11
Complex data structures

The focus in this chapter is on problems with complex data structure, in particular those involving dependence and censoring. Modelling is crucial for simplifying and clarifying the structure. Compromises with the likelihood, for example using the marginal, conditional or estimated likelihood, also become a necessity. With some examples we discuss the simplest thing we can do with the data, with as little modelling as possible. This will provide a safety check for results based on more complex models; similar results would give some assurance that the complex model is not off the mark. It is important to recognize in each application the price and the reward of modelling.

11.1 ARMA models

So far we have considered independent observations, where we only need to model the probability of single outcomes. From independence, the joint probability is simply the product of individual probabilities; this is not true with dependent data. Suppose we observe a time series $x_1, \ldots, x_n$; then, in general, their joint density can be written as

$$p_\theta(x_1, \ldots, x_n) = p_\theta(x_1)\ p_\theta(x_2|x_1) \ldots p_\theta(x_n|x_1, \ldots, x_{n-1}).$$

Difficulty in modelling arises as the list of conditioning variables gets long. Generally, even for simple parametric models, the likelihood of time series models can be complicated. To make the problems tractable we need to assume some special structures. The main objective of standard modelling assumptions is to limit the length of the conditioning list, while capturing the dependence in the series.

A minimal requirement for time series modelling is weak stationarity, meaning that x_t has a constant mean, and the covariance $\mathrm{cov}(x_t, x_{t-k})$ is only a function of lag k. To make this sufficient for likelihood construction, we usually assume a Gaussian model.

A large class of time series models that leads to a tractable likelihood is the class of autoregressive (AR) models. A time series is said to be an AR(p) series if

$$x_t = \theta_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + e_t,$$

where the so-called innovation or driving noise e_t is assumed to be an iid series. The stationarity and correlation structure of the time series are

determined solely by the parameters $(\phi_1, \ldots, \phi_p)$. Likelihood analysis of AR models typically assumes e_t's are iid normal.

Example 11.1: An AR(1) model specifies

$$x_t = \theta_0 + \phi_1 x_{t-1} + e_t,$$

where e_t's are iid $N(0, \sigma^2)$. This is equivalent to stating the conditional distribution of x_t given its past is normal with mean $\theta_0 + \phi_1 x_{t-1}$ and variance σ^2. Given $x_1, \ldots, x_n$, the likelihood of the parameter $\theta = (\theta_0, \phi_1, \sigma^2)$ is

$$\begin{aligned} L(\theta) &= p_\theta(x_1) \prod_{t=2}^{n} p_\theta(x_t | x_u, u < t) \\ &= p_\theta(x_1) \prod_{t=2}^{n} p_\theta(x_t | x_{t-1}) \\ &= p_\theta(x_1) \prod_{t=2}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} (x_t - \theta_0 - \phi_1 x_{t-1})^2 \right\} \\ &\equiv L_1(\theta)\ L_2(\theta), \end{aligned}$$

where $L_1(\theta) = p_\theta(x_1)$. The term $L_2(\theta)$ is a conditional likelihood based on the distribution of $x_2, \ldots, x_n$ given x_1. This conditional likelihood is commonly assumed in routine data analysis; it leads to the usual least-squares computations.

How much information is lost by ignoring x_1? Assuming $|\phi_1| < 1$, so the series is stationary with mean μ and variance σ_x^2, we find

$$Ex_t = \theta_0 + \phi_1 Ex_{t-1}$$

or $\mu = \theta_0/(1 - \phi_1)$. Iterated expansion of x_{t-1} in terms its past values yields

$$x_t = \mu + \sum_{i=0}^{\infty} \phi_1^i e_{t-i}$$

so

$$\mathrm{var}(x_t) = \sigma^2 \sum_{i=0}^{\infty} \phi_1^{2i}$$

or $\sigma_x^2 = \sigma^2/(1 - \phi_1^2)$. Therefore the likelihood based on x_1 is

$$L_1(\theta) = (2\pi\sigma^2)^{-1/2} (1 - \phi_1^2)^{1/2} \exp\left[-\frac{1 - \phi_1^2}{2\sigma^2} \{x_1 - \theta_0/(1 - \phi_1)\}^2 \right].$$

Hence x_1 creates a nonlinearity in the estimation of ϕ_1. If ϕ_1 is near one, which is the boundary of nonstationarity, the effect of nonlinearity can be substantial.

Figure 11.1 shows the likelihood of ϕ for simulated data with $n = 100$. For simplicity it is assumed $\theta_0 = 0$ and $\sigma^2 = 1$, the true values used in the simulation. For the time series in Figure 11.1(a), the full and conditional likelihoods show little difference. But when $\widehat{\phi}$ is near one, as shown in Figure 11.1(d), the difference is evident. The curvatures of the likelihoods at the maximum are similar, so we cannot measure the effect of x_1 using the Fisher information. □

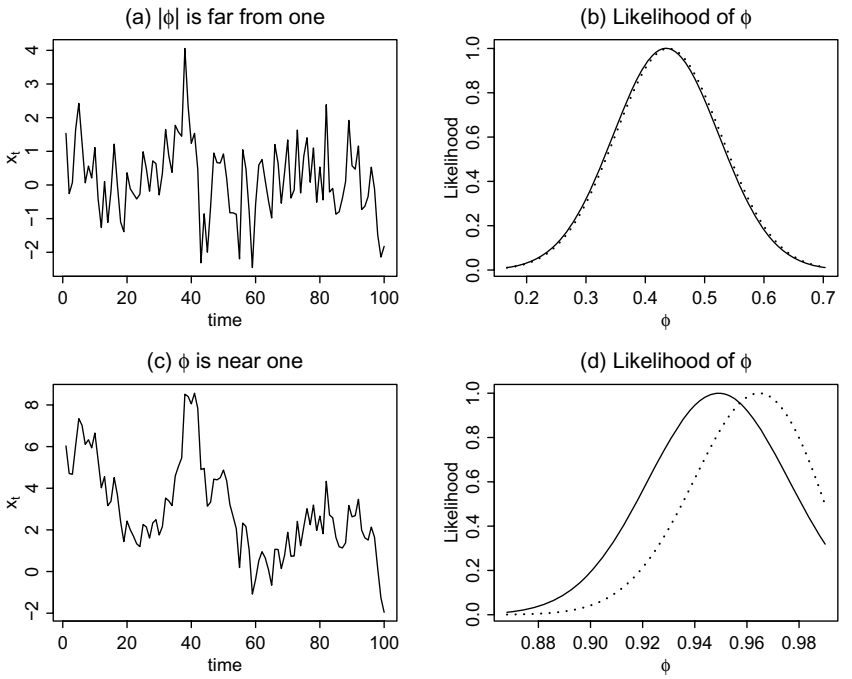

Figure 11.1: *(a) Simulated AR(1) time series, where $\widehat{\phi}$ is far from the boundary of nonstationarity. (b) The conditional (solid line) and full likelihood (dotted) of ϕ based on the time series data in (a). (c) and (d) The same as (a) and (b) for $\widehat{\phi}$ near one.*

A rich class of parametric time series models commonly used in practice is the autoregressive-moving average (ARMA) models (Box *et al.* 1994). A time series x_t that follows an ARMA(p, q) model can be represented as

$$x_t = \theta_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q},$$

where the e_t's are an iid series. Modelling time series data is more difficult than standard regression analysis, since we are not guided by any meaningful relationship between variables. There are a number of descriptive tools to help us, such as the autocorrelation and partial autocorrelation functions. Given a particular model choice, the derivation of a full Gaussian likelihood is tedious, but there are some fast algorithms based on the so-called state-space methodology (Mélard 1984).

11.2 Markov chains

A time series x_t is a first-order Markov chain if

$$p_\theta(x_t|x_1, \ldots, x_{t-1}) = p_\theta(x_t|x_{t-1}),$$

i.e. x_t depends on the past values only through x_{t-1}, or, conditional on x_{t-1}, x_t is independent of the past values. An AR(1) model is a Markov model of order one.

The simplest but still useful model is a two-state Markov chain, where the data are a dependent series of zeros and ones; for example, $x_t = 1$ if it is raining and zero otherwise. This chain is characterized by a matrix of transition probabilities

		x_t	
		0	1
x_{t-1}	0	θ_{00}	θ_{01}
	1	θ_{10}	θ_{11}

where $\theta_{ij} = P(X_t = j|X_{t-1} = i)$, for i and j equal to 0 or 1. The parameters satisfy the constraints $\theta_{00} + \theta_{01} = 1$ and $\theta_{10} + \theta_{11} = 1$, so there are two free parameters.

On observing time series data $x_1, \ldots, x_n$, the likelihood of the parameter $\theta = (\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})$ is

$$\begin{aligned} L(\theta) &= p_\theta(x_1) \prod_{t=2}^{n} p(x_t|x_{t-1}) \\ &= p_\theta(x_1) \prod_{t=2}^{n} \theta_{x_{t-1}0}^{1-x_t} \theta_{x_{t-1}1}^{x_t} \\ &= p_\theta(x_1) \prod_{ij} \theta_{ij}^{n_{ij}} \\ &\equiv L_1(\theta) L_2(\theta) \\ &\equiv L_1(\theta) L_{20}(\theta_{01}) L_{21}(\theta_{11}) \end{aligned}$$

where n_{ij} is the number of transitions from state i to state j. For example, if we observe a series

$$0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0$$

then $n_{00} = 2$, $n_{01} = 1$, $n_{10} = 1$ and $n_{11} = 1$. Again, here it is simpler to consider the conditional likelihood given x_1. The first term $L_1(\theta) = p_\theta(x_1)$ can be derived from the stationary distribution of the Markov chain, which requires a specialized theory (Feller 1968, Chapter XV).

The likelihood term $L_2(\theta)$ in effect treats all the pairs of the form (x_{t-1}, x_t) as if they are independent. All pairs with the same x_{t-1} are independent Bernoulli trials with success probability $\theta_{x_{t-1}1}$. Conditional on x_1, the free parameters θ_{01} and θ_{11} are orthogonal parameters, allowing separate likelihood analyses via

$$L_{20}(\theta_{01}) = (1 - \theta_{01})^{n_{00}} \theta_{01}^{n_{01}}$$

and

$$L_{21}(\theta_{11}) = (1-\theta_{11})^{n_{10}}\theta_{11}^{n_{11}}.$$

We would use this, for example, in (logistic) regression modelling of a Markov chain. In its simplest structure the conditional MLEs of the parameters are

$$\widehat{\theta}_{ij} = \frac{n_{ij}}{n_{i0}+n_{i1}}.$$

Example 11.2: Figure 11.2(a) shows the plot of asthma attacks suffered by a child during 110 winter-days. The data (read by row) are the following:

```
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0
0 0 0 0 1 1 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0
0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Assuming a first-order Markov model the observed transition matrix is

		x_t		
		0	1	Total
	0	82	8	90
x_{t-1}				
	1	8	11	19

For example, when the child is healthy today the estimated probability of an attack tomorrow is $\widehat{\theta}_{01} = 8/90 = 0.09$ (se $= 0.03$); if there is an attack today, the probability of another attack tomorrow is $\widehat{\theta}_{11} = 11/19 = 0.58$ (se $= 0.11$).

We can check the adequacy of the first-order model by extending the model to a second or higher order, and comparing the likelihoods of the different models; see Section 9.11. Derivation of the likelihood based on the higher-order models is left as an exercise. □

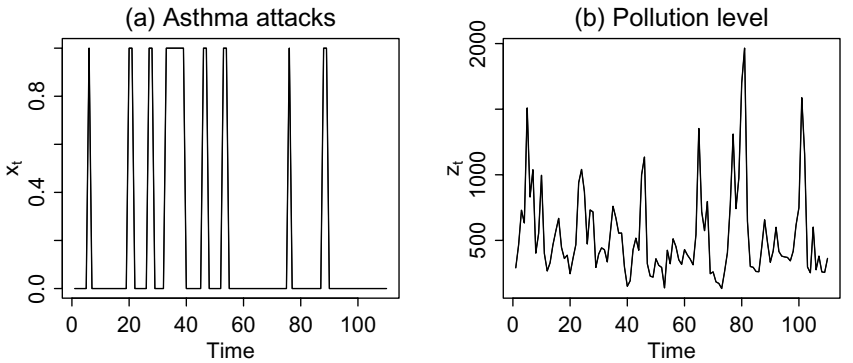

Figure 11.2: *A time series of asthma attacks and a related pollution series.*

Regression analysis

Figure 11.2(b) shows a time series of smoke concentration z_t, measured over the same 110-day period as the asthma series. The values of time series z_t are (read by row)

291	466	730	633	1509	831	1038	403	553	996	400	267	329
466	570	668	446	362	387	246	361	467	940	1041	871	473
732	717	294	396	443	429	336	544	760	672	555	556	298
150	192	428	517	425	1000	1135	322	228	220	360	310	294
138	425	322	512	453	352	317	430	389	357	314	544	1353
720	574	796	246	260	184	172	133	261	406	770	1310	742
976	1701	1965	646	301	295	263	261	450	657	486	333	419
600	415	380	374	370	344	418	617	749	1587	1157	297	253
601	276	380	260	256	363							

The association between asthma attacks and pollution level can be investigated by first-order Markov modelling. For example, we can model

$$\text{logit } \theta_{01t} = \beta_0 + \beta_1 \text{ pollution}_t,$$

so the pollution level modifies the transition probabilities. This model may be analysed using standard logistic regression analysis where the outcome data are pairs of (0,0)'s as failures and (0,1)'s as successes. A simpler model would have been

$$\text{logit } P(x_t = 1) = \beta_0 + \beta_1 \text{ pollution}_t,$$

but in this case the dependence in the asthma series has not been modelled. The actual estimation of the model is left as an exercise.

11.3 Replicated Markov chains

Classical applications of time series, for example in engineering, business or economics, usually involve an analysis of one long time series. New applications in biostatistics have brought short time series data measured on many individuals; they are called repeated measures or longitudinal data. The questions typically focus on comparison between groups of individuals; the dependence structure in the series is a nuisance rather than a feature of interest, and modelling is necessary to attain efficiency or correct inference.

Example 11.3: In a clinical trial of ulcer treatment, 59 patients were randomized to control or treatment groups. These patients were evaluated at baseline (week 0) and at weeks 2, 4, 6 and 8, with symptom severity coded on a six-point scale, where higher is worse. Table 11.1 shows the outcome data. Is there any significant benefit of the treatment? □

In real studies there are potential complications such as

- some follow-up data may be missing,
- length of follow-up differs between subjects, and
- there could be other covariates of interest such as age, sex, etc.

The proper likelihood treatment of these problems is left as Exercises 11.12, 11.13 and 11.15.

The simplest analysis can be obtained by assessing the improvement at the last visit (w8) relative to the baseline (w0):

	Control group						Treated group					
No.	w0	w2	w4	w6	w8	No.	w0	w2	w4	w6	w8
1	3	4	4	4	4	1	4	4	3	3	3
2	5	5	5	5	5	2	4	4	3	3	3
3	2	2	2	1	1	3	6	5	5	4	3
4	6	6	5	5	5	4	3	3	3	3	2
5	4	3	3	3	3	5	6	5	5	5	5
6	5	5	5	5	5	6	3	3	3	2	2
7	3	3	3	3	3	7	5	5	6	6	6
8	3	3	3	2	2	8	3	2	2	2	2
9	4	4	4	3	2	9	4	3	2	2	2
10	5	4	4	3	4	10	5	5	5	5	5
11	4	4	4	4	4	11	3	2	2	2	2
12	4	4	4	4	4	12	3	2	2	1	1
13	5	5	5	5	5	13	6	6	6	6	5
14	5	4	4	4	4	14	2	2	1	1	2
15	5	5	5	4	3	15	4	4	4	3	2
16	3	2	2	2	2	16	3	3	3	3	3
17	5	5	6	6	6	17	2	1	1	1	2
18	6	6	6	6	6	18	4	4	4	4	4
19	3	2	2	2	2	19	4	4	4	4	3
20	5	5	5	5	5	20	4	4	3	3	3
21	4	4	4	4	3	21	4	4	4	3	3
22	2	2	2	2	2	22	4	3	2	2	2
23	4	3	3	3	3	23	3	2	2	2	2
24	3	2	2	1	1	24	3	3	3	3	3
25	4	3	3	3	3	25	3	2	2	1	1
26	4	3	3	3	3	26	4	4	3	3	3
27	5	5	5	5	5	27	2	1	1	1	1
28	3	3	3	3	3	28	3	2	3	2	1
29	3	3	3	3	2	29	2	1	1	1	1
30	6	6	6	6	6						

Table 11.1: *Follow-up data on treatment of ulcers.*

Change	Control	Treated	Total
Better	16(53%)	22(76%)	38
No	14	7	21
Total	30	29	59

This table indicates some positive benefit. However, the standard χ^2 statistic for the observed 2×2 table is

$$\chi^2 = \frac{(16 \times 7 - 22 \times 14)^2 59}{30 \times 29 \times 38 \times 21} = 3.29,$$

with 1 degree of freedom; this gives a (one-sided) P-value=0.07, which is not quite significant.

How much has been lost by ignoring most of the data? To include the whole dataset in the analysis we need to consider a more complicated model. Assuming a first-order Markov model

$$P(x_k = j|x_t, t \leq k-1) = P(x_k = j|x_{k-1})$$

would spawn a 6×6 transition matrix with $6 \times (6-1) = 30$ independent parameters *for each group*; using this most general model, it would not be obvious how to compare the treatment versus the placebo groups.

This can be simplified by assuming a patient can only change by one level of severity in a two-week period. The transition probability from state i to state j is given by, for $i = 2, \ldots, 5$,

$$p_{ij} = \begin{cases} p & \text{if } j = i+1 \text{ (worse)} \\ q & \text{if } j = i-1 \text{ (better)} \\ 1-p-q & \text{if } j = i \text{ (same)} \\ 0 & \text{otherwise.} \end{cases}$$

The model is completed by specifying the boundary probabilities

$$p_{1j} = \begin{cases} p & \text{if } j = 2 \\ 1-p & \text{if } j = 1 \\ 0 & \text{otherwise.} \end{cases}$$

$$p_{6j} = \begin{cases} q & \text{if } j = 5 \\ 1-q & \text{if } j = 6 \\ 0 & \text{otherwise.} \end{cases}$$

In view of the model, the data can be conveniently summarized in the following table

Change	Control	Treated	Transition probability
$1 \to 1$	2	11	$1-p$
$1 \to 2$	0	2	p
$6 \to 6$	11	5	$1-q$
$6 \to 5$	1	3	q
other $+1$	3	2	p
other 0	84	63	$1-p-q$
other -1	19	30	q
Total	120	116	

where 'other $+1$' means a change of $+1$ from the states $2, \ldots, 5$. The data satisfy the assumption of no jump of more that one point; extension that allows for a small transition probability of a larger jump is left as an exercise. The likelihood of the parameters p and q for each group can be computed based on the table (Exercise 11.11). The two groups can now be compared, for example, in terms of the parameter q alone or the difference $q - p$.

If we are interested in comparing the rate of improvement we can compare the parameter q; for this we can reduce the data by conditioning on starting at states 2 or worse, so we can ignore the first two lines of the table and combine the rest into a 2×2 table

Change	Control	Treated	Total
Better (-1)	20(17%)	33(32%)	53
No (≥ 0)	98	70	168
Total	118	103	221

which now yields $\chi^2 = 6.87$ with P-value=0.009. The interpretation is different from the first analysis: there is strong evidence that the treatment has a short-term benefit in reducing symptoms.

11.4 Spatial data

Spatial data exhibit dependence like time series data, with the main difference that there is no natural direction in the dependence or time-causality. Spatially dependent data exhibit features such as

- clustering or smooth variation: high values tend to appear near each other; this is indicative of a local positive dependence;
- almost-regular pattern: high or low values tend to appear individually, indicating a negative spatial dependence or a spatial inhibition;
- streaks: positive dependence occurs along particular directions. Long streaks indicate global rather than local dependence.

We should realize, however, that a completely random pattern must show some local clustering. This is because if there is no clustering at all we will end up with a regular pattern. Except in extreme cases, our eyes are not good at judging if a particular clustering is real or spurious.

In many applications it is natural to model the value at a particular point in space in terms of the surrounding values. Modelling of spatial data is greatly simplified if the measurements are done on a regular grid or lattice structure, equivalent to equal-space measurement of time series. However, even for lattice-structured data the analysis is marked by compromises.

The spatial nature of an outcome y ('an outcome' here is a whole image) can be ignored if we model y in terms of spatially varying covariates x. Conditional on x we might model the elements of y as being independent; this assumes any spatial dependence in y is inherited from that in x. For example, the yield y of a plant may be spatially dependent, but the dependence is probably induced by fertility x. With such a model the spatial dependence is assumed to be fully absorbed by x, and it can be left unmodelled. So the technique we discuss in this section is useful if we are interested in the dependence structure itself, or if there is a residual dependence after conditioning on a covariate x.

One-dimensional Ising model

To keep the discussion simple, consider a linear lattice or grid system where we observe an array $y_1, \ldots, y_n$. What is a natural model to describe its probabilistic behaviour? We can still decompose the joint probability as

$$p(y_1, \ldots, y_n) = p(y_1)\, p(y_2|y_1) \ldots p(y_n|y_{n-1}, \ldots),$$

but such a decomposition is no longer natural, since it has an implied left-to-right direction. There are generalizations of the ARMA models to spatial processes. The key idea is that the probabilistic behaviour of the process at a particular location is determined by the nearby values or neighbours.

For example, we might want to specify a first-order model

$$p(y_k|y_1,\ldots,y_{k-1},y_{k+1},\ldots,y_n) = p(y_k|y_{k-1},y_{k+1}).$$

The problem is that, except for the Gaussian case, it is not obvious how to build the likelihood from such a specification. The product of such conditional probabilities for all values $y_1,\ldots,y_n$ is not a true likelihood. In its simplest form, if we observe (y_1,y_2), the product of conditional probabilities is

$$p(y_1|y_2)\ p(y_2|y_1),$$

while the true likelihood is

$$p(y_1)\ p(y_2|y_1).$$

To simplify the problem further, suppose y_i can only take 0–1 values. The famous Ising model in statistical mechanics specifies the joint distribution of $y_1,\ldots,y_n$ as follows. Two locations j and k are called neighbours if they are adjacent, i.e. $|j-k|=1$; for example, the neighbours of location $k=3$ are $j=2$ and $j=4$. Let n_k be the sum of y_j's from neighbours of k. The Ising model specifies that, conditional on the edges y_1 and y_n, the joint probability of $y_2,\ldots,y_{n-1}$ is

$$p(y_2,\ldots,y_{n-1}|y_1,y_n) = \exp\left\{\alpha\sum_{k=2}^{n-1} y_k + \beta\sum_{k=2}^{n-1} y_k n_k - h(\alpha,\beta)\right\},$$

where $h(\alpha,\beta)$ is a normalizing constant. The parameter β measures the local interaction between neighbours. If $\beta=0$ then $y_2,\ldots,y_{n-1}$ are iid Bernoulli with parameter e^α. Positive β implies positive dependence, so that values of y_i's tend to cluster; this will be obvious from the conditional probability below.

It is not easy to compute the true likelihood from the joint probability, since the normalizing factor $h(\alpha,\beta)$ is only defined implicitly. However, we can show (Exercise 11.16) that the conditional probability of y_k given all of the other values satisfies the logistic model

$$P(y_k=1|y_1,\ldots,y_{k-1},y_{k+1},\ldots,y_n) = \frac{\exp(\alpha+\beta n_k)}{1+\exp(\alpha+\beta n_k)},$$

so it is totally determined by the local neighbours. It is tempting to define a likelihood simply as the product of the conditional probabilities. That

is in fact the common approach for estimating the Ising model, namely we use

$$L(\alpha, \beta) = \prod_{k=2}^{n-1} p(y_k | y_1, \ldots, y_{k-1}, y_{k+1}, \ldots, y_n),$$

known as *pseudo-likelihood* (Besag 1974, 1975); it is a different likelihood from all we have defined previously, and inference from it should be viewed with care. Besag and Clifford (1989) developed an appropriate Monte Carlo test for the parameters. The practical advantage of the pseudo-likelihood is obvious: parameter estimation can be performed using standard logistic regression packages.

Example 11.4: As a simple illustration of the Ising model, consider analysing the dependence structure in the following spatial data:

```
0 0 0 1 1 1 0 0 1 0 0 0 1 1 0 1 1 1 0 1 0 0 1 1 0 0 1 1 1 1
```

The data for the logistic regression can be set up first in terms of data pairs (y_k, n_k) for $k = 2, \ldots, 29$. There are 30 total points in the data, and we have 28 points with two neighbours. Thus $(y_2 = 0, n_2 = 0)$, $(y_3 = 0, n_3 = 1)$, ... , $(y_{29} = 1, n_{29} = 2)$ and we can group the data into

n_k	$y_k = 0$	$y_k = 1$	Total
0	2	2	4
1	9	9	18
2	2	4	6

Estimating the logistic regression model with n_k as the predictor, the maximum pseudo-likelihood estimates and the standard errors of the parameters (α, β) are given in the following:

Effect	Parameter	Estimate	se
Constant	α	-0.27	0.79
n_k	β	0.38	0.65

The analysis shows that there is no evidence of local dependence in the series. □

Two-dimensional Ising model

Extension to true spatial data with a two-dimensional grid structure (i, j) is as follows. Suppose we observe a two-dimensional array y_{ij} of 0–1 values. The most important step is the definition of neighbourhood structure. For example, we may define the locations (i, j) and (k, l) as 'primary' neighbours if $|i - k| = 1$ and $j = l$, or $|j - l| = 1$ and $i = k$; the 'diagonal' or 'secondary' neighbours are those with $|i - k| = 1$ and $|j - l| = 1$. (Draw these neighbours to get a clear understanding.) Depending on the applications we might treat these different neighbours differently. Let n_{ij} be the sum of y_{ij}'s from the primary neighbours of location (i, j), and m_{ij} be the sum from the diagonal neighbours. Then a general Ising model of the joint distribution of y_{ij} (conditional on the edges) implies a logistic model

$$P(y_{ij} = 1 | \text{other } y_{ij}\text{'s}) = \frac{\exp(\alpha + \beta n_{ij} + \gamma m_{ij})}{1 + \exp(\alpha + \beta n_{ij} + \gamma m_{ij})}.$$

Estimation of the parameters based on the pseudo-likelihood can be done using standard logistic regression packages.

Gaussian models

To describe the Gaussian spatial models it is convenient to first vectorize the data into $y = (y_1, \ldots, y_n)$. The neighbourhood structure can be preserved in an $n \times n$ matrix W with elements $w_{ii} = 0$, and $w_{ij} = 1$ if i and j are neighbours, and zero otherwise. The domain of y can be irregular.

The Gaussian conditional autoregressive (CAR) model specifies that the conditional distribution of y_i given the other values is normal with mean and variance

$$\begin{aligned} E(y_i|y_j,\ j \neq i) &= \mu_i + \sum_j c_{ij}(y_j - \mu_j) \\ \mathrm{var}(y_i|y_j,\ j \neq i) &= \sigma_i^2. \end{aligned}$$

Let $C \equiv [c_{ij}]$ with $c_{ii} = 0$ and $D \equiv \mathrm{diag}[\sigma^2]$. The likelihood can be derived from the unconditional distribution of y, which is $N(\mu, \Sigma)$ with

$$\begin{aligned} \mu &= (\mu_1, \ldots, \mu_n) \\ \Sigma &= (I_n - C)^{-1} D, \end{aligned}$$

provided Σ is a symmetric positive definite matrix; I_n is an $n \times n$ identity matrix. The symmetry is satisfied if $c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2$. For modelling purposes, some simplifying assumptions are needed, such as equal variance and a first-order model, which specifies $c_{ij} = \phi$ if i and j are neighbours and zero otherwise. Some model fitting examples can be found in Ripley (1988).

An important application of spatial data analysis is in smoothing sparse disease maps, where the raw data exhibit too much noise for sensible reading. The input for these maps are count data, which are not covered by the Ising model. However, the pseudo-likelihood approach can be extended for such data (Exercise 11.20). The problem can also be approached using mixed model in Section 18.10. For this purpose it is useful to have a nonstationary process for the underlying smooth function. Besag *et al.* (1991) suggest a nonstationary CAR model defined by the joint distribution of a set of differences. The log-density is

$$\begin{aligned} \log p(y) &= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{ij} w_{ij}(y_i - y_j)^2 \\ &= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} y' R^{-1} y \end{aligned}$$

where $N = \sum w_{ij}$ is the number of neighbour pairs and $\sigma^{-2}R^{-1}$ is the inverse covariance matrix of y. It can be shown that $R^{-1} = [r^{ij}]$, where $r^{ii} = \sum_j w_{ij}$ and $r^{ij} = -w_{ij}$ if $i \neq j$.

11.5 Censored/survival data

In most clinical trials or reliability studies it is not possible to wait for all experimental units to reach their 'end-point'. An end-point in a survival study is the time of death, or appearance of a certain condition; in general it is the event that marks the end of follow-up for a subject. Subjects are said to be censored if they have not reached the end-point when the study is stopped, or are lost during follow-up. In more general settings, censored data are obtained whenever the measurement is not precise; for example, a binomial experiment where (i) the number of successes is known only to be less than a number, or (ii) the data have been grouped.

Example 11.5: Two groups of rats were exposed to carcinogen DBMA, and the number of days to death due to cancer was recorded (Kalbfleisch and Prentice 1980).

Group 1 :	143, 164, 188, 188, 190, 192, 206, 209, 213,
	216, 220, 227, 230, 234, 246, 265, 304, 216+, 244+
Group 2 :	142, 156, 163, 198, 205, 232, 232, 233, 233,
	233, 233, 239, 240, 261, 280, 280, 296, 296,
	323, 204+, 344+

Values marked with '+' are censored. Is there a significant difference between the two groups? □

In this example four rats were 'censored' at times 216, 244, 204 and 344; those rats were known not to have died of cancer by those times. Possible reasons for censoring are

- deaths due to other causes;
- being alive when the study ends.

The group comparison problem is simple, although the censored data presents a problem. How do we treat these cases? We can

- ignore the censoring information, i.e. we treat all the data as if they are genuine deaths;
- drop the censored cases, so we are dealing with genuine deaths;
- model the censored data properly.

The first two methods can be very biased and misleading if the censoring patterns in the groups differ. The second method is inefficient even if the censoring patterns in the two groups are similar. With a correct model, the last method is potentially the best as it would take into account whatever information is available in the censored data.

The censored data can be written as pairs $(y_1, \delta_1), \ldots, (y_n, \delta_n)$, where δ_i is the *last-known status* or *event indicator*: $\delta_i = 1$ if y_i is a true event time, and zero otherwise. If t_i is the true lifetime of subject i, then $\delta_i = 0$ iff $t_i > y_i$. We would be concerned with modeling the true lifetime t_i rather

than the observed y_i, since censoring is usually a nuisance process that does not have any substantive meaning, for example it can be determined by the study design.

Suppose $t_1, \ldots, t_n$ are an iid sample from $p_\theta(t)$. The likelihood contribution of the observation (y_i, δ_i) is

$$L_i(\theta) = P_\theta(T_i > y_i) \quad \text{if } \delta_i = 0$$

or

$$L_i(\theta) = p_\theta(t_i) \quad \text{if } \delta_i = 1.$$

The probability $P_\theta(T_i > y_i)$ is called the survival function of T_i. The overall likelihood is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} L_i(\theta) \\ &= \prod_{i=1}^{n} \{p_\theta(y_i)\}^{\delta_i} \{P_\theta(T_i > y_i)\}^{1-\delta_i}. \end{aligned}$$

As an example, consider a simple exponential model, which is commonly used in survival and reliability studies, defined by

$$\begin{aligned} p_\theta(t) &= \frac{1}{\theta} e^{-t/\theta} \\ P_\theta(T > t) &= e^{-t/\theta}. \end{aligned}$$

In this case

$$L(\theta) = \left(\frac{1}{\theta}\right)^{\sum \delta_i} \exp(-\sum y_i/\theta).$$

Upon taking the derivative of the log-likelihood we get the score function

$$S(\theta) = -\frac{\sum \delta_i}{\theta} + \frac{\sum_i y_i}{\theta^2}$$

and setting it to zero, we get

$$\widehat{\theta} = \frac{\sum y_i}{\sum \delta_i}.$$

Note that $\sum y_i$ is the total observation times including both the censored and uncensored cases, while $\sum \delta_i$ is the number of events. The inverse $1/\widehat{\theta}$ is an estimate of the event rate. This is a commonly used formula in epidemiology, and known as the person-year method. For example, the Steering Committee of the Physicians' Health Study Research Group (1989) reported the rate of heart attacks as 254.8 per 100,000 person-years in the aspirin group, compared with 439.7 in the placebo group; see Section 1.1.

With some algebra the observed Fisher information of θ is

$$I(\widehat{\theta}) = \frac{\sum \delta_i}{\widehat{\theta}^2},$$

so the standard error of $\widehat{\theta}$ is

$$\text{se}(\widehat{\theta}) = \frac{\widehat{\theta}}{(\sum \delta_i)^{1/2}}.$$

Example 11.5: continued. Assume an exponential model for excess lifetime over 100 days (in principle we can make this cutoff value another unknown parameter), so from Group 1 we get $n = 19$, $\sum y_i = 2195$, $\sum \delta_i = 17$ and

$$L(\theta_1) = \left(\frac{1}{\theta_1}\right)^{17} \exp(-2195/\theta_1),$$

which yields $\widehat{\theta}_1 = 2195/17 = 129.1$ (se = 31.3). The plot of the likelihood function is given in Figure 11.3(a). Similarly, from Group 2 we have $n = 21$, $\sum y_i = 2923$, $\sum \delta_i = 19$ and

$$L(\theta_2) = \left(\frac{1}{\theta_2}\right)^{19} \exp(-2923/\theta_2),$$

which yields $\widehat{\theta}_2 = 2923/19 = 153.8$ (se = 35.3). There is some indication that rats from Group 2 live longer than those from Group 1. The standard error of $\widehat{\theta}_1 - \widehat{\theta}_2$ is

$$\text{se}(\widehat{\theta}_1 - \widehat{\theta}_2) = \{\text{se}(\widehat{\theta}_1)^2 + \text{se}(\widehat{\theta}_2)^2\}^{1/2} = 47.2.$$

The Wald statistic for comparing the mean of Group 1 with the mean of Group 2 is $z = (129.1 - 153.8)/47.2 = -0.53$.

The following table compares the results of the three methods mentioned earlier. The normal model will be described later. □

Method	Sample mean Group 1	Sample mean Group 2	t- or Wald statistic
Ignore	115.5	139.2	−1.65
Drop cases	113.8	135.5	−1.49
Exp. model	129.1	153.8	−0.53
Normal model	119.1	142.6	−1.56

The result based on the exponential model is the least significant, which of course does not mean it is the best result for this dataset. Such a result is strongly dependent on the chosen model. Is the exponential model a good fit for the data? Figure 11.3(b) shows the QQ-plot of the uncensored observations versus the theoretical exponential distribution, indicating that the exponential model is doubtful. (A proper QQ-plot that takes the censored data into account requires an estimate of the survival function, which is given by the so-called Kaplan–Meier estimate below.)

The mean–variance relationship implied by the exponential model also does not hold: the variance is much smaller than the square of the mean.

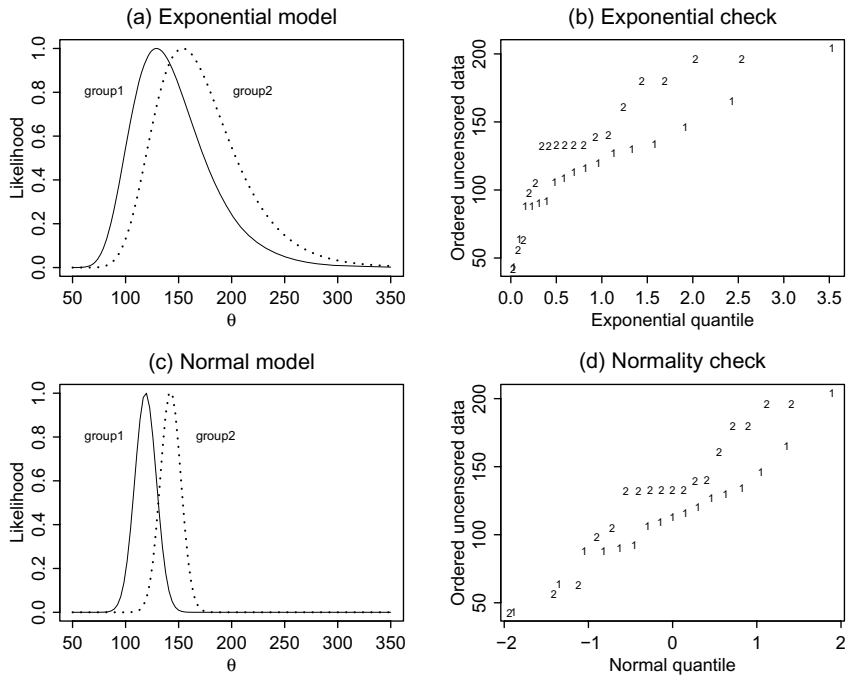

Figure 11.3: *Analysis of the rat data: the first row assumes exponential model and the second row assumes normal model.*

For the uncensored data there is an underdispersion factor of around 0.11. This means that the exponential-based likelihood is too wide. A proper model that takes the dispersion factor into account is given by the general survival regression model in the next section.

Normal model

Suppose t_i's are an iid sample from $N(\mu, \sigma^2)$. The likelihood contribution of an uncensored observation y_i is

$$p_\theta(y_i) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\}$$

and the contribution of a censored observation is

$$P_\theta(y_i) = 1 - \Phi\left(\frac{y_i - \mu}{\sigma}\right)$$

where $\Phi(\cdot)$ is the standard normal distribution function. The functions are analytically more difficult than those in the exponential model, and there are no explicit formulae.

For the rat data, the likelihood of μ based on the normal model is shown in Figure 11.3(c). To simplify the computations it is assumed that the two

groups have equal variance, and the variance is known at the estimated value from the uncensored data. The QQ-plot suggests the normal model seems to be a better fit. A formal comparison can be done using the AIC.

Kaplan–Meier estimate of the survival function

The most commonly used display of survival data is the Kaplan–Meier estimate of the survival function. It is particularly useful for a graphical comparison of several survival functions. Assume for the moment that there is *no censored data* and no tie, and let $t_1 < t_2 < \cdots < t_n$ be the ordered failure times. The empirical distribution function (EDF)

$$F_n(t) = \frac{\text{number of failure times } t_i\text{'s} \le t}{n}$$

is the cumulative proportion of mortality by time t. In Section 15.1 this is shown to the nonparametric MLE of the underlying distribution, therefore the MLE of the survival function is

$$\widehat{S}(t) = 1 - F_n(t).$$

The function $\widehat{S}(t)$ is a step function with a drop of $1/n$ at each failure time, starting at $\widehat{S}(0) \equiv 1$.

We can reexpress $\widehat{S}(t)$ in a form that is extendable to censored data. Let n_i be the number 'at risk' (yet to fail) just prior to failure time t_i. If there is no censoring $n_1 = n$, $n_2 = n - 1$, and $n_i = n - i + 1$. It is easily seen that

$$\widehat{S}(t) = \prod_{t_i \le t} \frac{n_i - 1}{n_i}.$$

For example, for $t_3 \le t < t_4$,

$$\begin{aligned} \widehat{S}(t) &= \frac{n_1 - 1}{n_1} \times \frac{n_2 - 1}{n_2} \times \frac{n_3 - 1}{n_3} \\ &= \frac{n-1}{n} \times \frac{n-2}{n-1} \times \frac{n-3}{n-2} \\ &= \frac{n-3}{n} = 1 - \frac{3}{n} = 1 - F_n(t). \end{aligned}$$

If there are ties, we only need a simple modification. Let $t_1, \ldots, t_k$ be the observed failure times, and $d_1, \ldots, d_k$ be the corresponding number of failures. Then

$$\widehat{S}(t) = \prod_{t_i \le t} \frac{n_i - d_i}{n_i}.$$

Exactly the same formula applies for censored data, and it is called Kaplan–Meier's product limit estimate of the survival function. It can be shown that the Kaplan–Meier estimate is the MLE of the survival distribution; see Kalbfleisch and Prentice (1980, page 11), and also Section 11.10.

Information from the censored cases is used in computing the number at risk n_i's. With uncensored data

$$n_i = n_{i-1} - d_{i-1},$$

i.e. the number at risk prior to time t_i is the number at risk prior to the previous failure time t_{i-1} minus the number that fails at time t_{i-1}. With censored data,

$$n_i = n_{i-1} - d_{i-1} - c_{i-1},$$

where c_{i-1} is the number censored between failure times t_{i-1} and t_i. (If there is a tie between failure and censoring times, it is usually assumed that censoring occurs after failure.)

These ideas can be grasped easily using a toy dataset $(y_1, \ldots, y_6) = (3, 4, 6+, 8, 8, 10)$. Here, we have

i	t_i	n_i	d_i	c_i	$\widehat{S}(t)$
0	$t_0 \equiv 0$	6	0	0	1
1	3	6	1	0	5/6
2	4	5	1	1	$5/6 \times 4/5 = 4/6$
3	8	3	2	0	$5/6 \times 4/5 \times 1/3 = 2/9$
4	10	1	1	0	0

For larger datasets the computation is tedious, but there are many available softwares. Figure 11.4 shows the Kaplan–Meier estimates of the survival functions of the rat groups in Example 11.5. The plot indicates a survival advantage of group 2 over group 1.

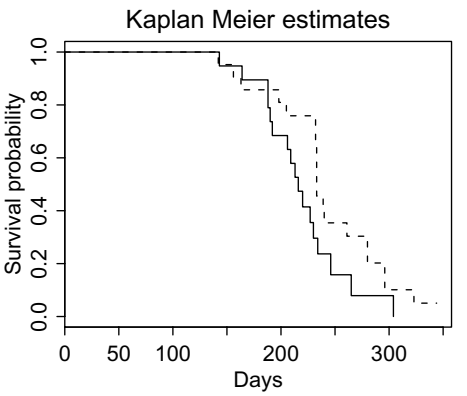

Figure 11.4: *Kaplan–Meier estimates of the survival function of group 1 (solid) and group 2 (dashed) of the rat data in Example 11.5.*

11.6 Survival regression models

In the same spirit as the models we develop in Chapter 6, the previous example can be extended to a general regression framework. Suppose we

want to analyse the effect of some characteristics x on survival; for example, x is a group indicator ($x = 0$ or 1 in the rat data example). Using an exponential model

$$t_i \sim \text{Exponential}(\theta_i),$$

where the mean θ_i is a function of the covariates x_i, connected via a link function $h(\cdot)$, such that

$$h(\theta_i) = x_i'\beta.$$

We might consider the identity link

$$\theta_i = x_i'\beta$$

or the log-link

$$\log \theta_i = x_i'\beta.$$

The log-link function is more commonly used since $\theta_i > 0$.

Based on the observed data $(y_1, \delta_1, x_1), \ldots, (y_n, \delta_n, x_n)$ the likelihood function of the regression parameter β can written immediately as

$$L(\beta) = \prod_{i=1}^{n} \{p_{\theta_i}(y_i)\}^{\delta_i} \{P_{\theta_i}(T_i > y_i)\}^{1-\delta_i},$$

where θ_i is a function of β, and

$$p_{\theta_i}(y_i) = \theta_i^{-1} e^{-y_i/\theta_i},$$

and

$$P_{\theta_i}(y_i) = e^{-y_i/\theta_i}.$$

As we have seen before the exponential model specifies a rigid relationship between the mean and the variance. An extension that allows a flexible relationship is essential. To motivate a natural development, note that if T is exponential with mean θ, then

$$\log T = \beta_0 + W$$

where $\beta_0 = \log \theta$ and W has the standard extreme-value distribution with density

$$p(w) = e^w \exp(-e^w)$$

and survival function

$$P(W > w) = \exp(-e^w).$$

A more flexible regression model between a covariate vector x_i and outcome T_i can be developed by assuming

$$\log T_i = \log \theta_i + \sigma W_i$$

where W_i's are iid with standard extreme-value distribution, and σ is a scale parameter. This extension is equivalent to using the Weibull model.

As usual, the parameter θ_i is related to the covariate vector x_i via a link function. For example, using the log-link

$$\log \theta_i = x_i' \beta.$$

This is the so-called *accelerated failure time model*: the parameter e^β is interpreted as the multiplicative effect of a unit change in x on average lifetime.

The likelihood can be computed from the Weibull survival function

$$\begin{aligned} P_{\theta_i}(T_i > y_i) &= P\{\log(T_i/\theta_i) > \log(y_i/\theta_i)\} \\ &= P\{W_i > \log(y_i/\theta_i)^{1/\sigma}\} \\ &= \exp\{-(y_i/\theta_i)^{1/\sigma}\} \end{aligned}$$

and density function

$$p_{\theta_i}(y_i) = \sigma^{-1} y_i^{1/\sigma - 1} \theta_i^{-1/\sigma} \exp\{-(y_i/\theta_i)^{1/\sigma}\}.$$

Example 11.6: For the rat data in Example 11.5, suppose we model the mean θ_i as

$$\log \theta_i = \beta_0 + \beta_1 x_i$$

where $x_i = 0$ or 1, for Group 1 or 2, respectively. The following table summarizes the analysis using the exponential model ($\sigma = 1$) and the general model by letting σ be free.

		Exponential		General	
Effect	Parameter	Estimate	se	Estimate	se
Constant	β_0	4.861	0.243	4.873	0.079
Group	β_1	0.175	0.334	0.213	0.108
Scale	σ	1	–	0.32	–

For the exponential model the estimated mean ratio of the two groups is $e^{\widehat{\beta_1}} = 1.19 = 153.8/129.1$ as computed before. The estimated scale $\widehat{\sigma} = 0.32$ gives a dispersion factor $\widehat{\sigma}^2 = 0.10$, as expected from the previous discussion of this example. Since the exponential model ($\sigma = 1$) is a special case of the general model, we can test its adequacy by the likelihood ratio test: find the maximized likelihood under each model, and compute the likelihood ratio statistic $W = 44.8$, which is convincing evidence against the exponential model.

Under the general model, we obtain $z = 0.213/0.108 = 1.98$ for the observed group difference, which is borderline significant. Checking the appropriateness of the quadratic approximation for the profile likelihood of β_1 is left as an exercise. □

11.7 Hazard regression and Cox partial likelihood

The *hazard function* is indispensable in the analysis of censored data. It is defined as

$$\lambda(t) = p(t)/P(T > t),$$

and interpreted as the rate of dying at time t among the survivors. For example, if T follows an exponential distribution with mean θ, the hazard function of T is

$$\lambda(t) = 1/\theta.$$

This inverse relationship is sensible, since a short lifetime implies a large hazard. Because the hazard function is constant the exponential model may not be appropriate for living organisms, where the hazard is typically higher at both the beginning and the end of life. Models can be naturally put in hazard form, and the likelihood function can be computed based on the following relationships:

$$\lambda(t) = -\frac{d \log P(T > t)}{dt} \tag{11.1}$$

$$\log P(T > t) = -\int_0^t \lambda(u) du \tag{11.2}$$

$$\log p(t) = \log \lambda(t) - \int_0^t \lambda(u) du. \tag{11.3}$$

Given censored data $(y_1, \delta_1, x_1), \ldots, (y_n, \delta_n, x_n)$, where δ_i is the event indicator, and the underlying t_i has density $p_{\theta_i}(t_i)$, the log-likelihood function contribution of (y_i, δ_i, x_i) is

$$\begin{aligned} \log L_i &= \delta_i \log p_{\theta_i}(y_i) + (1 - \delta_i) \log P_{\theta_i}(y_i) \\ &= \delta_i \log \lambda_i(y_i) - \int_0^{y_i} \lambda_i(u) du \end{aligned} \tag{11.4}$$

where the parameter θ_i is absorbed by the hazard function. Only uncensored observations contribute to the first term.

The most commonly used model in survival analysis is the proportional hazard model of the form

$$\lambda_i(t) = \lambda_0(t) e^{x_i' \beta}. \tag{11.5}$$

In survival regression or comparison studies the baseline hazard $\lambda_0(t)$ is a nuisance parameter; it must be specified for a full likelihood analysis of the problem. Here is a remarkable property of the model that avoids the need to specify $\lambda_0(t)$: if lifetimes T_1 and T_2 have proportional hazards

$$\lambda_i(t) = \lambda_0(t) \eta_i$$

for $i = 1, 2$, respectively, then

$$P(T_1 < T_2) = \eta_1/(\eta_1 + \eta_2)$$

regardless of the shape of the baseline hazard function (Exercise 11.28). We can interpret the result this way: if $\eta_1 > \eta_2$ then it is more likely that subject 1 will die first.

Such a probability can be used in a likelihood function based on knowing the event $[T_1 < T_2]$ alone, namely only the ranking information is used, not

the actual values. Such a likelihood is then a marginal likelihood. If x_i is the covariate vector associated with T_i and we model

$$\eta_i = e^{x_i'\beta},$$

then the likelihood of β based on observing the event $[T_1 < T_2]$ is

$$L(\beta) = \frac{e^{x_1'\beta}}{e^{x_1'\beta} + e^{x_2'\beta}},$$

which is free of any nuisance parameter.

In general, given a sample $T_1, \ldots, T_n$ from a proportional hazard model

$$\lambda_i(t) = \lambda_0(t) e^{x_i'\beta},$$

the probability of a particular configuration $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$ is

$$P(T_{i_1} < T_{i_2} < \cdots < T_{i_n}) = \prod_{j=1}^{n} e^{x_{i_j}'\beta} / \Big(\sum_{k \in R_j} e^{x_k'\beta} \Big), \tag{11.6}$$

where R_j is the list of subjects where $T \geq T_{i_j}$, which is called the 'risk set' at time T_{i_j}. It is easier to see this in an example for $n = 3$: for $(i_1, i_2, i_3) = (2, 3, 1)$ we have

$$P(T_2 < T_3 < T_1) = \frac{e^{x_2'\beta}}{e^{x_1'\beta} + e^{x_2'\beta} + e^{x_3'\beta}} \times \frac{e^{x_3'\beta}}{e^{x_1'\beta} + e^{x_3'\beta}} \times \frac{e^{x_1'\beta}}{e^{x_1'\beta}}.$$

The likelihood of the regression parameter β computed from this formula is called the Cox partial likelihood (Cox 1972, 1975), the main tool of survival analysis. With this likelihood we are only using the ranking observed in the data. In the example with $n = 3$ above we only use the information that subject 2 died before subject 3, and subject 3 died before subject 1. If the baseline hazard can be any function, it seems reasonable that there is very little extra information beyond the ranking information in the data. In fact, it has been shown that, for a wide range of underlying hazard functions, the Cox partial likelihood loses little or no information (Efron 1977).

Extension to the censored data case is intuitively obvious by thinking of the risk set at each time of death: a censored value only contributes to the risk sets of the *prior* uncensored values, but it cannot be compared with later values. For example, if we have three data values 1, 5+ and 7, where 5+ is censored, we only know that 1 is less than 5+ and 7, but we cannot distinguish between 5+ and 7. In view of this, the same likelihood formula (11.6) holds.

Example 11.7: The following toy dataset will be used to illustrate the construction of the Cox likelihood for a two-group comparison.

i	x_i	y_i	δ_i
1	0	10	0
2	0	5	1
3	0	13	1
4	1	7	1
5	1	21	1
6	1	17	0
7	1	19	1

Assume a proportional hazard model

$$\lambda_i(t) = \lambda_0(t)e^{x_i\beta}.$$

We first sort the data according to the observed y_i.

i	x_i	y_i	δ_i
2	$x_2 = 0$	5	1
4	$x_4 = 1$	7	1
1	$x_1 = 0$	10	0
3	$x_3 = 0$	13	1
6	$x_6 = 1$	17	0
7	$x_7 = 1$	19	1
5	$x_5 = 1$	21	1

The Cox partial likelihood of β is

$$\begin{aligned} L(\beta) &= \frac{e^{x_2\beta}}{\sum_{i=1}^{7} e^{x_i\beta}} \\ &\times \frac{e^{x_4\beta}}{e^{x_4\beta} + e^{x_1\beta} + e^{x_3\beta} + e^{x_6\beta} + e^{x_7\beta} + e^{x_5\beta}} \\ &\times \frac{e^{x_3\beta}}{e^{x_3\beta} + e^{x_6\beta} + e^{x_7\beta} + e^{x_5\beta}} \\ &\times \frac{e^{x_7\beta}}{e^{x_7\beta} + e^{x_5\beta}}. \end{aligned}$$

Note that only uncensored cases can appear in the numerator of the likelihood. The Cox partial likelihood is generally too complicated for direct analytical work; in practice most likelihood quantities such as MLEs and their standard errors are computed numerically. In this example, since x_i is either zero or one, it is possible to simplify the likelihood further. □

Example 11.8: Suppose we assume the proportional hazard model for the rat data in Example 11.5:

$$\lambda(t) = \lambda_0(t)e^{\beta_1 x_i},$$

where $x_i = 0$ or 1 for Group 1 or 2, respectively. We do not need an intercept term β_0 since it is absorbed in the baseline hazard $\lambda_0(t)$. The estimate of β_1 is

$$\widehat{\beta}_1 = -0.569 \text{ (se} = 0.347),$$

giving a Wald statistic of $z = -1.64$, comparable with the result based on the normal model. The minus sign means the hazard of Group 2 is $e^{-0.569} = 0.57$ times the hazard of Group 1; recall that the rats in Group 2 live longer.

The estimated hazard ratio, however, is much smaller than what we get using a constant hazard assumption (i.e. the exponential model); in Example 11.5 we

obtain a ratio of $129.1/153.8 = e^{-0.175} = 0.84$. Since the shape of the hazard function in the Cox regression is free the result here is more plausible, although the ratio 0.57 is no longer interpretable as a ratio of mean lifetimes. The proportional hazard assumption itself can be checked; see, for example, Grambsch and Therneau (1994). □

11.8 Poisson point processes

Many random processes are events that happen at particular points in time (or space); for example, customer arrivals to a queue, times of equipment failures, times of epilepsy attacks in a person, etc. Poisson point or counting processes form a rich class of models for such processes. Let $N(t)$ be the number of events up to time t, dt a small time interval, and $o(dt)$ a quantity of much smaller magnitude than dt in the sense $o(dt)/dt \to 0$ as $dt \to 0$. The function $N(t)$ captures the point process and we say $N(t)$ is a Poisson point process with intensity $\lambda(t)$ if

$$N(t+dt) - N(t) = \begin{cases} 1 & \text{with probability } \lambda(t)dt \\ 0 & \text{with probability } 1 - \lambda(t)dt \\ > 1 & \text{with probability } o(dt) \end{cases}$$

and $N(t+dt) - N(t)$ is independent of $N(u)$ for $u < t$; the latter is called the independent increment property, which is a continuous version of the concept of independent trials.

It is fruitful to think of a Poisson process informally this way: $N(t+dt) - N(t)$, the number of events between t and $t+dt$, is Poisson with mean $\lambda(t)dt$. In a regular Poisson process, as defined above, there is a maximum of one event that can conceivably happen in an interval dt, so $N(t+dt) - N(t)$ is also approximately Bernoulli with probability $\lambda(t)dt$. Immediately identifying $N(t+dt) - N(t)$ as Poisson leads to simpler heuristics.

What we observe is determined stochastically by the intensity parameter $\lambda(t)$. Statistical questions can be expressed via $\lambda(t)$. It is intuitive that there is a close connection between this intensity function and the hazard function considered in survival analysis. The developments here and in the next sections are motivated by Whitehead (1983), Lawless (1987) and Lindsey (1995).

The first problem is, given a set of observations $t_1, \ldots t_n$, what is the likelihood of $\lambda(t)$? To answer this we first need two results from the theory of point processes. Both results are intuitively obvious; to get a formal proof see, for example, Diggle (1983). For the first result, since the number of events in each interval is Poisson, the sum of events on the interval $(0, T)$ is

$$\begin{aligned} N(T) &= \sum_{0<t<T} \{N(t+dt) - N(t)\} \\ &\sim \text{Poisson with mean } \sum_t \lambda(t)dt. \end{aligned}$$

So, in the limit we have:

Theorem 11.1 *$N(T)$ is Poisson with mean $\int_0^T \lambda(t)dt$.*

Furthermore, the way $t_1, \ldots, t_n$ are arranged will be determined by $\lambda(t)$, as shown by the following.

Theorem 11.2 *Given $N(T) = n$, the times $t_1, \ldots, t_n$ are distributed like the order statistics of an iid sample from a distribution with density proportional to $\lambda(t)$.*

To make it integrate to one, the exact density is given by

$$\frac{\lambda(t)}{\int_0^T \lambda(u)du}.$$

Suppose we model $\lambda(t) = \lambda(t, \theta)$, where θ is an unknown parameter. For example,

$$\lambda(t) = \alpha e^{\beta t}$$

with $\theta = (\alpha, \beta)$. For convenience, let

$$\Lambda(T) \equiv \int_0^T \lambda(t)dt$$

be the cumulative intensity. Then, given the observation times $0 < t_1, \ldots, t_n < T$, the likelihood of the parameters is

$$\begin{aligned} L(\theta) &= P\{N(T) = n) \times p\{t_1, \ldots, t_n | N(t) = n\} \\ &= e^{-\Lambda(T)} \frac{\Lambda(T)^n}{n!} \times n! \prod_{i=1}^n \frac{\lambda(t_i)}{\Lambda(T)} \\ &= e^{-\Lambda(T)} \prod_{i=1}^n \lambda(t_i). \end{aligned}$$

It is instructive to follow a heuristic derivation of the likelihood, since it applies more generally for point processes. First, partition the time axis into tiny intervals of length dt, such that only a single event can conceivably occur. On each interval let $y(t) \equiv N(t + dt) - N(t)$; then the time series $y(t)$ is an independent Poisson series with mean $\lambda(t)dt$. Observing $N(t)$ for $0 < t < T$ is equivalent to observing the series $y(t)$, where $y(t) = 1$ at $t_1, \ldots, t_n$, and zero otherwise; since we are thinking of dt as very small, the series $y(t)$ is mostly zero. For example, using dt as the time unit, events at times $t_1 = 3$ and $t_2 = 9$ during observation period $T = 10$ mean the series $y(t)$ is (0,0,1,0,0,0,0,0,0,1,0).

Given the observation times $0 < t_1, \ldots, t_n < T$, we obtain the likelihood

$$L(\theta) = \prod_t p(y_t)$$

$$\begin{aligned} &= \prod_t \exp\{-\lambda(t)dt\}\lambda(t)^{y(t)} \\ &\approx \exp\{-\sum_t \lambda(t)dt\}\prod_{i=1}^n \lambda(t_i) \\ &\approx \exp\{-\int_0^T \lambda(t)dt\}\prod_{i=1}^n \lambda(t_i), \end{aligned}$$

as we have just seen.

The last heuristic can be put into more technical notation. Let $dN(t) \equiv y(t) = N(t+dt) - N(t)$; then the log-likelihood can be written as

$$\begin{aligned} \log L(\theta) &= -\int_0^T \lambda(t)\ dt + \sum_{i=1}^n \log\lambda(t_i) \qquad (11.7) \\ &= -\int_0^T \lambda(t)\ dt + \int_0^T \log\lambda(t)\ dN(t), \end{aligned}$$

where for any function $h(t)$

$$\begin{aligned} \int_0^T h(t)\ dN(t) &\approx \sum_t h(t)\ dN(t) \\ &= \sum_{i=1}^n h(t_i). \end{aligned}$$

Note that, because of the way $dN(t)$ is defined, the intensity function $\lambda(t)$ can include values of the process $N(t)$, or any other data available prior to time t, without changing the likelihood. Andersen et al. (1993) provide a rigorous likelihood theory for counting processes.

The close connection between Poisson intensity modelling and hazard modelling of survival data in Section 11.7 is now clear: the likelihood (11.7) reduces to (11.4) if we limit ourselves to absorbing events (events that can occur only once, such as deaths, or events that end the observation period). It is also clear that the Poisson process models are more general than the survival models as they allow multiple end-points per subject.

Example 11.9: The following data (from Musa *et al.* 1987) are the times of 136 failures (in CPU seconds) of computer software during a period of $T = 25.4$ CPU hours. At each failure the cause is removed from the system. The questions are how fast can bugs be removed from the system, and how many bugs are still in the system. The histogram in Figure 11.5(a) shows that the number of failures decreases quickly over time. □

3	33	146	227	342	351	353	444	556	571	709
759	836	860	968	1056	1726	1846	1872	1986	2311	2366
2608	2676	3098	3278	3288	4434	5034	5049	5085	5089	5089
5097	5324	5389	5565	5623	6080	6380	6477	6740	7192	7447
7644	7837	7843	7922	8738	10089	10237	10258	10491	10625	10982

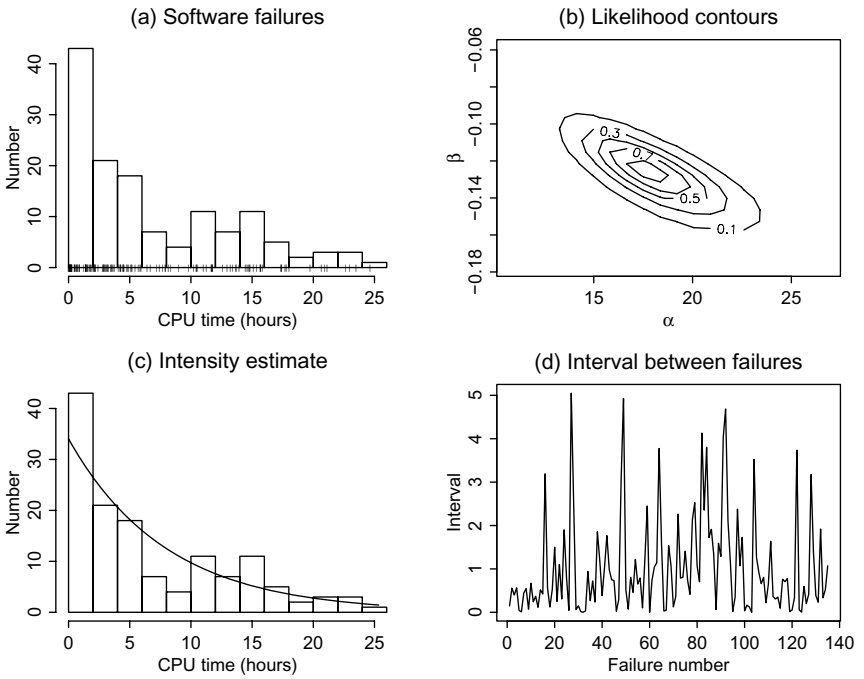

Figure 11.5: *(a) The vertical lines on the x-axis indicate the failure times; the histogram simply shows the number of failures in each 2-hour interval. (b) The likelihood of the parameters of an exponential decline model* $\lambda(t) = \alpha e^{\beta t}$. *(c) The fitted intensity function compared with the histogram. (d) The time series of scaled inter-failure times.*

```
11175 11411 11442 11811 12559 12559 12791 13121 13486 14708 15251
15261 15277 15806 16185 16229 16358 17168 17458 17758 18287 18568
18728 19556 20567 21012 21308 23063 24127 25910 26770 27753 28460
28493 29361 30085 32408 35338 36799 37642 37654 37915 39715 40580
42015 42045 42188 42296 42296 45406 46653 47596 48296 49171 49416
50145 52042 52489 52875 53321 53443 54433 55381 56463 56485 56560
57042 62551 62651 62661 63732 64103 64893 71043 74364 75409 76057
81542 82702 84566 88682
```

Assume a model $\lambda(t) = \alpha e^{\beta t}$, so

$$\begin{aligned} \Lambda(T) &= \int_0^T \alpha e^{\beta t} dt \\ &= \frac{\alpha}{\beta}(e^{\beta T} - 1) \end{aligned}$$

and, defining $\theta = (\alpha, \beta)$, we obtain

$$L(\theta) = \exp\left\{-\frac{\alpha}{\beta}(e^{\beta T} - 1)\right\} \prod_{i=1}^{136} (\alpha e^{\beta t_i}).$$

The contours of the likelihood function are given in Figure 11.5(b); as usual, they represent 10% to 90% confidence regions. Using a numerical optimization routine we find $\widehat{\alpha} = 17.79$ and $\widehat{\beta} = -0.13$, which means that for every CPU hour of testing the rate of failures is reduced by $(1-e^{-0.13})\times 100 = 12\%$. In Figure 11.5(c) the fit of the parametric model is compared with the histogram of failure times, showing a reasonable agreement. The parametric fit is useful for providing a simple summary parameter β, and prediction of future failures.

To check the Poisson assumption, we know theoretically that if a point process is Poisson, then the intervals between failures are independent exponentials with mean $1/\lambda(t)$, or the scaled intervals $\lambda(t_i)(t_i - t_{i-1})$ are iid exponentials with mean one. Figure 11.5(d) shows the plot of these scaled intervals, and Figure 11.6 shows the autocorrelation and the exponential QQ-plot, indicating reasonable agreement with Poisson behaviour.

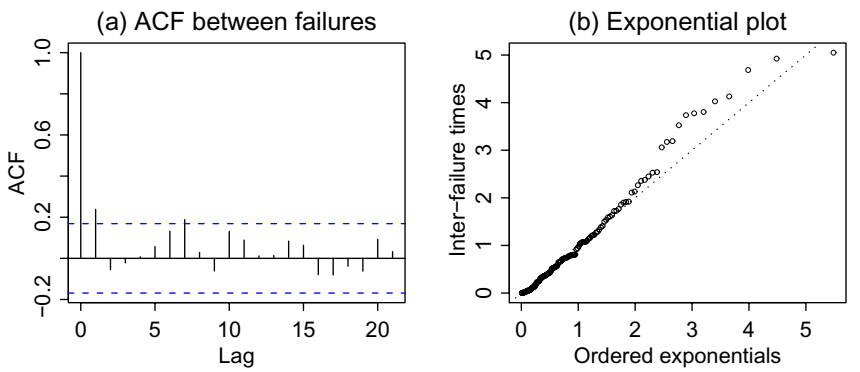

Figure 11.6: *Diagnostic check of the Poisson assumption: (a) autocorrelation between inter-event times and (b) exponential plot of inter-event times.*

11.9 Replicated Poisson processes

In biomedical applications we often deal with multiple processes, each generated by a subject under study. Since the notation can become cumbersome we will discuss the general methodology in the context of a specific example. Figure 11.7 shows a dataset from a study of treatment of epilepsy, where patients were randomized to either active or placebo groups. Because of staggered entry to the study, patients have different follow-up periods. The patients' families were asked to record the time of epileptic attacks during follow-up.

We will analyse this dataset using several methods of increasing complexity. Note the flexibility offered by the more complex methods in handling general intensity functions and covariates. Also, we will meet the Cox partial likelihood again.

Subject	x_i	T_i	n_i	Time of events
1	active	12	3	2.6 3.3 7.2
2	active	5	2	3.5 4.4
3	active	7	4	1.5 1.6 2.2 6.1
4	active	14	3	12.1 12.4 13.4
5	active	10	5	0.7 2.6 3.9 6.9 7.8
6	active	10	2	5.3 6.3
7	active	12	1	10.2
8	active	8	3	0.2 3.2 7.7
9	active	11	3	0.1 2 3.2
10	active	8	3	0.1 3.2 3.7
11	placebo	11	4	2.3 7.9 8 8.8
12	placebo	11	7	5.1 5.2 6.1 6.5 7.9 9.9 10.9
13	placebo	8	6	0.5 0.8 1.9 2.7 5.4 7.2
14	placebo	16	8	1.4 4.3 5 6 7.8 8.4 9.2 11.2
15	placebo	11	11	0.3 0.3 1.9 1.9 2.7 3.1 3.9 5.3 7 8.8 10.1
16	placebo	7	8	1.2 2.6 3.5 4.7 5.3 5.7 5.9 6.1
17	placebo	15	7	0.8 1.5 4.3 4.4 5.1 12.1 14
18	placebo	9	7	0.1 0.1 1 3.6 5.4 6.3 8.7
19	placebo	7	4	0.9 2.2 5.2 6.6
20	placebo	4	2	2.2 3.2
21	placebo	6	6	0.5 1.3 1.3 1.7 2.9 5.6
22	placebo	4	1	1.4

Figure 11.7: *The epilepsy data example. Patient i was followed for T_i weeks, and there were n_i events during the follow-up.*

Method 1

The first method will take into account the different follow-up periods among patients, but not use the times of attacks, and it cannot be generalized if we want to consider more covariates. Assume the event times within each patient follow a Poisson point process. Let λ_a and λ_p be the rate of attacks in the active and the placebo groups, i.e. assume that the rate or intensity is constant over time. Let

$$\begin{aligned} n_i &= \text{total number of attacks patient } i \\ y_a &= \sum_{i \in \text{active}} n_i \\ y_p &= \sum_{i \in \text{placebo}} n_i. \end{aligned}$$

Then

$$\begin{aligned} n_i &\sim \text{Poisson}(T_i\lambda) \\ y_a &\sim \text{Poisson}(\sum_{i \in \text{active}} T_i\lambda_a) \\ y_p &\sim \text{Poisson}(\sum_{i \in \text{placebo}} T_i\lambda_p), \end{aligned}$$

where λ in the first equation is either λ_a or λ_p. The parameter of interest is

$$\theta = \lambda_a/\lambda_p. \tag{11.8}$$

We can summarize the observed data in the following table:

	Active	Placebo
y_a or y_p	29	71
$\sum T_i$	97	109

Proceeding as in the aspirin data example in Section 4.7, conditional on $y_a + y_p$ the distribution of y_a is binomial with parameter $y_a + y_p = 100$ and probability

$$\frac{97\lambda_a}{97\lambda_a + 109\lambda_p} = \frac{97\theta}{97\theta + 109}.$$

So, the conditional likelihood is

$$L(\theta) = \left(\frac{97\theta}{97\theta + 109}\right)^{29} \left(1 - \frac{97\theta}{97\theta + 109}\right)^{71},$$

shown in Figure 11.8(a). We can verify that the MLE of θ is

$$\widehat{\theta} = (29/97)/(71/109) = 0.46.$$

The standard error is $\text{se}(\widehat{\theta}) = 0.10$, but the quadratic approximation is poor; the reader can verify that $\log\theta$ has a more regular likelihood. The likelihood of the null hypothesis H_0: $\theta = 1$ is tiny (approximately $e^{-6.9}$), leading to the conclusion that the active treatment has led to fewer attacks of epilepsy.

Method 2: Poisson regression

We now use the Poisson regression method, which could easily accommodate some covariates, but still does not use any information about event times. Let $x_i = 1$ if patient i belongs to the active group and zero otherwise. Using the same assumptions as in Method 1, the number of attacks n_i is Poisson with mean

$$\mu_i = T_i \exp(\beta_0 + \beta_1 x_i)$$

or

$$\log \mu_i = \log T_i + \beta_0 + \beta_1 x_i.$$

In generalized linear modelling $\log T_i$ is called an offset term. The likelihood of (β_0, β_1) can be computed as before, and the reader can verify the following summary table. Note that $e^{\widehat{\beta}_1} = e^{-0.78} = 0.46 = \widehat{\theta}$ as computed by Method 1.

Effect	Parameter	Estimate	se
Intercept	β_0	-0.43	0.11
Treatment	β_1	-0.78	0.21

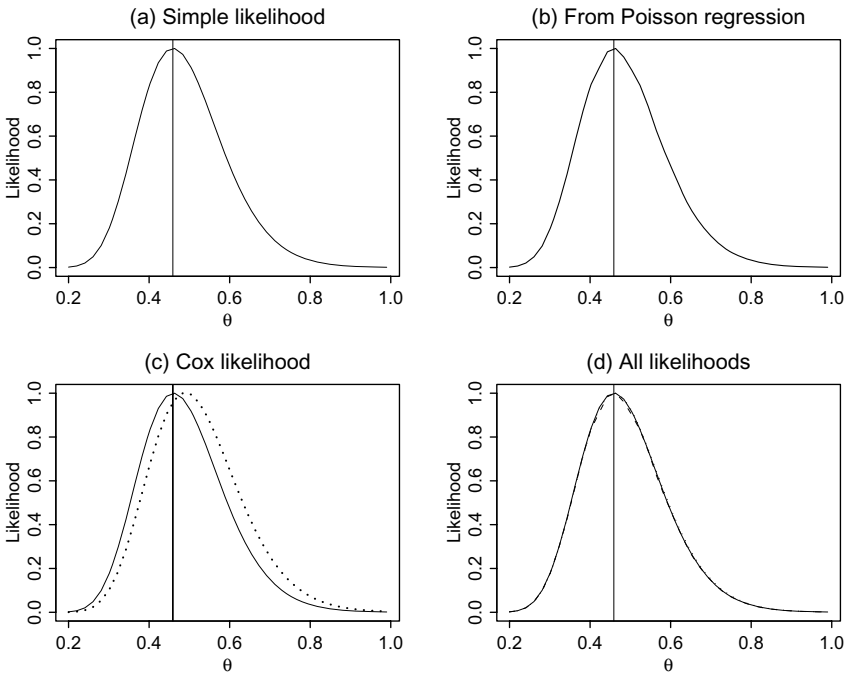

Figure 11.8: *Analysis of the epilepsy data using three methods: (a) Likelihood of θ using Method 1. (b) Profile likelihood of θ from Poisson regression. (c) Approximate (dotted) and true (solid) Cox partial likelihood. (d) All three likelihoods.*

To get a comparable likelihood function for the parameter $\theta = e^{\beta_1}$, we can compute the profile likelihood

$$L(\beta_1) = \max_{\beta_0} L(\beta_0, \beta_1),$$

and then evaluate $L(\theta) = L(\beta_1)$ at $\theta = e^{\beta_1}$. This likelihood function is given in Figure 11.8(b); it coincides with the likelihood given by Method 1.

Method 3

The previous methods make no use of the times of attacks, and they assume constant intensity for the Poisson processes for each subject. Neither can be generalized to overcome these limitations, so we will now consider a method that will address them at the expense of more complicated modelling.

We start by assuming the attacks for a patient follow a Poisson point process with intensity $\lambda_x(t)$, where x is the covariate vector. A useful general model for problems of this type is a proportional intensity model

$$\lambda_x(t) = \lambda_0(t, \alpha) g(x, \beta),$$

where $\lambda_0(t, \alpha)$ is the baseline intensity function with unknown parameter α. This is the analogue of the proportional hazard model in survival analysis.

The effect of covariate x is to modify the baseline intensity proportionally by a constant $g(x, \beta)$. The unknown regression parameter β expresses the effect of the covariate on the intensity level, for example using the usual log-linear model

$$\lambda_x(t) = \lambda_0(t, \alpha)e^{x'\beta}.$$

The baseline intensity $\lambda_0(t, \alpha)$ requires a parameter α, which is a nuisance parameter. The proportional intensity assumption is also analogous to the parallel regression assumption in the analysis of covariance; it may or may not be appropriate depending on the actual intensity functions.

From our previous theory, denoting $t_{i1}, \ldots, t_{in_i}$ to be the event times of subject i, the contribution of this subject to the likelihood is

$$L_i(\alpha, \beta) = e^{-\Lambda_{x_i}(T_i)} \prod_{j=1}^{n_i} \lambda_{x_i}(t_{ij}),$$

where

$$\begin{aligned} \Lambda_{x_i}(T_i) &= \int_0^{T_i} \lambda_{x_i}(t)dt \\ &= g(x_i, \beta) \int_0^{T_i} \lambda_0(t, \alpha)dt \\ &= g(x_i, \beta)\Lambda_0(T_i, \alpha). \end{aligned}$$

So

$$\begin{aligned} L_i(\alpha, \beta) &= e^{-g(x_i,\beta)\Lambda_0(T_i,\alpha)}\{g(x_i, \beta)\Lambda_0(T_i, \alpha)\}^{n_i} \prod_{j=1}^{n_i} \frac{\lambda_0(t_{ij}, \alpha)}{\Lambda_0(T_i, \alpha)} \\ &\equiv L_{1i}(\alpha, \beta)L_{2i}(\alpha), \end{aligned}$$

where

$$L_{1i}(\alpha, \beta) \equiv e^{-g(x_i,\beta)\Lambda_0(T_i,\alpha)}\{g(x_i, \beta)\Lambda_0(T_i, \alpha)\}^{n_i}.$$

The total likelihood from all, say m, individuals is

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{m} L_i(\alpha, \beta) \\ &= \prod_{i=1}^{m} L_{1i}(\alpha, \beta) \prod_i L_{2i}(\alpha) \\ &\equiv L_1(\alpha, \beta)L_2(\alpha). \end{aligned}$$

Hence, in proportional intensity models, the information about β is contained in the first term $L_1(\alpha, \beta)$; this is the likelihood based on the number of events from each individual

$$N_i \sim \text{Poisson}(\Lambda_0(T_i, \alpha)g(x_i, \beta)). \tag{11.9}$$

If we assume constant intensity, this reduces to the first method.

Having to model the baseline intensity $\lambda_0(t, \alpha)$ is literally a nuisance, since it is not directly relevant to the question of treatment comparisons. The data for estimating $\lambda_0(t, \alpha)$ are provided by the set of event times t_{ij}'s. The data structure makes it difficult to specify an appropriate model for $\lambda_0(t, \alpha)$; for example, we cannot simply plot the histogram of the event times. Fitting a model for the current example is left as Exercise 11.33. Given such a model, the decomposition of the likelihood above suggests the following method of estimation. (Actual implementation of the procedure is left as Exercise 11.34.)

1. Estimate α from $L_2(\alpha)$, which is a conditional likelihood given n_i's. This is fully determined by the set of event times t_{ij}'s.
2. Use $\widehat{\alpha}$ to compute $\Lambda_0(T_i, \widehat{\alpha})$.
3. Estimate β in the Poisson regression (11.9) based on the data (n_i, x_i) with $\Lambda_0(T_i, \widehat{\alpha})$ as an offset term.

To get further simplification, in particular to remove the nuisance parameter α, let us assume for the moment that $T_i \equiv T$. We first need the result that if X_i, for $i = 1, \ldots, m$, are independent Poisson(λ_i), then the conditional distribution of $(X_1, \ldots, X_m)$ given $\sum X_i$ is multinomial with parameters $(\pi_1, \ldots, \pi_m)$, where $\pi_i = \lambda_i / \sum_{j=1}^m \lambda_j$. This is applied to

$$N_i \sim \text{Poisson}(\Lambda_0(T, \alpha) g(x_i, \beta)).$$

Letting $n = \sum_i n_i$, we now have

$$\begin{aligned} L_1(\alpha, \beta) &= P(N_1 = n_1, \ldots, N_m = n_m) \\ &= P(N_1 = n_1, \ldots, N_m = n_m | \sum N_i = n) P(\sum N_i = n) \\ &= \prod_{i=1}^m \left(\frac{g(x_i, \beta)}{\sum_{j=1}^m g(x_j, \beta)} \right)^{n_i} P(\sum N_i = n) \\ &\equiv L_{10}(\beta) L_{11}(\alpha, \beta), \end{aligned} \tag{11.10}$$

where

$$L_{10}(\beta) = \prod_{i=1}^m \left(\frac{g(x_i, \beta)}{\sum_{j=1}^m g(x_j, \beta)} \right)^{n_i}.$$

Finding the exact formula for $L_{11}(\alpha, \beta)$ is left as Exercise 11.35. Now $L_{10}(\beta)$ is only a function of the parameter of interest β. Intuitively, if the baseline intensity $\lambda_0(t, \alpha)$ can be of *any shape*, then there is little information about β in the total number of events $\sum n_i$. This reasoning is similar to conditioning on the sum in the comparison of two Poisson means.

If we use the common log-linear model

$$\lambda_x(t) = \lambda_0(t, \alpha) e^{x'\beta},$$

then

$$L_{10}(\beta) = \prod_{i=1}^{m} \left(\frac{e^{x_i'\beta}}{\sum_{j=1}^{m} e^{x_j'\beta}} \right)^{n_i},$$

exactly the Cox partial likelihood for this particular setup.

In general, when $T_i \neq T$, and assuming a Poisson process with proportional intensity, the Cox partial likelihood is defined as the following. Note that it allows the covariate to change over time. Let

$$\begin{aligned} t_{ij} &= \text{the } j\text{'th event of subject } i, \\ x_{ij} &= \text{the covariate value of subject } i \text{ at time } t_{ij}, \\ R_{ij} &= \text{the set of subjects still at risk at time } t_{ij}. \end{aligned}$$

Then

$$L_{10}(\beta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \left(\frac{e^{x_{ij}'\beta}}{\sum_{k \in R_{ij}} e^{x_{kj}'\beta}} \right).$$

We can arrive at this likelihood by partitioning the time axis at T_i's and using, under the Poisson process, the independence of non-overlapping time intervals. The Cox partial likelihood looks more complicated, but it does correspond to the formula we derive before for the proportional hazard model. The significant contribution of the current approach is that it allows for multiple outcomes from each subject, i.e. the end-point does not have to be death, but it can be a recurrent event such as relapse. Furthermore, subjects can go in and out of the risk set depending whether they are being followed (events are being recorded) or not.

To apply this approach to the epilepsy data, consider first the approximate Cox partial likelihood, *pretending* that the follow-up period T_i's are the same for all subjects. Let the covariate $x_i = 1$ if i belongs to the active therapy group and $x_i = 0$ otherwise. Then

$$L_{10}(\beta) = \prod_{i=1}^{m} \left(\frac{e^{x_i'\beta}}{\sum_{j=1}^{m} e^{x_j'\beta}} \right)^{n_i},$$

where $e^{x_i'\beta} = \theta$ if $x_i = 1$ and $e^{x_i'\beta} = 1$ otherwise; then $\sum_{j=1}^{m} e^{x_j'\beta} = 10\theta + 12$, so

$$L_{10}(\beta) = \left(\frac{\theta}{10\theta + 12} \right)^{y_a} \left(\frac{1}{10\theta + 12} \right)^{y_p},$$

the same as the likelihood given by Method 1 had we used $T_i \equiv T$. Figure 11.8(c) shows the approximate and the true Cox partial likelihood for the dataset. Computation of the true Cox partial likelihood is left as an exercise. As shown in Figure 11.8(d), in this case the likelihoods from all methods virtually coincide.

11.10 Discrete time model for Poisson processes

In this section we describe in detail the close connection between censored survival data and the Poisson process, and the unifying technique of Poisson regression to handle all cases. To avoid unnecessary technicalities we consider discrete time models. While continuous time models are more elegant mathematically, in real studies we measure time in a discrete way, say hourly or daily. Hence a discrete Poisson process model is actually quite natural, and its statistical analysis reduces to the standard Poisson regression. Given a natural time unit dt we can model the outcomes $y(t)$ as a Poisson series with mean $\lambda(t)dt$. This model is valid not just as an approximation of the continuous model; in particular, and there is no need for special handling of ties, which are a problem under continuous time models. The stated results here generally extend to continuous time as a limit of discrete time as dt tends to zero.

The main advantage of the Poisson regression approach for survival-type data is the flexibility in specifying models. Moreover, the model elements generally have a clear interpretation. The price is that each outcome value, a single number y_i, generates a long array $y_i(t)$, so the overall size of the problem becomes large. For a single Poisson process the observed $y(t)$ is a Poisson time series; if dt is small enough then $y(t)$ takes 0-1 values, but this is not a requirement. The log-likelihood contribution from a single series is

$$\sum_t y(t) \log \lambda(t) - \sum_t \lambda(t)dt.$$

To see the generality of this setup, suppose we observe survival data (y_i, δ_i) as described in Section 11.5; δ_i is the event indicator, which is equal to one if y_i is a true event time. After partitioning time by interval unit dt we convert each observed y_i into a 0–1 series $y_i(t)$. For example, $(y_i = 5, \delta_i = 1)$ converts to $y_i(t) = (0, 0, 0, 0, 1)$, while $(y_i = 3, \delta_i = 0)$ converts to (0,0,0). Therefore, the log-likelihood contribution of (y_i, δ_i) is

$$\log L_i = \delta_i \log \lambda_i(y_i) - \sum_{t=1}^{y_i} \lambda_i(t)dt.$$

In the limit it can be written as

$$\log L_i = \delta_i \log \lambda_i(y_i) - \int_0^{y_i} \lambda_i(t)dt,$$

exactly the log-likelihood (11.4) we derive in Section 11.7 for general survival data. Any survival regression model that can be expressed in terms of a hazard function has a discrete version as a Poisson model.

Now for convenience we set the time unit $dt \equiv 1$, and let $y_i(t)$ be Poisson with mean $\lambda_i(t)$. To analyse the effect of covariate x_i on the intensity function $\lambda_i(t)$ we may consider, for example, the log-link

$$\log \lambda_i(t) = \alpha(t) + x_i'\beta,$$

which implies a proportional intensity model with baseline intensity

$$\lambda_0(t) = e^{\alpha(t)}.$$

The nuisance parameter $\alpha(t)$ needs further modelling.

Example 11.10: The exponential regression model for survival data (with a maximum of one event per subject) is equivalent to specifying a constant baseline intensity $\lambda_0(t) \equiv \lambda$, or

$$\log \lambda_i(t) = \alpha_0 + x_i'\beta,$$

with $\alpha(t) = \alpha_0 = \log \lambda$. The general extreme-value model, or the Weibull model, is equivalent to specifying a log-linear model

$$\log \lambda_i(t) = \alpha_0 + \alpha_1 \log t + x_i'\beta$$

or

$$\log \lambda_0(t) = \alpha(t) \equiv \alpha_0 + \alpha_1 \log t.$$

To get a specific comparison, recall the survival regression model in Section 11.6 for the underlying survival time t_i as

$$\log t_i = \beta_0 + x_i'\beta_w + \sigma W_i,$$

where W_i's are iid with standard extreme-value distribution. Then we have the following relationships (Exercise 11.37):

$$\begin{aligned} \alpha_0 &= -\beta_0/\sigma - \log \sigma \\ \alpha_1 &= 1/\sigma - 1 \\ \beta &= -\beta_w/\sigma. \end{aligned}$$

For multiple events, we can simply modify $\log t$ to $\log z_t$, where z_t is the time elapsed since the last event. Other models can be specified for the baseline intensity function. □

The Cox proportional hazard model is associated with nonparametric $\alpha(t)$. We obtain it by setting $\alpha(t)$ to be a categorical parameter, with one parameter value for each time t. This is useful for modelling, since we can then compare the parametric versus Cox models.

The Cox partial likelihood itself can be derived as a profile likelihood. Suppose the length of observation time is T_i and $y_i(t)$ is nonzero at event times $t_{i1}, \ldots, t_{in_i}$ and zero otherwise, and denote $y_{ij} \equiv y_i(t_{ij})$. If there is no tie, $y_{ij} = 1$ for all i and j. The notation will become very cumbersome here; see the epilepsy data example to get a concrete idea. The log-likelihood contribution of the i'th series is

$$\begin{aligned} \log L_i &= -\sum_{t=1}^{T_i} \lambda_i(t) + \sum_{j=1}^{n_i} y_{ij} \log \lambda_i(t_{ij}) \\ &= -\sum_{t=1}^{T_i} e^{\alpha(t)+x_{it}'\beta} + \sum_{j=1}^{n_i} y_{ij}\{\alpha(t_{ij}) + x_{ij}'\beta\}, \end{aligned}$$

where x_{it} is the value of the covariate x_i at time t and x_{ij} is the value at time t_{ij}. The overall log-likelihood is

$$\log L(\alpha, \beta) = \sum_i \log L_i.$$

To get the profile likelihood of β, we need to find the MLE of $\alpha(t)$ at each value β. To this end we analyse the event and nonevent times separately. At a nonevent time $t \neq t_{ij}$ the estimate $\widehat{\alpha}(t)$ satisfies

$$-\sum_i e^{\widehat{\alpha}(t)+x'_{it}\beta} = 0$$

or $e^{\widehat{\alpha}(t)} = 0$.

At an event time $t = t_{ij}$ we need to keep track of the multiplicity of $\alpha(t_{ij})$ in the first term of the log-likelihood. This can be seen graphically by drawing parallel lines to represent the observation intervals for the subjects. We define the risk set R_{ij} as the set of subjects still under observations at time t_{ij}. We can write the total log-likelihood as

$$\begin{aligned}\log L(\alpha, \beta) &= -\sum_{i=1}^{n} \sum_{t \neq (t_{i1} \ldots t_{in_i})} e^{\alpha(t)+x'_{it}\beta} \\ &\quad -\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k\in R_{ij}} e^{\alpha(t_{ij})+x'_{kj}\beta} \\ &\quad +\sum_{i=1}^{n}\sum_{j=1}^{n_i} y_{ij}\{\alpha(t_{ij}) + x'_{ij}\beta\},\end{aligned}$$

so, in this case $\widehat{\alpha}(t_{ij})$ satisfies

$$\sum_{k\in R_{ij}} e^{\widehat{\alpha}(t_{ij})+x'_{kj}\beta} = y_{ij}$$

or

$$e^{\widehat{\alpha}(t_{ij})} = \frac{y_{ij}}{\sum_{k\in R_{ij}} e^{x'_{kj}\beta}}.$$

Substituting $\widehat{\alpha}(t)$ for all values of t we get the profile likelihood of β, which is proportional to

$$L(\beta) = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left(\frac{e^{x'_{ij}\beta}}{\sum_{k\in R_{ij}} e^{x'_{kj}\beta}}\right)^{y_{ij}}.$$

If there is no tie, $y_{ij} = 1$ for all i and j, and

$$L(\beta) = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left(\frac{e^{x'_{ij}\beta}}{\sum_{k\in R_{ij}} e^{x'_{kj}\beta}}\right),$$

exactly the Cox partial likelihood stated in the previous section. Cox's (1972) treatment of ties does not correspond to the result shown above; see Whitehead (1980) for a discussion.

From the derivation it is clear that we can estimate the baseline intensity or hazard function by

$$\widehat{\lambda}_0(t) = e^{\widehat{\alpha(t)}},$$

which takes nonzero values only at the event times. A sensible continuous estimate can be found by smoothing; see Chapter 18.

Example 11.11: For two-sample problems we can arrange the data to limit the problem size. This is because there are multiples of $e^{\alpha(t)+x_i\beta}$ according to $x_i = 0$ or 1. The multiplicity is simply the number of subjects at risk at each event time; this will enter as an offset term in the Poisson regression. Furthermore, to produce Cox regression results we should only consider the statistics at the event times; including time points where there is no event will produce numerical problems, since in this case $e^{\widehat{\alpha(t)}} = 0$.

For the epilepsy data example in the previous section suppose we choose the time unit $dt = 1$ week. Then the data can be summarized as in Table 11.2. The index g in the table now refers to week-by-treatment grouping, with a total of 29 groups with at least one event; R_g is the corresponding number at risk at the beginning of the week; y_g is the number of events during the week.

g	x_g	week$_g$	R_g	y_g	g	x_g	week$_g$	R_g	y_g
1	1	0	10	4	15	0	0	12	9
2	1	1	10	2	16	0	1	12	11
3	1	2	10	4	17	0	2	12	7
4	1	3	10	7	18	0	3	12	5
5	1	4	10	1	19	0	4	10	4
6	1	5	9	1	20	0	5	10	12
7	1	6	9	3	21	0	6	9	6
8	1	7	8	3	22	0	7	7	5
9	1	8	6	0	23	0	8	6	5
10	1	9	6	0	24	0	9	5	2
11	1	10	4	1	25	0	10	5	2
12	1	11	3	0	26	0	11	2	1
13	1	12	1	2	27	0	12	2	1
14	1	13	1	1	28	0	13	2	0
					29	0	14	2	1

Table 11.2: *Data setup for Poisson regression of the epilepsy example. The time intervals are of the form* $[k, k+1)$*; for example, an event at* $t = 1.0$ *is included in week-1, not week-0.*

Now assume y_g is Poisson with mean μ_g and use the log-link

$$\log \mu_g = \log R_g + \alpha(\text{week}_g) + \beta_1 x_g,$$

where $\log R_g$ is an offset, and β_1 is the treatment effect. The function $\alpha(\text{week}_g)$ is a generic function expressing the week effect. For example,

$$\alpha(\text{week}_g) \equiv \alpha_0$$

Effect	Parameter	Estimate	se
intercept	α_0	−0.247	0.284
treatment	β_1	−0.767	0.221
week-1	α_1	0.000	0.392
week-2	α_2	−0.167	0.410
week-3	α_3	−0.080	0.400
week-4	α_4	−0.828	0.526
week-5	α_5	0.160	0.392
week-6	α_6	−0.134	0.434
week-7	α_7	−0.045	0.450
week-8	α_8	−0.317	0.526
week-9	α_9	−1.112	0.760
week-10	α_{10}	−0.580	0.641
week-11	α_{11}	−0.975	1.038
week-12	α_{12}	0.444	0.641
week-13	α_{13}	−0.655	1.037
week-14	α_{14}	−0.446	1.040

Table 11.3: *Estimates from Poisson regression of the epilepsy data example.*

for constant intensity function, or

$$\alpha(\text{week}_g) \equiv \alpha_0 + \alpha_1 \times \log \text{week}_g$$

for log-linear time effect, etc. The most general function is setting 'week' to be a categorical variable, with a single parameter for each week; for convenience we set $\alpha_0 \equiv \alpha(0)$, and for $j > 0$ define $\alpha_j = \alpha(j) - \alpha(0)$ as the week-j effect relative to week-0. These options can be compared using the AIC.

For categorical week effects the parameter estimates are given in Table 11.3. The estimate of the treatment effect $\widehat{\beta}_1 = -0.767$ (se $= 0.221$) is similar to the estimates found by different methods in the previous section. Fitting various other models for $\alpha(\text{week}_g)$ is left as an exercise. □

11.11 Exercises

Exercise 11.1: Simulate AR(1) processes shown in Example 11.1, and verify the likelihoods given in Figure 11.1.

Exercise 11.2: In the previous exercise, compare the Fisher information of $\widehat{\phi}_1$ based on the full likelihood $L(\theta)$ and the conditional likelihood $L_2(\theta)$.

Exercise 11.3: For the AR(2) model

$$x_t = \theta_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + e_t,$$

where e_t's are iid $N(0, \sigma^2)$. Derive the full and conditional likelihood of the parameters.

Exercise 11.4: Suppose we observe time series data $x_1, \ldots, x_n$, which we identify as an MA(1) series

$$x_t = \theta_0 + e_t - \theta_1 e_{t-1}.$$

Assuming e_t is $N(0, \sigma^2)$ derive the likelihood of the parameters. Is there a form of likelihood that is simpler to compute as in the AR models? Discuss a practical

method to compute the likelihood. (Hint: to get the joint density of $x_1, \ldots, x_n$, first consider a transformation from $e_0, e_1, \ldots, e_n$ to $e_0, x_1, \ldots, x_n$, then evaluate the marginal density of $x_1, \ldots, x_n$.)

Exercise 11.5: Verify the standard errors given for the MLEs in Example 11.2.

Exercise 11.6: Use the likelihood approach to test a general hypothesis H_0 : $\theta_{01} = \theta_{11}$ to the data in Example 11.2. Compare the result using the standard error approximation. Interpret what the hypothesis means.

Exercise 11.7: A time series x_t is Markov of order 2, if the conditional distribution of x_t given its past depends only on the last two values. That is,

$$p(x_t|x_{t-1}, \ldots) = p(x_t|x_{t-1}, x_{t-2}).$$

(a) Describe the parameters of the simplest second-order Markov chain with 0-1 outcomes.

(b) Find the likelihood of the parameters given an observed series $x_1, \ldots, x_n$. Identify the 'easy' part of the likelihood.

(c) Estimate the parameters for the data in Example 11.2.

(d) Using the likelihood ratio test, check whether the first-order model is adequate.

Exercise 11.8: Investigate the association between the asthma episodes and the pollution level using the data given in Section 11.2. Compare the results of the two types of regression mentioned in the section.

Exercise 11.9: Verify the summary tables given in Section 11.3.

Exercise 11.10: Draw the profile likelihood of the odds-ratio parameters for the data shown by the 2×2 tables in Section 11.3. Compute the approximate 95% CI for the odds ratio. Compare the χ^2 tests with the Wald tests.

Exercise 11.11: For the data in Section 11.3, derive the likelihood of the parameters p and q for both groups given the summary table (there is a total of four parameters: p_0 and q_0 for the control group and p_1 and q_1 for the treatment group). Use the likelihood ratio test to test the hypothesis H_0: $q_0 = q_1$, and compare with the result given in the section.

Exercise 11.12: In Example 11.3 suppose the follow-up observation of patient 2 of the control group is missing at week 6, so we observe

w0	w2	w4	w6	w8
5	5	5	–	5

Derive the likelihood contribution of this patient.

Exercise 11.13: In Example 11.3, describe a regression model to take into account some possible baseline differences in the two groups such as in age or prior history of other related diseases. Note that the purpose of the analysis is still to compare the response to treatment.

Exercise 11.14: Describe ways to check whether the simplified transition probability matrix given in Section 11.3 is sensible for the data.

Exercise 11.15: The data in Figure 11.9 were collected by Dr. Rosemary Barry in a homeophatic clinic in Dublin from patients suffering from arthritis. Baseline information include age, sex (1= male), arthritis type (RA= rheumathoid arthritis, OA= ostheo-arthritis), and the number of years with the symptom. The pain score was assessed during a monthly followup and graded from 1 to 6 (high is

worse) with −9 to indicate missing. All patients were under treatment, and only those with a baseline pain score greater than 3 and a minimum of six visits are reported here. Investigate the trend of pain score over time and the effect of the covariates.

No	Age	Sex	Type	Years	Pain scores										
1	41	0	RA	10	4	4	5	3	3	5					
2	34	0	RA	0	5	2	2	3	2	3	2				
3	53	1	RA	1	5	4	3	2	2	2	1	2	1		
4	38	0	RA	12	6	6	5	6	5	5	5	6			
5	51	0	RA	2	5	5	5	5	4	4					
6	70	0	RA	40	5	4	4	3	3	3	3	3			
7	74	0	RA	5	4	4	5	4	4	3	5	−9	6	5	4	3
8	56	0	RA	1	5	4	−9	5	5	5	4	4	3	3	3	−9
9	57	0	RA	33	5	5	6	5	6	6	−9	6	5	5	5	
10	65	1	RA	18	5	6	6	6	6	6						
11	61	0	RA	12	6	3	5	2	2	2	5	5				
12	64	0	RA	10	4	3	4	4	4	4	3	−9	6	6		
13	47	0	RA	10	5	3	4	4	3	3						
14	59	0	RA	1	6	5	5	6	4	4						
15	54	1	RA	2	6	5	3	4	4	6						
16	74	0	RA	14	4	4	4	4	−9	−9	−9					
17	57	0	RA	2	4	3	3	2	1	2	2	2				
18	86	0	RA	5	4	3	3	2	4	4	4	4	4	5	3	
19	69	0	RA	39	4	2	4	4	5	2	4	6	5			
20	45	0	RA	7	4	4	4	−9	4	4	4					
21	45	0	RA	20	5	5	4	4	5	4	3					
22	70	1	RA	6	6	6	6	6	6	6	6					
23	38	0	RA	0	4	5	4	2	2	3	3	2	2	4		
24	68	0	RA	16	6	−9	4	4	4	3						
25	18	1	RA	1	4	4	3	1	−9	−9	−9	1	1			
26	58	0	RA	1	5	4	3	4	3	3						
27	62	0	RA	1	4	3	5	4	3	4	2					
28	56	0	OA	6	4	5	6	3	5	4	3	3	5			
29	68	0	OA	10	5	5	4	3	2	2	2	2	2	1		
30	64	1	OA	5	5	4	3	5	−9	4						
31	49	0	OA	8	4	4	4	3	3	3	3					
32	66	0	OA	5	4	3	4	3	3	4						
33	70	0	OA	7	4	2	1	1	1	1	2	1				
34	61	0	OA	5	6	−9	3	3	3	5	−9					
35	41	0	OA	15	5	4	3	4	3	2	1	1	2			
36	57	0	OA	4	5	6	6	3	5	6	5	3	5			
37	49	0	OA	4	5	2	2	3	1	2	1	1	1	1	2	1
38	57	0	OA	14	4	4	4	3	5	3	5	3	4	4	3	2
39	78	0	OA	4	5	5	3	4	−9	5						
40	61	0	OA	20	4	4	3	3	3	3	2	3	3	2	2	
41	82	0	OA	40	5	−9	−9	3	3	3	4	5	6	6		
42	48	0	OA	1	5	3	1	1	1	1	2					
43	51	0	OA	2	5	3	3	3	2	3	3					
44	54	0	OA	2	4	4	4	3	2	2	2	2	2			
45	54	0	OA	1	5	5	3	6	6	6						
46	68	0	OA	15	6	5	5	6	3	2						
47	70	1	OA	15	4	4	4	5	5	4						
48	63	1	OA	1	4	4	4	3	2	2	2	1				
49	56	0	OA	4	6	5	3	3	3	3	4	4				
50	66	1	OA	5	4	4	3	3	2	2						
51	64	1	OA	2	5	2	1	1	1	1	1	−9				
52	53	1	OA	2	4	3	3	3	3	3	2	3				
53	58	0	OA	5	4	4	4	4	3	3	2	2	2	2	3	
54	65	1	OA	30	5	5	6	6	6	6	6	6				
55	74	0	OA	40	6	3	1	1	2	1	2	1	2			
56	60	0	OA	57	4	3	4	2	3	3	2	2	2	−9	−9	
57	88	0	OA	10	4	5	5	5	6	4	3	3	3	5		
58	66	0	OA	10	4	2	2	2	1	4	4	3	4			
59	71	0	OA	20	4	4	4	3	4	3						
60	66	0	OA	6	5	−9	4	5	4	3	4					

Figure 11.9: *Arthritis data*

Exercise 11.16: Show that the Ising model in Section 11.4 does imply the conditional probability in the logistic form in both the one-dimensional and two-dimensional models.

Exercise 11.17: Verify the maximum pseudo-likelihood estimates and the standard errors of the parameters (α, β) given in Example 11.4.

Exercise 11.18: To appreciate that the product of conditional probabilities is not really a probability specification, describe how you might simulate a realization of an Ising model given a certain parameter (α, β). (Being able to simulate data is important for Monte Carlo tests.)

Exercise 11.19: Bartlett (1971) reported the absence/presence of a certain plant in a 24×24 square region, reflecting some interaction or competition among *Carex arenaria* plant species.

```
0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 0 1 0 0 1 1 0 1 0
0 1 0 0 1 1 1 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 0 1
1 1 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0
0 0 1 1 1 1 1 0 1 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0
0 1 1 0 1 1 1 1 1 0 0 0 0 1 1 1 0 1 1 0 0 0 0 0
0 1 0 0 0 1 0 1 0 1 1 1 1 1 0 0 0 0 1 1 0 0 0 0
0 1 0 1 1 0 1 0 1 0 0 0 1 0 0 1 1 0 0 1 0 0 0 0
0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0
0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0
0 0 1 0 0 0 1 1 0 1 0 1 1 1 1 1 1 1 0 0 0 0 1 1
0 1 0 0 1 1 0 0 0 0 0 0 0 1 1 0 0 1 0 1 0 1 0 1
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 1 1 0 0 1
1 0 0 0 0 0 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 1 1 0 1 0 1
0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 1 1 1 1 1 0 1
1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 1
0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1
1 0 0 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 0 1
0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
```

(a) Perform and interpret the logistic regression analysis of the Ising model on the dataset.

(b) Test the randomness of the pattern by grouping the data, say into 3×3 or 4×4 squares, and test whether the number of plants on each square follows a Poisson model. What do you expect to see if there is clustering (positive dependence) or if there is negative dependence?

Exercise 11.20: Extend the pseudo-likelihood approach to analyse spatial count data, i.e. y_k is Poisson rather than Bernoulli as in the text. Simulate simple one- and two-dimensional Poisson data and apply the methodology. For a more advanced exercise, a real dataset is given in Table 6 of Breslow and Clayton (1993).

Exercise 11.21: Find the profile likelihood of $\theta = \theta_1/\theta_2$ for the rat data example using the exponential model; report the approximate 95% CI for θ. Compare the Wald statistic based on $\widehat{\theta}$ and $\log \widehat{\theta}$; which is more appropriate?

Exercise 11.22: Perform the group comparison using the normal model as described in Section 11.5. For simplicity, assume the two groups have common variance σ^2, and eliminate σ^2 by replacing it with its estimate from uncensored data (i.e. use the estimated likelihood). Find the profile likelihood of $\mu_1 - \mu_2$.

Exercise 11.23: Repeat the previous exercise, but with the unknown σ^2 estimated using the likelihood from all the data.

Exercise 11.24: Verify the regression analysis performed in Example 11.6.

Exercise 11.25: Derive theoretically the Fisher information for the regression parameters (β_0, β_1) under the exponential and the general extreme-value distributions in Section 11.6.

Exercise 11.26: Compute the Fisher information for the observed data, and verify the standard errors given in Example 11.6. Check the quadratic approximation of the profile likelihood of the slope parameter β_1.

Exercise 11.27: Find the profile likelihood of the scale parameter σ in Example 11.6, and verify the likelihood ratio statistic $W = 44.8$. Report also the Wald test of H_0: $\sigma = 1$, and discuss whether it is appropriate.

Exercise 11.28: If independent lifetimes T_1 and T_2 have proportional hazards, say $\lambda_i(t) = \lambda_0(t)\eta_i$ for $i = 1, 2$ respectively, then show that

$$P(T_1 < T_2) = \eta_1/(\eta_1 + \eta_2)$$

regardless of the shape of the baseline hazard function $\lambda_0(t)$. Generalize the result for $P(T_{i_1} < T_{i_2} < \cdots < T_{i_n})$.

Exercise 11.29: To repeat the exercise given Example 11.7, suppose we observe the following dataset:

i	x_i	y_i	δ_i
1	0	10	1
2	0	5	1
3	0	13	0
4	1	7	1
5	1	21	1
6	1	17	1
7	1	19	0

Assume a proportional hazard model

$$\lambda_i(t) = \lambda_0(t)e^{x_i\beta}.$$

(a) Assume the baseline hazard $\lambda_0(t) \equiv \lambda$, i.e. a constant hazard. Compute the profile likelihood of β.

(b) Compare the profile likelihood in part (a) with the Cox partial likelihood.

(c) Simplify the Cox partial likelihood as much as possible and try to interpret the terms.

(d) Compute the Fisher information of β from the Cox partial likelihood and compare with the information if there is no censoring (i.e. replace the two censored cases with $\delta_i = 1$).

Exercise 11.30: Verify the Cox regression analysis in Example 11.8. Compare the Cox partial likelihood with the profile likelihood found using the exponential model in Exercise 11.21. Check the quadratic approximation of the Cox partial likelihood.

Exercise 11.31: Verify the analysis of software failure data given in Example 11.9. Provide the standard error for $\widehat{\beta}$. Predict the number of bugs still to be found in the system.

Exercise 11.32: Different methods to analyse the epilepsy data in Section 11.9 lead to very similar likelihoods. Provide some explanation, and think of other situations where the three methods would give different results.

Exercise 11.33: For the data in Figure 11.7 find a reasonably simple parametric model for the baseline intensity $\lambda_0(t, \alpha)$ assuming a proportional hazard model.

Exercise 11.34: Assume an exponential decay model as the underlying parameter model for the epilepsy data in Section 11.9. Apply the iterative procedure suggested in Method 3 to estimate the regression parameter β. Compare the likelihood with those obtained by the other methods.

Exercise 11.35: Find the exact formula for $L_{11}(\alpha, \beta)$ in (11.10).

Exercise 11.36: Compare the three methods of analysis for the data in Figure 11.10. The structure is the same as in the previous dataset, but now there is also an additional covariate.

Subject	age_i	x_i	T_i	n_i	Time of events
1	16	active	14	1	1.6
2	22	active	14	3	2 3.2 4.7
3	20	active	1	0	
4	22	active	6	1	0.8
5	21	active	27	1	5
6	17	active	1	0	
7	18	active	1	0	
8	18	active	1	1	0
9	17	active	1	0	
10	22	active	15	2	1.1 5.9
11	21	placebo	13	5	3.9 3.9 4.3 5.5 11.8
12	21	placebo	4	1	0.9
13	22	placebo	6	6	0.5 0.5 0.8 1.2 1.2 2
14	17	placebo	18	1	9.2
15	18	placebo	13	3	2.2 2.5 4.7
16	21	placebo	8	3	0.3 1.4 4.5
17	20	placebo	4	2	3.1 3.4
18	21	placebo	14	0	
19	24	placebo	5	0	
20	24	placebo	17	0	
21	22	placebo	8	4	1.8 6.8 7.2 7.6
22	24	placebo	1	1	0.5

Figure 11.10: *Another sample dataset from an epilepsy clinical trial.*

Exercise 11.37: Prove the relationships stated in Example 11.10 between the coefficients of the Weibull and Poisson regression models.

Exercise 11.38: Verify the relationships stated in Example 11.10 using the rat data example given in Section 11.5.

Exercise 11.39: Verify the data setup in Table 11.2 for the Poisson regression of the epilepsy data. Analyse the data using a constant and log-linear week effect, and compare the different models using the AIC.