Today

- Regression with survival data
- In the news: biomarkers and death; PLOS 1 paper
- Model formulation CD Ch. 6

- April 4: Questions re any HW, re grading, re 2012 final test; Summary of course notes

- Office Hours: April 8, 9, 10; 3 – 5
- HW 4: due April 11
- Final Exam: April 11 2:00 – 5:00 pm SS 1085
  - 4 questions
  - one theory question
  - one applied question
  - one question from HW
  - one question about a study
  - one question with computer output
- SM: 9.1, 9.2.1, 9.2.2 (to end p.431), 9.3.1, 9.4.2; 10.1, 10.2, 10.3, 10.4, 10.6, 10.7.1 (skip p.529-530), 10.7.2, 10.7.3, 10.8.1, 10.8.2 (skip log rank test, time-dependent covariates)
- C& D: from slides only – Ch 1, 2, 7.2, 7.3, 6.5

Regression with survival data

SM §10.8

- response y^0 is time to 'failure', or 'survival' time $y^0 \geq 0$
- density function $f(\cdot)$, distribution function $F(\cdot)$

- survivor function $S(\cdot) = 1 - F(\cdot)$
- hazard function $h(\cdot) = f(\cdot)/\{S(\cdot)\}$
- cumulative hazard function $H(y) = \int_0^y h(u)du = -\log S(y)$

- parametric models: exponential, Weibull, Gamma, log-normal, log-logistic

  SM §5.4

- random censoring: $C \sim G(\cdot)$, independently of Y^0
- observe $(y_j, d_j), j = 1, \ldots, n$

- $y_j = \min(y_j^0, c_j), \qquad d_j = \mathbb{1}(y_j^0 \leq c_j)$

... regression with survival data

SM §10.8

- $y_j = \min(y_j^0, c_j), \qquad d_j = \mathbb{1}(y_j^0 \leq c_j)$
- data $(y_j, d_j, x_j), j = 1, \ldots, n;$ x_j explanatory variables

- survivor $S(\cdot; x, \beta)$, density $f(\cdot; x, \beta)$, hazard $h(\cdot; x, \beta)$

- log-likelihood

$$\ell(\beta; y, d) = \sum_{j=1}^{n} \{d_j \log h(y_j; x_j, \beta) - H(y_j; x_j, \beta)\}$$

SM (5.26)

- maximum likelihood estimates $\hat{\beta}$
  observed information function $-\ell''(\hat{\beta})$

- residuals

$$r_j = H(y_j; x_j, \hat{\beta}) + 1 - d_j$$

Example 10.36

Group 1						Group 2					
	x	y		x	y		x	y		x	y
1	3.36	65	10	3.85	143	18	3.64	56	27	4.45	3
2	2.88	156	11	3.97	56	19	3.48	65	28	4.49	8
3	3.63	100	12	4.51	26	20	3.60	17	29	4.41	4
4	3.41	134	13	4.54	22	21	3.18	7	30	4.32	3
5	3.78	16	14	5.00	1	22	3.95	16	31	4.90	30
6	4.02	108	15	5.00	1	23	3.72	22	32	5.00	4
7	4.00	121	16	4.72	5	24	4.00	3	33	5.00	43
8	4.23	4	17	5.00	65	25	4.28	4			
9	3.73	39				26	4.43	2			

Table 10.22 Survival times y (weeks) for two groups of acute leukaemia patients, together with $x = \log_{10}$ white blood cell count at time of diagnosis (Feigl and Zelen, 1965). Patients in group 1 had Auer rods and/or significant granulation of the leukaemic cells in the bone marrow at the time of diagnosis; those in group 2 did not.

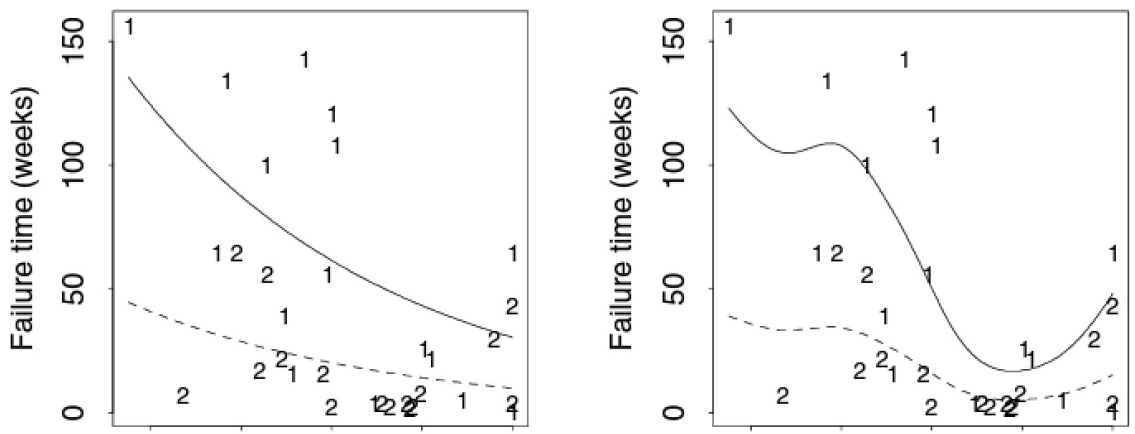


Figure 10.21 Plots of data and fitted means for generalized linear (left) and generalized additive (right) models fitted to two groups of survival times for leukaemia patients: group 1 (solid); group 2 (dashed).

... example 10.36

```
> library(SMPracticals)
> library(survival)
> data(leuk)
> head(leuk)
    wbc      ag time
1  2300 present   65
2   750 present  156
3  4300 present  100
4  2600 present  134
5  6000 present   16
6 10500 present  108
> with(leuk,log10(wbc[1:5]))
[1] 3.361728 2.875061 3.633468 3.414973 3.778151
leuk.glm <- glm(time ~ ag + log10(wbc), data = leuk, family = Gamma(link = "log"))
> summary(leuk.glm, dispersion = 1)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.8154     1.2932   4.497 6.89e-06 ***
agpresent     1.0176     0.3492   2.914  0.00357 **
log10(wbc)   -0.7009     0.3036  -2.308  0.02097 *
---

(Dispersion parameter for Gamma family taken to be 1)

    Null deviance: 58.138  on 32  degrees of freedom
Residual deviance: 40.319  on 30  degrees of freedom

> summary(leuk.glm)
...
(Dispersion parameter for Gamma family taken to be 1.087715)
```

... example 10.36

```
> leuk.surv <- survreg(Surv(time, rep(1,length(time))) ~ log10(wbc) + ag, data = leuk, dist= "
> summary(leuk.surv)

Call:
survreg(formula = Surv(time, rep(1, length(time))) ~ log10(wbc) +
    ag, data = leuk, dist = "exponential")
             Value Std. Error     z        p
(Intercept)  5.815      1.263  4.60 4.15e-06
log10(wbc)  -0.701      0.286 -2.45 1.44e-02
agpresent    1.018      0.364  2.80 5.14e-03

Scale fixed at 1

> leuk.surv2 <- survreg(Surv(time,rep(1,length(time)))~pspline(log10(wbc),df=0) + ag,
+ data = leuk, dist= "exponential" )

## see help file for survreg

> leuk.gam <- gam(time ~ s(log10(wbc)) + ag, data = leuk, family = Gamma(link = "log") )
> summary(leuk.gam, dispersion = 1)
...
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.7270     0.2524   10.80  < 2e-16 ***
agpresent     1.1424     0.3547    3.22  0.00128 **
...
Approximate significance of smooth terms:
                edf Ref.df Chi.sq p-value
s(log10(wbc)) 3.236  3.967  14.59 0.00553 **
```

Proportional hazards model

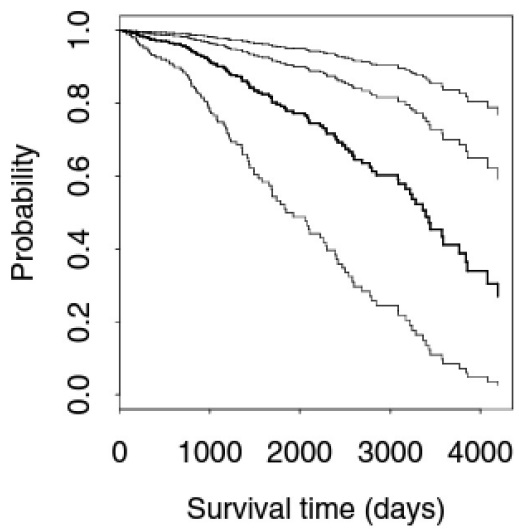
- hazard function $h(y; x, \beta) = h_0(y) \exp(x^{\mathrm{T}}\beta)$
- survivor function $S(y; x, \beta) = S_0(y)^{\exp(x^{\mathrm{T}}\beta)}$
- log-likelihood $\sum_{j=1}^{n} \{d_j x_j^{\mathrm{T}}\beta + \log h_0(y_j) - H_0(y_j)\exp(x_j^{\mathrm{T}}\beta)\}$
- partial likelihood

$$L_{\text{part}}(\beta) = \prod_{j=1}^{n} \left\{ \frac{\exp(x_j^{\mathrm{T}}\beta)}{\sum_{i \in \mathcal{R}_j} \exp(x_i^{\mathrm{T}}\beta)} \right\}^{d_j}$$

- derived in SM §10.8 as profile likelihood, treating $h_0(y_1), \ldots, h_0(y_n)$ as n nuisance parameters
- $\mathcal{R}_j$ risk set at time y_j^-
- all observations available to fail just before the time of the jth failure
- adjustments for ties, see p.544

... proportional hazards model

- estimation of the hazard function and survival probability
- $$\widehat{H}_0(y) = \sum_{j:y_j \leq y} \frac{d_j}{\sum_{i \in \mathcal{R}_j} \exp(x_i^{\mathrm{T}} \hat{\beta})}$$
- $$\widehat{S}_0(y) = \prod_{j:y_j \leq y} \left(1 - \frac{d_j}{\sum_{i \in \mathcal{R}_j} \exp(x_i^{\mathrm{T}} \hat{\beta})}\right)$$



... example 10.36

```
> leuk.ph <- coxph(Surv(time,rep(1,length(time)))~  ag + log10(wbc) , data = leuk)
> summary(leuk.ph)
...
              coef exp(coef) se(coef)      z Pr(>|z|)
agpresent  -1.0691    0.3433   0.4293 -2.490  0.01276 *
log10(wbc)  0.8467    2.3318   0.3132  2.703  0.00687 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
agpresent     0.3433     2.9126     0.148    0.7964
log10(wbc)    2.3318     0.4288     1.262    4.3083



> leuk.ph2 <- coxph(Surv(time,rep(1,length(time)))~  ag + log10(wbc) ,
+ ties="breslow", data = leuk)
> summary(leuk.ph2)
...
              coef exp(coef) se(coef)      z Pr(>|z|)
agpresent  -1.0176    0.3614   0.4235 -2.403  0.01626 *
log10(wbc)  0.8296    2.2924   0.3120  2.659  0.00785 **
---


> plot(survfit(leuk.ph))
```

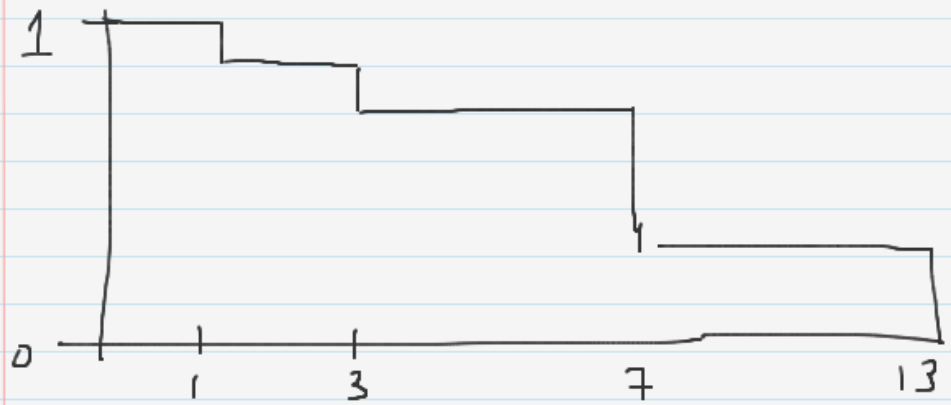
Kaplan-Meier estimation of $S(\cdot)$

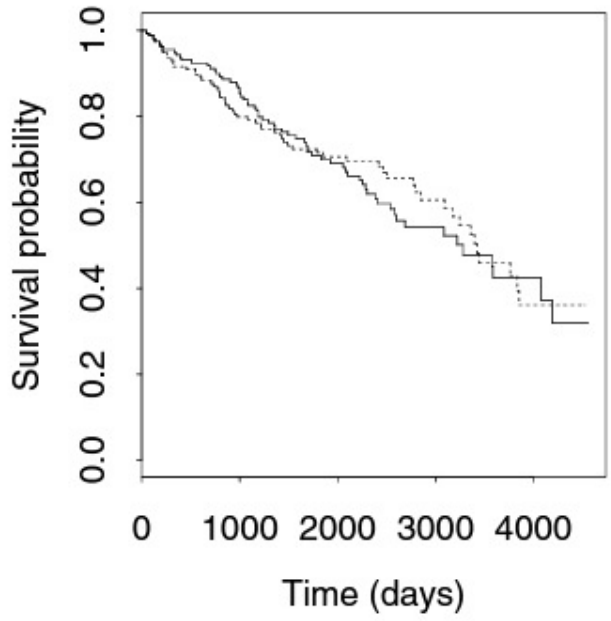
- nonparametric estimation of survivor function
- censored data analogue of empirical cumulative distribution function
- $$\widehat{S}(y) = \prod_{i:y_i \leq y} \left(1 - \frac{d_i}{r_i}\right)$$
- with grouped data:
  - d_i number of items failing at time y_i
- with continuous data
- $$\widehat{S}(y) = \prod_{i:y_i \leq y} \left(1 - \frac{1}{r_i}\right)^{d_i}$$
- $$\widehat{\text{var}}\{\log \widehat{S}(y)\} = \sum_{i:y_i \leq y} \frac{d_i}{r_i(r_i - d_i)}$$

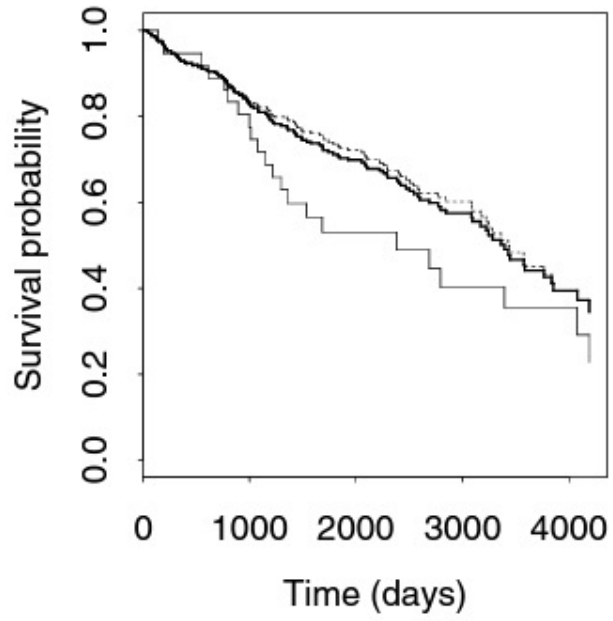
Obs^{+}: 1, 3, 5^{+}, 7, 7^{+}, 13

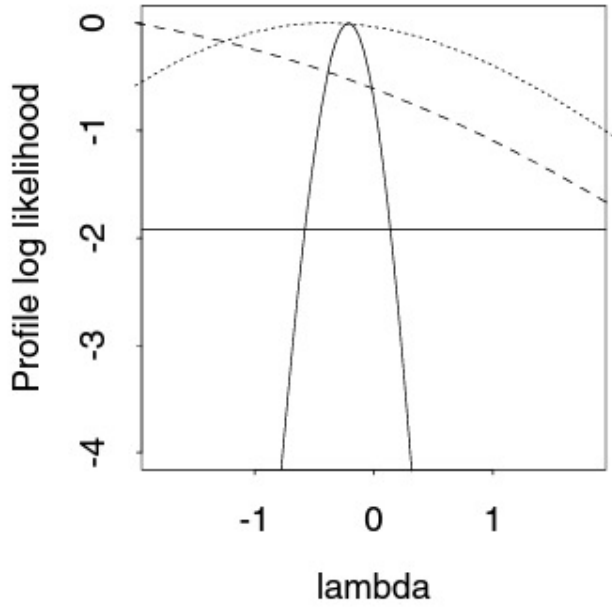
$$\hat{S}(t) = \prod_{y_j \le t} \left(1 - \frac{1}{r_j}\right)^{d_j}$$

$$= \begin{cases} 1 & t < 1 \\ \frac{5}{6} & 1 < t < 3 \\ \frac{5}{6} \times \frac{4}{5} = \frac{4}{6} & 3 < t < 7 \\ \frac{4}{6} \times \frac{2}{3} = \frac{4}{9} & 7 < t < 13 \\ \frac{4}{9} \times \frac{0}{1} = 0 & t > 13 \end{cases}$$









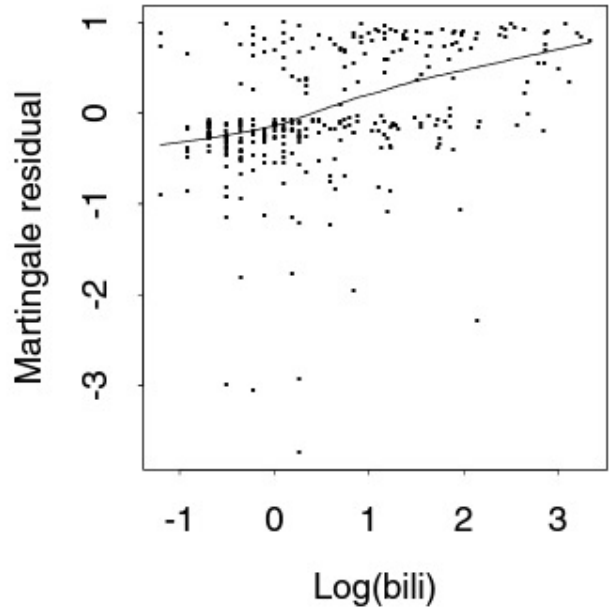


Figure 10.22 PBC data analysis (Fleming and Harrington, 1991). Top left: product-limit estimates for control (solid) and treatment (dots) groups. Top right: estimates of baseline survivor function for data stratified by sex, men (dots), women (solid). The heavy line shows the unstratified estimate. Lower left: profile likelihood for Box–Cox transformations of bilirubin (solid), albumin (dots), and prothrombin time (dashes); the horizontal line indicates 95% confidence limits for the transformation parameter. Lower right: martingale residuals from the model with terms `age`, `log(alb)`, `edtrt`, `log(protime)` against log bilirubin, and lowess smooth with $p = 2/3$.