hat{S}_1$ and $\hat{S}_2$ should differ approximately by a uniform horizontal shift.

The proportional hazards assumption is best tested by plotting $\log(-\log \hat{S}_i(x))$. Under PH, $S_i(x) = S_0(x)^{\rho_i}$, which implies that

$$\log(-\log S_i(x)) = \log(-\rho_i \log S_0(x)) = \log(\rho_i) + \log(-\log S_0(x)).$$

Thus, if $\log(-\log \hat{S}_i(x))$ is plotted against x, the two graphs should differ approximately by a constant vertical shift if the two groups satisfy the PH assumption. The same is true if we plot $\log(-\log \hat{S}_i(x))$ against any function of x. Thus, if we plot $\log(-\log \hat{S}_i(x))$ against $\log x$, we will see a constant vertical shift reflecting the PH assumption, and a constant horizontal shift reflecting the AL assumption.

(c) The computations for the Kaplan–Meier estimator are given in Table B.1. In figure B.1 we plot the two survival curves (red for control, black for treatment), as $\log(-\log \hat{S})$ against $\log x$. Both look reasonably close to lines, so it would be reasonable to suppose that they came from Weibull models. The lines are approximately parallel, suggesting that the α parameters are approximately the same. This means that one curve may be obtained from another by a horizontal or vertical shift, suggesting that PH or AL would be appropriate. (Weibull curves with the same α parameter, it should be noted, satisfy both hypotheses.)

We test the hypothesis by finding maximum likelihood estimators. The log likelihood for the exponential distribution are

$$\ell(\lambda) = \sum_i (-\lambda x_i) + d \log \lambda,$$

where d is the number of uncensored observations. Since the maximum likelihood estimator is $\hat{\lambda} = d/\sum x_i$, we get maximum likelihoods of

$$\ell^*_{exp} = d\left(\log d - 1 - \log \sum x_i\right) .qqq$$

For the Weibull distribution we have

$$\ell(\rho, \kappa) = -\sum (\rho x_i)^\alpha + d(\kappa \log \rho + \log \kappa) + \sum_{i \text{ uncensored}} (\kappa - 1) \log x_i.$$

There is no closed form solution, but we can optimise numerically, yielding estimates

t_j	d_j	n_j	$\hat{h}_j$	$\hat{S}(t_j)$
1	2	21	0.095	0.905
2	2	19	0.105	0.810
3	1	17	0.059	0.762
4	2	16	0.125	0.667
5	2	14	0.143	0.572
8	4	12	0.333	0.381
11	2	8	0.250	0.286
12	2	6	0.333	0.191
15	1	4	0.250	0.143
17	1	3	0.333	0.095
22	1	2	0.500	0.048
23	1	1	1.000	0.000

t_j	d_j	n_j	$\hat{h}_j$	$\hat{S}(t_j)$
6	3	21	0.143	0.857
7	1	17	0.059	0.806
10	1	15	0.067	0.752
13	1	12	0.083	0.690
16	1	11	0.091	0.627
22	1	7	0.143	0.537
23	1	6	0.167	0.448

Table B.1: Estimates for control group (left) and treatment group (right) in Gehan study.

	Treatment	Control
$\hat{\lambda}$	0.025	0.12
ℓ^*_{exp}	-42.17	-66.35
$\hat{\rho}$	0.030	0.11
$\hat{\kappa}$	1.35	1.37
ℓ^*_{weib}	-41.66	-64.92

The likelihood ratio statistic for the treatment group is thus $2((-41.66)-(-42.17)) = 1.02$, and for the control group it is 2.86. Comparing these to the χ^2 distribution with 1 degree of freedom, we see that the cutoff for rejecting the null hypothesis that $\kappa = 1$ at the 0.05 significance level would be 3.84. Thus, we cannot reject the null hypothesis for either group.

2. We give below `R` code for computing this in two different ways: Using the function **`survdiff`**, which does the computation automatically, and by extracting the relevant quantities from the survival object and doing the computation directly.

   We get $Z = -1.67$, which corresponds to a p-value of 0.09.

   Using **`survdiff`** we get the same result, but it is reported as a chi-squared statistic of 2.8 (which is 1.67^2) on 1 degree of freedom.

   SURVDIFF CODE

```
> require('survival')
> require('KMsurv')
> data(tongue)
> attach(tongue)

>
  > tongue.surv=Surv(time,delta)
> tongue.fit=survfit(tongue.surv~type)
  > tdiff=survdiff(tongue.surv~type)
  > tdiff
Call:
  survdiff(formula = tongue.surv ~ type)
```

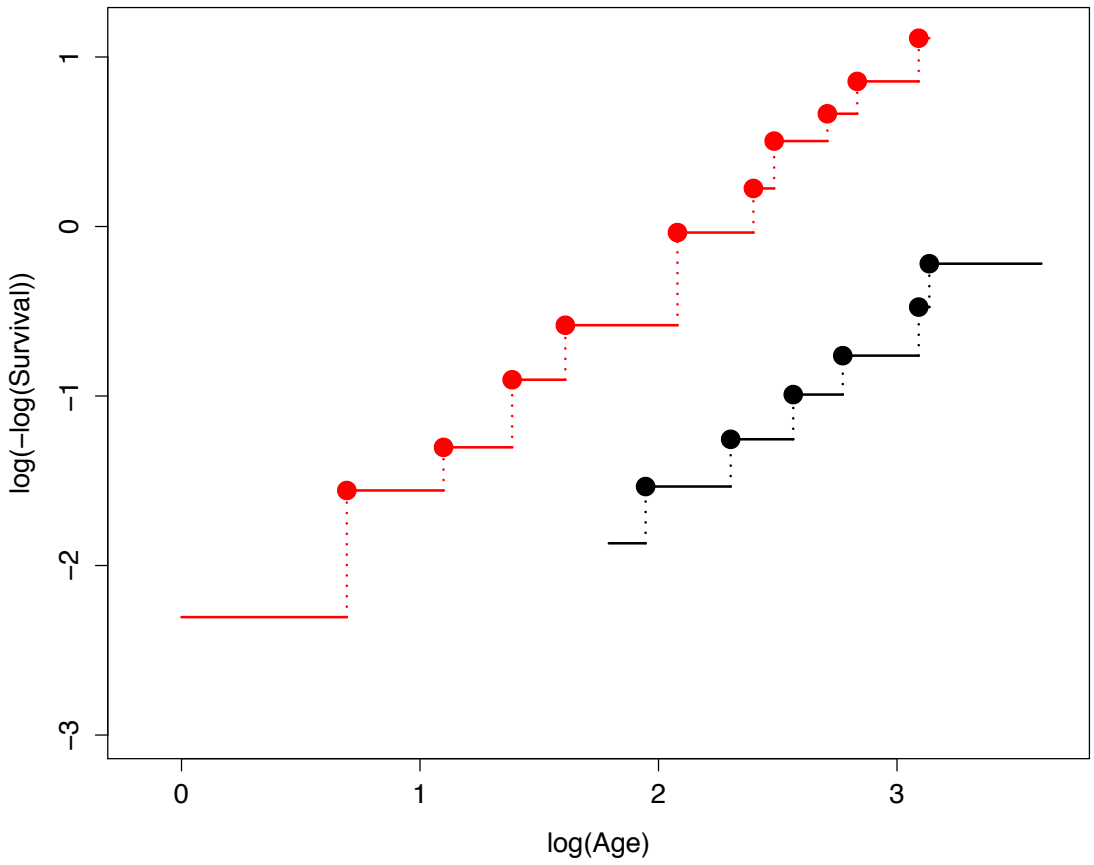


Figure B.1: Plot of estimated survival for Gehan leukaemia data. The control group is in red, the treatment group is black.

```
N Observed Expected (O-E)^2/E (O-E)^2/V
type=1 52       31     36.6     0.843      2.79
type=2 28       22     16.4     1.873      2.79

Chisq= 2.8  on 1 degrees of freedom, p= 0.0949
```

DIRECT COMPUTATION

```
# Problem sheet 4, question 1
require('survival')
require('KMsurv')
data(tongue)
attach(tongue)

tongue.surv=Surv(time,delta)
tongue.fit=survfit(tongue.surv~type)
```

```
n1=tongue.fit$strata[1]
n2=tongue.fit$strata[2]

# Input two vectors of times t1,t2, and
# numbers at risk n1,n2 whose length is 1 longer than the t's
#   Output four vectors I1, I2, (of same length as t1,t2) and Y1,Y2
#   I1[k] gives an index of I2 corresponding to
#   the last time in t2 that precedes t1[k]
#       Thus, we have t2[I1[k]]<=t1[k] < t2[I1[k]+1],
#   and r2[I1[k]+1] is the number of type 2 individuals at risk
#   at the time t1[k] (when there are r1[k] type 1 individuals)
#   Y1=r1[I1]

crossrisk=function(t1,t2,r1,r2){
  I1=rep(0,length(t1))
  I2=rep(0,length(t2))
  for(i in seq(length(t1))){
    I1[i]=1+sum(t1[i]>t2)
  }
  for(i in seq(length(t2))){
    I2[i]=1+sum(t2[i]>t1)
  }
  list(I1,I2,r1[I2],r2[I1])
}

r1=tongue.fit$n.risk[seq(n1)]
r2=tongue.fit$n.risk[seq(n1+1,n1+n2)]

r1=c(r1,r1[n1]-tongue.fit$n.event[n1]-tongue.fit$n.censor[n1])
r2=c(r2,r2[n2]-tongue.fit$n.event[n1+n2]-tongue.fit$n.censor[n1+n2])
t1=tongue.fit$time[seq(n1)]
t2=tongue.fit$time[seq(n1+1,n1+n2)]

cr=crossrisk(t1,t2,r1,r2)

Y1=c(r1[-n1],cr[[3]])
Y2=c(cr[[4]],r2[-n2])
  # Note: r1 and r2 had an extra count added on to make crossrisk work
d1=c(tongue.fit$n.event[seq(n1)],rep(0,n2))
d2=c(rep(0,n1),tongue.fit$n.event[seq(n1+1,n1+n2)])

t=c(t1,t2)

# We have to deal with the problem of ties between times for the two groups


dup1=which(duplicated(t,fromLast=TRUE))
dup2=which(duplicated(t))
ndup=length(dup1)

# Type 2 Event counts are removed from the second appearance
# and placed in the first appearance
d2[dup1]=d2[dup2]
```

```
d2=d2[-dup2]
d1=d1[-dup2]

# Type 2 at-risk counts are removed from the second appearance
# and placed in the first appearance
Y2[dup1]=Y2[dup2]
Y2=Y2[-dup2]
Y1=Y1[-dup2]
t=t[-dup2]

tord=order(t)
t=t[tord]  #put times in order
## Now put everything else in the same order
Y=Y[tord]
Y1=Y1[tord]
Y2=Y2[tord]
d=d[tord]
d1=d1[tord]
d2=d2[tord]

Y=Y1+Y2
d=d1+d2

# Product of number at risk
atriskprod=Y1*Y2
includes=(atriskprod>0)&(d>0)
# We only get contributions if someone's at risk and events occurred at that time

Y=Y[includes]
Y1=Y1[includes]
Y2=Y2[includes]
d=d[includes]
d2=d2[includes]
d1=d1[includes]

t=t[includes]

wLR=Y1*Y2/Y
p=1
q=0

S=c(1,cumprod((Y-d)/Y))[-length(Y)]  #K-M estimator for survival
wFH=(1-S)^q*S^p*wLR

# Now compute the test statistic

w=wLR

M=w*(d1/Y1-d2/Y2)
sigma=w*w*d*(Y-d)/Y2/Y1/(Y-1)
sK=d*Y1*Y2*(Y-d)/Y^2/(Y-1)

Z=sum(M)/sqrt(sum(sigma))
```

```
> Z
[1] -1.670246
```

3. (a) Crude estimates from the data are subject to stochastic fluctuation. Smoothing (graduating) the estimates may make more reliable predictions.

   (b) $\mu_x = a + be^{\alpha x}$ for Gompertz–Makeham. This is generally considered a reasonable model for the hazard rate (force of mortality) from middle age onward. Note, though, that the mortality rate doubling times (which would be approximately constant under Gompertz–Makeham) lengthen progressively. The parameters a, b, α will have to be fitted from the data.

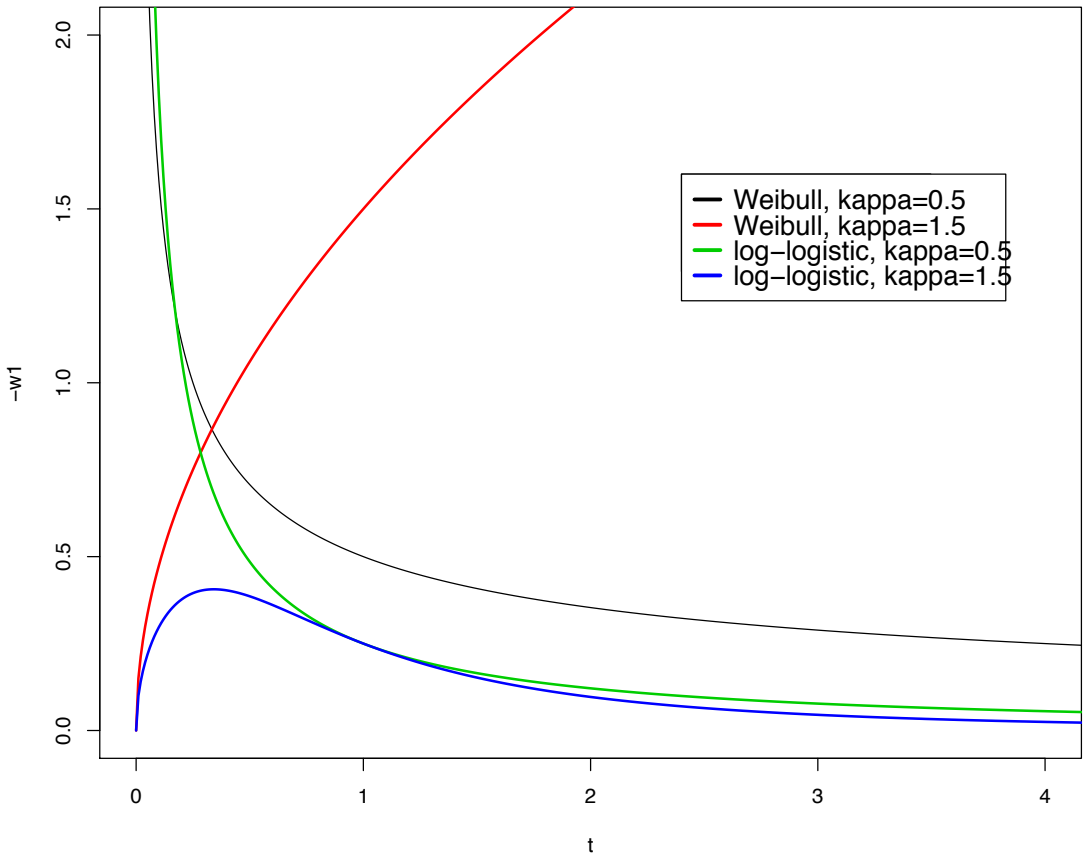
   We apply the chi-squared test. To begin with, we combine the last two rows to have ≥ 5 expected deaths in each row. The last row becomes

   $$99 \qquad 17.5 \qquad 5 \qquad 0.2857 \qquad 0.3027 \qquad -0.1293$$

   (We interpolate by weighting the two rows by their central exposed to risk.) The χ^2 statistic is then 4.96 on 8 observations. Since we have estimated 3 parameters, we compare this to the table with 5 degrees of freedom, obtaining p-value 0.42.

   To test for bias we use the cumulative deviations test, obtaining $Z = 0.96$, and a p-value of 0.3375. Thus, the model seems to fit. Notice that graduated hazard is generally lower — it is strongly affected by the mortality plateau a very late ages — which would lead to an overestimate of benefits paid. This is a relatively good error to make, though it would be reversed if the company were selling life insurance!

4. (a) The plot is:



(b) One could consider using the log-logistic or the log-normal.

(c) The hazard function is $h(t) = \kappa(\mathrm{e}^{\beta\cdot x})^\kappa t^{\kappa-1}$ and the survival function is $\mathrm{e}^{-(\mathrm{e}^{\beta\cdot x}t)^\kappa}$. Hence the log likelihood is

$$\ell(,\kappa,\beta) = n\log\kappa + \kappa\sum \beta\cdot x_i + (\kappa-1)\sum \log t_i - \sum\left(\mathrm{e}^{\beta\cdot x_i}t_i\right)^\kappa.$$

We include $x_{i0} = 1$ for each individual, so that e^{β_0} will be the baseline value of ρ.
The MLE must satisfy

$$\begin{aligned} 0 &= \frac{\partial\ell}{\partial\kappa} = \frac{n}{\kappa} + \sum \beta\cdot x_i + \sum \log t_i - \kappa\sum \mathrm{e}^{\kappa\beta\cdot x_i}t_i^{\kappa-1}, \\ 0 &= \frac{\partial\ell}{\partial\beta_j} = \kappa\sum_i x_{ij}\Big(1 - \left(\mathrm{e}^{\beta\cdot x_i}t_i\right)^\kappa\Big). \end{aligned}$$

Asymptotically, the estimators will be normally distributed. If some observations are right-censored, the log likelihood becomes

$$\ell(\kappa,\beta) = n_d\log\kappa + n_d\kappa\log\rho_0 + \kappa\sum \delta_i\beta\cdot x_i + (\kappa-1)\sum \delta_i\log t_i - \sum\left(\rho_0\mathrm{e}^{\beta\cdot x_i}t_i\right)^\kappa$$

where n_d is the number of (uncensored) events observed.

5. (a) Assuming no ties, the partial likelihood is constructed by computing the probability that the subjects failed in exactly the order observed, conditioned on the times observed.
The proportional hazards (PH) assumption says that subject i has hazard rate $h_i(x) = r_i h_0(x)$ at time x, where h_0 is an unspecified baseline hazard. In the regression approach, we think of r_i as a function $r(y_i)$ of a vector y_i of covariates. The linear approach is to suppose $\phi(r(y)) = \beta\cdot y$, where ϕ is the *link function* and β is a vector of parameters to estimate. In the Cox model we use the logarithmic link function, so that $r(y) = e^{\beta\cdot y}$. The partial likelihood is defined as

$$L_P(\beta; y) := \prod_{t_i} \frac{e^{\beta y_{(i)}}}{\sum_{j\in R_j} e^{\beta y_j}},$$

where $x_{(i)}$ represents the covariates of the subject failing at time t_i and R_i is the *risk set*, of those subjects at risk at t_i.
We use L_P as though it were a likelihood. We compute the parameters $\hat\beta$ that maximise L_P. Under the assumption that the observations came from the distribution given by this model with some (unknown) parameter β, the estimate $\hat\beta$ is asymptotically normal, with mean β and variance matrix that may be estimated by

$$\left[E\left(-\frac{\partial^2\ell_P}{\partial\beta\partial\beta^T}\right)\right]^{-1}, \text{ where } \ell_P = \log L_P.$$

(b) The hazard ratio is

$$\begin{aligned} \frac{h(\text{clinic} = 1, \text{prison=}0)}{h(\text{clinic} = 0, \text{prison=}1)} &= \frac{e^{\hat\beta\cdot y_1}}{e^{\hat\beta\cdot y_2}} \\ &= \frac{e^{-1.009}}{e^{0.327}} \\ &= 0.263. \end{aligned}$$

(c) The log hazard ratio for prison/no prison is 0.327, with standard error 0.167. A 95% confidence interval for the coefficient is $0.327 \pm 1.96\cdot 0.167 = (0.0, 0.654)$. Thus a 95% confidence intervalfor the hazard ration is $e^{(0.0,0.654)} = (1.00, 1.92)$.

6. (a) The times t_i are $50, 52, 58, 61, 67, 68, 70, 72, 75$. A full description of the risk sets requires that we describe exactly which individuals are at risk. We number the males as $M1, \ldots, M12$ and $F1, \ldots, F12$. We have then the risk sets

$$\begin{aligned}
R_1 &= \{M1, M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, F1, F2, F3, F4, F5, F6, \\
&\qquad F7, F8, F9, F10, F11, F12\} \\
R_2 &= \{M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, F1, F2, F3, F4, F5, F6, F7, \\
&\qquad F8, F9, F10, F11, F12\} \\
R_3 &= \{M4, M5, M6, M7, M8, M9, M10, M11, M12, F1, F3, F4, F5, F6, F7, F8, F9, \\
&\qquad F10, F11, F12\} \\
R_4 &= \{M4, M5, M6, M7, M8, M9, M10, M11, M12, F1, F3, F4, F5, F6, F7, F8, F9, \\
&\qquad F10, F12\} \\
R_5 &= \{M4, M5, M6, M7, M8, M9, M10, M12, F3, F4, F5, F6, F7, F8, F9, F10, F12\} \\
R_6 &= \{M4, M5, M6, M7, M9, M10, M12, F3, F4, F5, F6, F7, F8, F9, F10, F12\} \\
R_7 &= \{M5, M7, M9, M10, M12, F3, F4, F5, F7, F8, F9, F10, F12\} \\
R_8 &= \{M5, M9, M10, M12, F4, F5, F9, F10, F12\} \\
R_9 &= \{M10, M12, F4, F10\}.
\end{aligned}$$

Note that there is some ambiguity in breaking ties. When an observation is censored at time t_j we must decide whether to treat the censoring as having occurred just after or just before t_j: that is, was the individual available to have been counted if they had died at time t_j or not? By convention, unless otherwise indicated, we choose the former: Thus, for instance, R_9 is the set of individuals at risk at time 75, and it includes M12, who was censored at age 75. Either one is acceptable, depending on details of the study.

Since we are interested only in the binary covariate of gender, we need only consider the risk sets as counting the numbers of males and females, coded as $R_j = (m_j, f_j)$. We may then summarise them as

$$\begin{aligned}
R_1 = (11, 12)\ R_2 = (10, 12)\ R_3 = (9, 11)\ R_4 = (9, 10)\ R_5 = (8, 9) \\
R_6 = (7, 9)\ R_7 = (5, 8)\ R_8 = (4, 5)\ R_9 = (2, 2).
\end{aligned}$$

(b) Using the notation as above, the covariates at the event times are $x_{i_j} = (0, 1, 1, 0, 0, 0, 1, 1, 1)$ – coding female as 1 and male as 0 — we have the partial likelihood being

$$L_P = \prod_{j=1}^{9} \frac{e^{\beta x_{i_j}}}{m_j + e^{\beta} f_j} = e^{5\beta} \prod_{j=1}^{9} \left(m_j + e^{\beta} f_j \right)^{-1}. \tag{B.9}$$

A plot of this function is in Figure B.2. The maximum likelihood is attained at $\beta = -0.042$.

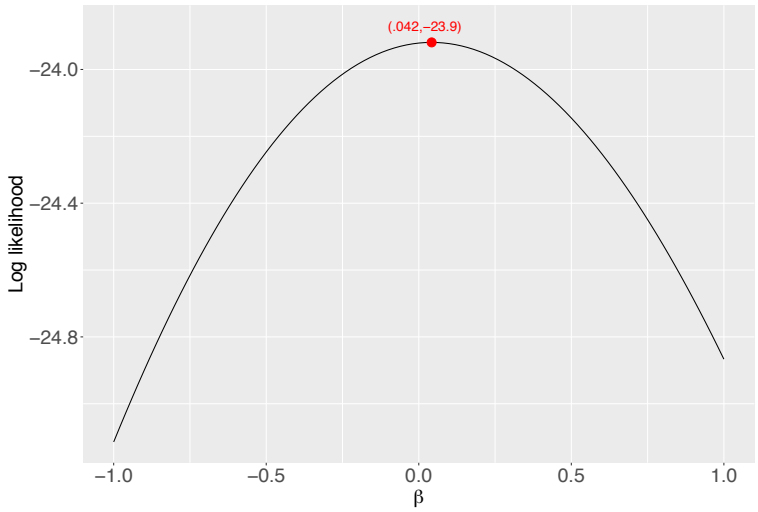


Figure B.2: Plot of logarithm of partial likelihood given by (B.9).

(c) We collect the relevant data in a table.

		event time								
		50	52	58	61	67	68	70	72	75
Male	d_j^m	1	0	0	1	1	1	0	0	0
	m_j	11	10	9	9	8	7	5	4	2
Female	d_j^f	0	1	1	0	0	0	1	1	1
	f_j	12	12	11	10	9	9	8	5	2
Total	d_j	1	1	1	1	1	1	1	1	1
	n_j	23	22	20	19	17	16	13	9	4
	$\hat{S}_{\text{male}}(t_j)$	0.909	0.909	0.909	0.808	0.707	0.606	0.606	0.606	0.606
	$\hat{S}(t_{j-1})$	1	0.957	0.913	0.867	0.822	0.773	0.725	0.669	0.595

Male mortality is estimated by the Kaplan–Meier estimator.

(d) To compute the Fleming–Harrington statistic we will need to compute $\hat{S}$ for the combined population, which we have done at the end of the above table. (Note that we have shifted these by one, since the FH statistic uses the survival **up to** time t_j, which is $\hat{S}(t_{j-1})$.) Plugging these into the formula

$$Z_{\text{LR}} = \frac{\sum_{j=1}^{9}\left(d_j^m - n_j^m \frac{d_j}{n_j}\right)}{\sqrt{\sum_{j=1}^{9} \frac{n_j^m n_j^f (n_j - d_j) d_j}{n_j^2 (n_j - 1)}}},$$

we get $Z = -.063$, which should be like a draw from a normal distribution if the male and female survival times were drawn from the same distribution. In fact, we get a p-value of $1 - 2\Phi(.063) = .95$.

(e) For the Fleming–Harrington test we down-weight the later times, when very few are at risk,

substituting

$$Z_{FH} = \frac{\sum_{j=1}^{9} \hat{S}(t_{j-1}) \left(d_j^m - n_j^m \frac{d_j}{n_j} \right)}{\sqrt{\sum_{j=1}^{9} \hat{S}(t_{i-1})^2 \frac{n_j^m n_j^f (n_j - d_j) d_j}{n_j^2 (n_j - 1)}}} = 0.105,$$

yielding a p-value for the two-sided test of 0.92. In either case, of course, we would not reject the null hypothesis. Of course, this is not surprising, as the sample is very small.

Note that this analysis could be improved by taking account of the pairing of twins.

(f) Death due to other causes is unlikely to be independent of CHD. Hence, non-informative censoring is questionable.