

B.6 Model testing; Proportional-hazards and accelerated life-times

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)

$$\sup_{(\rho, \alpha)} \ell(\rho, \alpha) - \sup_{\rho} \ell(\rho, 1) \sim \chi_1^2.$$

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, $S_i(x) = S_0(\rho_i x)$ 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.5. In figure B.2 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

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

For the treatment group this is For the Weibull distribution we have

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

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

t_i	n_i	d_i	\hat{h}_i	$\hat{S}(t)$
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_i	n_i	d_i	\hat{h}_i	$\hat{S}(t)$
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.5: 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
$\hat{\alpha}$	1.35	1.37
ℓ_{weib}^*	-41.66	-64.92

The log likelihood ratio for the treatment group is thus $(-41.66) - (-42.17) = 0.51$, and for the control group it is 1.43. Comparing these to the χ^2 distribution with 1 degree of freedom, we see that the cutoff for rejecting the null hypothesis that $\alpha = 1$ at the 0.05 significance level would be 3.84. Thus, we cannot reject the null hypothesis for either group.

2. (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

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

where $x_{(i)}$ represents the covariates of the subject failing at time t_i and R_i is the *risk set* 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

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

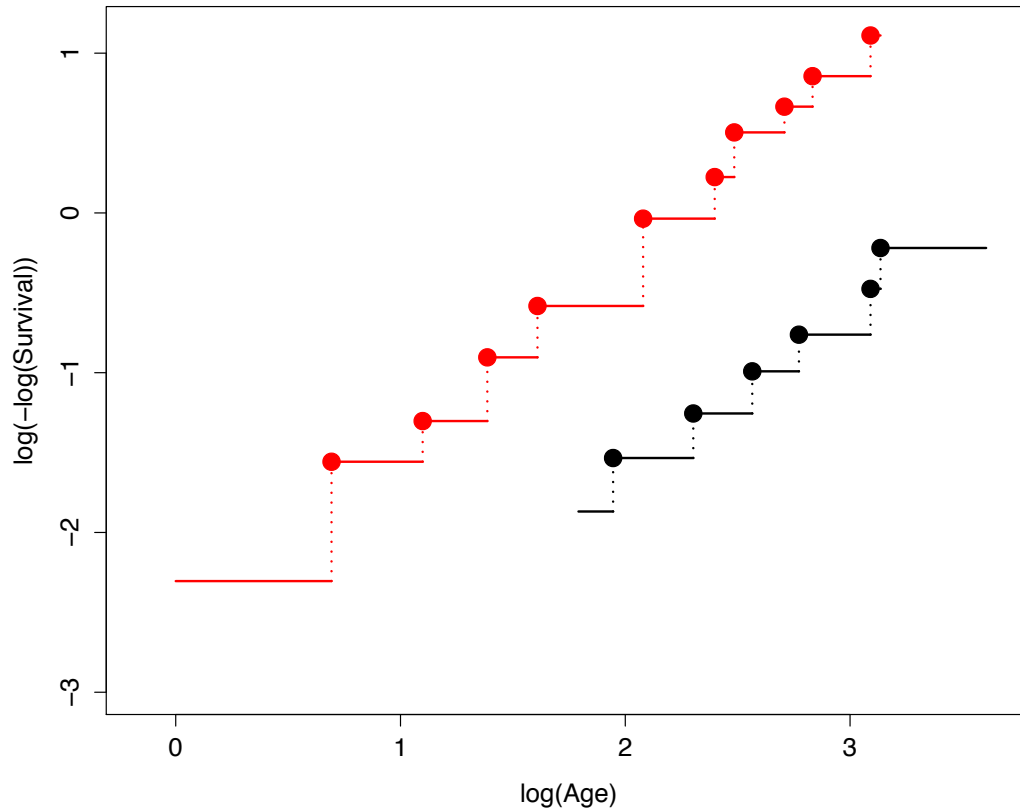


Figure B.2: Plot of estimated survival for Gehan leukaemia data. The control group is in red, the treatment group is black.

(b) The hazard ratio is

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

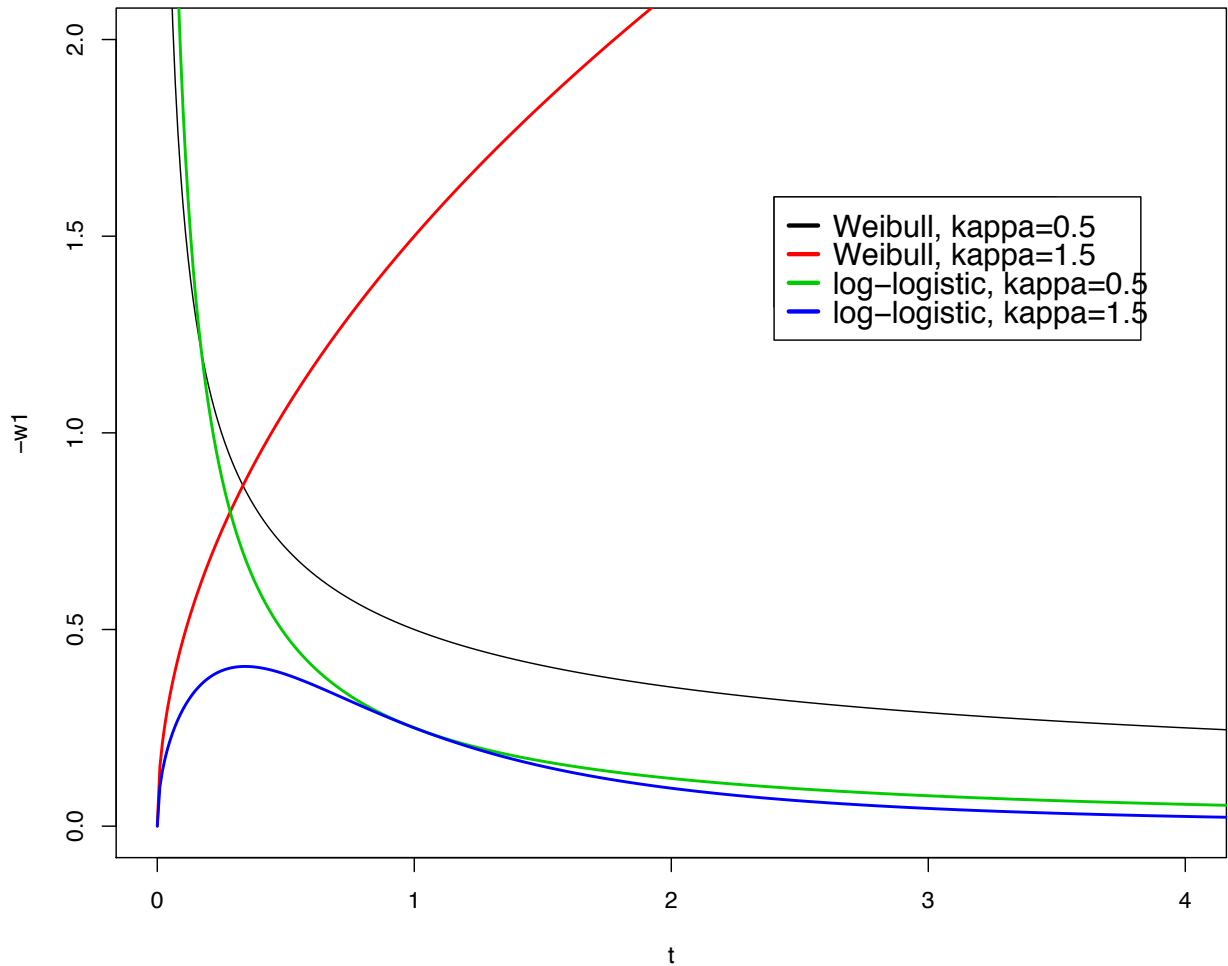
(c) The log hazard ratio for prison/no prison is 0.327, with standard error 0.167. A 95% confidence interval for the coefficient is $0.327 \pm 1.96 \cdot 0.167 = (0.0, 0.654)$. Thus a 95% confidence interval for the hazard ratio is $e^{(0.0, 0.654)} = (1.00, 1.92)$.

3. (a) The plot is:

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

(c) For a given choice of the parameters α, β , we transform $Z_i := \alpha(\log y_i + \beta \cdot x_i)$. Then Z_i has the standard survival function $S_Z(z) = e^{-e^z}$. The log likelihood is

$$\ell(\alpha, \beta) = \sum z_i(\alpha, \beta) - e^{z_i(\alpha, \beta)}.$$



The MLE must satisfy

$$0 = \sum \frac{\partial z_i}{\partial \alpha} (1 - e^{z_i}) \Rightarrow \sum e^{z_i} z_i = \sum z_i$$

$$0 = \sum \frac{\partial z_i}{\partial \beta} (1 - e^{z_i}) \Rightarrow \bar{x} = \frac{1}{n} \sum x_i e^{z_i}.$$

Note that if x_i is a single binary categorical variable (=0 or 1), then we are looking for a simultaneous solution to

$$e^{\beta} = \left(\frac{1}{n_1} \sum_{i:x_i=1} T_i^{\alpha} \right)^{1/\alpha},$$

$$1 = \frac{1}{n_0} \sum_{i:x_i=0} T_i^{\alpha},$$

where n_k is the number of i such that $x_i = k$. Asymptotically, the estimators will be normally distributed. If some observations are right-censored, their contribution to the log likelihood is $-e^{z_i}$ in place of $z_i - e^{z_i}$.

4. (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 $M1, \dots, M12$ and $F1, \dots, F12$. We have then the risk sets

$$R_1 = \{M1, M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, F1, F2, F3, F4, F5, F6, \\ F7, F8, F9, F10, F11, F12\}$$

$$R_2 = \{M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, F1, F2, F3, F4, F5, F6, F7, \\ F8, F9, F10, F11, F12\}$$

$$R_3 = \{M4, M5, M6, M7, M8, M9, M10, M11, M12, F1, F3, F4, F5, F6, F7, F8, F9, \\ F10, F11, F12\}$$

$$R_4 = \{M4, M5, M6, M7, M8, M9, M10, M11, M12, F1, F3, F4, F5, F6, F7, F8, F9, \\ F10, F12\}$$

$$R_5 = \{M4, M5, M6, M7, M8, M9, M10, M12, F3, F4, F5, F6, F7, F8, F9, F10, F12\}$$

$$R_6 = \{M4, M5, M6, M7, M9, M10, M12, F3, F4, F5, F6, F7, F8, F9, F10, F12\}$$

$$R_7 = \{M5, M7, M9, M10, M12, F3, F4, F5, F7, F8, F9, F10, F12\}$$

$$R_8 = \{M5, M9, M10, M12, F4, F5, F9, F10, F12\}$$

$$R_9 = \{M10, M12, F4, F10\}.$$

Note that there is some ambiguity in breaking ties. When an observation is censored at time t_i we must decide whether to treat the censoring as having occurred just after or just before t_i : that is, was the individual available to have been counted if they had died at time t_i or not? We have chosen 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 — though details of the study may suggest one or the other interpretation — but it should be specified.

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 $R_i = (m_i, f_i)$. We may then summarise them as

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

- (b) Using the notation as above, and setting the vector of covariates to be $x = (1, 0, 0, 1, 1, 1, 0, 0, 0)$ — coding female as 0 and male as 1 — we have the partial likelihood being

$$L_P = \prod_{i=1}^9 \frac{e^{\beta x_i}}{f_i + e^{\beta m_i}} = e^{4\beta} \prod_{i=1}^9 (f_i + e^{\beta m_i})^{-1}. \quad (3)$$

A plot of this function is in Figure B.3. The maximum likelihood is attained at $\beta = -0.042$.

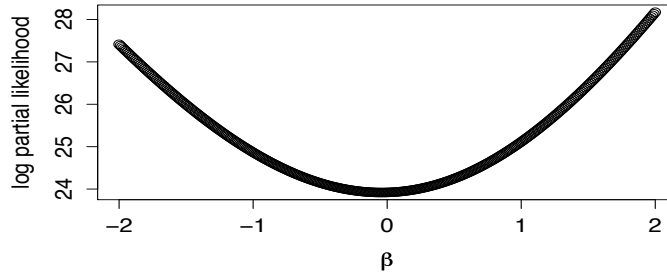


Figure B.3: Plot of negative logarithm of partial likelihood given by (3).

(c) We need to compute first \hat{S} for the combined population. We have

		event time								
		50	52	58	61	67	68	70	72	75
Male	d_m	1	0	0	1	1	1	0	0	0
	m_i	11	10	9	9	8	7	5	4	2
Female	d_i^f	0	1	1	0	0	0	1	1	1
	f_i	12	12	11	10	9	9	8	5	2
Total	d_i	1	1	1	1	1	1	1	1	1
	n	23	22	20	19	17	16	13	9	4
$\hat{S}(t_{i-1})$		1	0.957	0.913	0.867	0.822	0.773	0.725	0.669	0.595

Plugging these into the formula

$$Z = \frac{\sum_{i=1}^9 \left(d_i^m - n_i^m \frac{d_i}{n_i} \right)}{\sqrt{\sum_{i=1}^9 \frac{n_i^m n_i^f (n_i - d_i) d_i}{n_i^2 (n_i - 1)}}$$

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 $1 - 2\Phi(.063) = .95$.

For the Fleming-Harrington test we down-weight the later times, when very few are at risk, substituting

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

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.

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

5. (a) This is an accelerated lifetime model, since $S(t; x) = S_0(te^{\beta x}) = \exp(-H_0(te^{\beta x}))$. We have

$$\begin{aligned} P(T(x) \geq t) &= S_0(te^{\beta x}) \\ &= P(T(0) \geq te^{\beta x}) \\ &= P(e^{-\beta x} T(0) \geq t). \end{aligned}$$

- (b) It follows that $\ln T(x)$ and $\ln T(0) - \beta x$ have identical distributions. Set $\ln T(0) = E \ln T(0) + \epsilon$ (so $E(\epsilon) = 0$). It follows that

$$\ln T(x_i) + \beta x_i = E[\ln T(0)] + \epsilon_i,$$

where ϵ_i has the same distribution as ϵ . Consequently

$$\ln T(x_i) = E[\ln T(0)] - \beta x_i + \epsilon_i,$$

and $\alpha = E \ln T(0)$.

- (c) Assuming that $\text{Var} \epsilon = \sigma^2 < \infty$, we can use least-squares estimation. Let $Z_j = \ln T(x_j)$. Then

$$\mathbf{Z} = \begin{pmatrix} 1 & -x_1 \\ \vdots & \vdots \\ 1 & -x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \epsilon =: \mathbf{X} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \epsilon, \quad \mathbf{X}^T \mathbf{X} = \begin{pmatrix} n & -n\bar{x} \\ -n\bar{x} & S_{xx} \end{pmatrix},$$

where $S_{xx} = \sum x_i^2$. We solve these to yield

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \frac{1}{nS_{xx} - n^2\bar{x}^2} \begin{pmatrix} S_{xx} & n\bar{x} \\ n\bar{x} & n \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \\ -x_1 & \cdots & -x_n \end{pmatrix} \mathbf{Z} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Z}$$

$$\text{var}(\hat{\alpha}, \hat{\beta}) = (\mathbf{X}^T \mathbf{X})^{-1} \text{var}(\epsilon_i) = (\mathbf{X}^T \mathbf{X})^{-1} \text{var}(\ln T(0)).$$

We can check for AL as usual by plotting \hat{S} against $\ln t$ if we are able to group the variables x_j into levels. But also note that $E \ln T(x_j) = \alpha - \beta x_j$, so plotting $\ln T(x_j)$ against x_j would be helpful. We might look at residuals, though this is problematic, as ϵ is unlikely to be normally distributed