29
Conditional logistic regression

In an individually matched case-control study, it is necessary to introduce a new parameter for every case-control set, if the matching is to be preserved in the analysis. This means that the number of parameters in the model exceeds the number of cases and in this case the profile likelihood does not lead to sensible estimates. Instead the nuisance parameters must be eliminated using a conditional likelihood. In Chapter 19 we indicated how this is done for a simple binary exposure. In this chapter we show how to use a conditional likelihood with the logistic regression model.

29.1 The logistic model

Suppose we wish to fit a logistic regression model which contains parameters for the case-control sets in addition to parameters for the effects of two explanatory variables A and B. Using a categorical variable to define the set to which each subject belongs, the model would be written

$$\log(\text{Odds}) = \text{Corner} + \text{Set} + \text{A} + \text{B}.$$

The model can also be written in the multiplicative form as

$$\text{Odds} = \text{Corner} \times \text{Set} \times \text{A} \times \text{B}.$$

For the case where A has three levels and B has two levels, the parameters in this model are Corner, A(1), A(2), B(1), together with

$$\text{Set}(1),\ \text{Set}(2),\ \cdots,\ \text{Set}(N-1)$$

where N is the number of case-control sets. These set parameters are those used in standard logistic regression models, but they are no longer the most convenient choice. It is now more convenient to choose a separate corner for each set, namely the odds parameter for each set when A and B are at level 0. The corner for the first case-control set is the corner parameter referred to above, the corner for the second case-control set is

$$\text{Corner} \times \text{Set}(1),$$

and so on. This corresponds to splitting the terms in the model into two groups, as follows:

$$\text{Odds} = \boxed{\text{Corner} \times \text{Set}} \times \boxed{\text{A} \times \text{B}}.$$

The first part of the model contains the separate corners, and these are the nuisance parameters to be eliminated, while the second part contains the effects of interest. When a conditional logistic program is used to fit this model the nuisance parameters are eliminated using conditional likelihood and estimates of the effects of A and B are reported. No estimates of either the corner or the set parameters are obtained in this method, so none can be reported.

To see how the nuisance parameters are eliminated using conditional likelihood it is convenient to return to the algebraic notation for parameters using Greek letters. For any particular case-control set let the corner parameter be ω_{C}. Let the odds for any subject in the set be ω_i, where $i = 1, 2, \ldots$, indexes the subjects within the case-control set, and write

$$\omega_i = \omega_{\text{C}}\theta_i,$$

so that θ_i is the ratio of the odds for subject i to the corner odds. The way θ is related to the effects of A and B is determined by the $\boxed{\text{A}\times\text{B}}$ part of the model. The corner parameter refers to subjects within the set with both A and B at level 0, so that the value of θ for such subjects is 1. For subjects with A at level 1 and B at level 0,

$$\theta = \text{A}(1),$$

for subjects with A at level 1 and B at level 1,

$$\theta = \text{A}(1) \times \text{B}(1),$$

and so on.

To be specific about which case-control set is being referred to, the parameters should be written with superscripts t, as in

$$\omega_i^t = \omega_{\text{C}}^t\theta_i^t.$$

where $t = 0, 1, 2, \ldots$ refers to the levels of the variable defining set membership. The parameters ω_{C}^t correspond to the

$$\boxed{\text{Corner} \times \text{Set}}$$

part of the model, and are the nuisance parameters to be eliminated. In the rest of this chapter we shall derive the contribution to the conditional

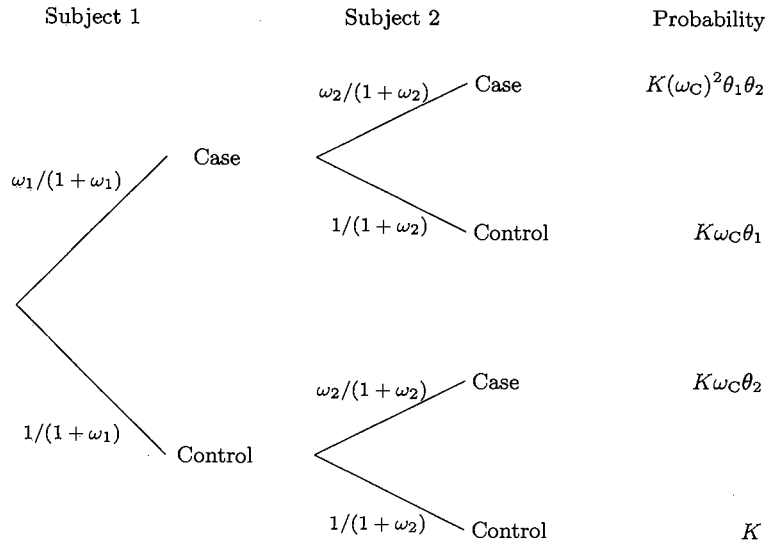

Fig. 29.1. Disease status for two subjects in a case-control study.

log likelihood for a single case-control set, and shall therefore omit the t superscript. The total log likelihood is found by adding the contributions from the single sets.

29.2 The conditional likelihood for 1:1 matched sets

First we derive the contribution for case-control studies with one case and one control in each set. The possible case or control status for any two subjects are represented as a probability tree in Fig. 29.1. Using the relationship between odds and probability, the probabilities that subject 1 is a case or a control are $\omega_1/(1+\omega_1)$ and $1/(1+\omega_1)$ respectively. Similarly, the probabilities for subject 2 are $\omega_2/(1+\omega_2)$ and $1/(1+\omega_2)$. The probabilities of the outcomes for the pair of subjects are obtained by multiplying along branches of the tree in the usual way. The last column of the figure shows such probabilities, after writing

$$\omega_1 = \omega_C\theta_1, \qquad \omega_2 = \omega_C\theta_2,$$

and

$$K = \frac{1}{1+\omega_1} \times \frac{1}{1+\omega_2}.$$

These probabilities refer to any two subjects from the study base. Conditional on the fact that one of the subjects is a case and the other is a control, the probability that subject 1 is the case is

$$\frac{K\omega_C\theta_1}{K\omega_C\theta_1 + K\omega_C\theta_2} = \frac{\theta_1}{\theta_1+\theta_2}.$$

and the probability that subject 2 is the case is

$$\theta_2/(\theta_1+\theta_2).$$

The contribution to the log likelihood of the case-control set is, therefore

$$\log\left(\frac{\theta_{(\text{for case})}}{\theta_{(\text{for case})} + \theta_{(\text{for control})}}\right).$$

This way of writing the log likelihood makes it clear that it does not depend on the arbitrary numbering of the subjects in the pair but only on the expressions for θ in terms of A(1), A(2) and B(1), the parameters to be estimated. The total log likelihood thus depends only on A(1), A(2), and B(1), and the nuisance parameters ω_C^t have been eliminated.

Exercise 29.1. Table 29.1 shows the data for the first two case-control sets in a 1:1 matched study. The set variable indicates which set each subject belongs to, and case or control status is indicated using a variable taking the value 1 for cases and 0 for controls. Illustrative parameter values for the multiplicative effects of the explanatory variables age and exposure, where age has three levels ($< 55, 55-64, 65-74$) and exposure has two levels, are shown below.

Parameter	Value
Age (1)	×1.5
Age (2)	×3.0
Exposure (1)	×5.0

The corner is defined as unexposed and age < 55. Calculate the values of θ predicted by the model for these four subjects. Calculate the log likelihood contributions for the two sets.

Before leaving the 1:1 case we shall verify that the method of obtaining the log likelihood described above gives the same answer as the method described in Chapter 19, for a binary exposure. The model is now

$$\text{Odds} = \boxed{\text{Corner} \times \text{Set}} \times \boxed{\text{Exposure}}$$

which has only one parameter, Exposure(1), apart from the nuisance parameters. This parameter is the multiplicative effect of exposure and we shall refer to it as ϕ. The values of θ for the case and control are determined

Table 29.1. Data file for a 1:1 matched case-control study

Subject	Set	Case/control	Age	Exposure
1	1	1	48	1
2	1	0	64	0
3	2	1	52	1
4	2	0	70	1
...				

Table 29.2. Likelihood contributions for the 1:1 matched study

Exposure	θ for case	θ for control	Likelihood
Neither	1	1	$1/(1+1) = 1/2$
Both	ϕ	ϕ	$\phi/(\phi+\phi) = 1/2$
Case only	ϕ	1	$\phi/(\phi+1)$
Control only	1	ϕ	$1/(1+\phi)$

by whether or not they were exposed. For example, if the case was not exposed then $\theta = 1$, while if the case was exposed then $\theta = \phi$. Similarly for the control. Table 29.2 sets out the four possible outcomes for each case-control set and the corresponding contributions to the log likelihood. The first two outcomes, in which the exposure status of case and control is the same, lead to log likelihood contributions which do not depend upon the parameter, and can be ignored. If N_1 and N_2 are the frequency of occurrence of the remaining outcomes, the total log likelihood is

$$N_1 \log\left(\frac{\phi}{1+\phi}\right) + N_2 \log\left(\frac{1}{1+\phi}\right)$$

which is the same as we obtained in Chapter 19, except that here we have called the effect ϕ rather than θ to avoid confusion.

29.3 The conditional likelihood for 1: m matched sets

We now extend the above argument to sets with one case and m controls. If the sampling had not been carried out deliberately so as to obtain a single case and m controls in the set, the probability that subject 1 is a case and the remaining m subjects are controls would be

$$\frac{\omega_1}{1+\omega_1} \times \frac{1}{1+\omega_2} \times \frac{1}{1+\omega_3} \times \cdots,$$

and making the substitutions

$$\begin{aligned} \omega_i &= \omega_C \theta_i \\ K &= \frac{1}{1+\omega_1} \times \frac{1}{1+\omega_2} \times \frac{1}{1+\omega_3} \times \cdots \end{aligned}$$

this may be written as $K\omega_C\theta_1$. Similarly, the probability that subject 2 is a case and all other subjects controls is $K\omega_C\theta_2$, and so on. The sum of probabilities for all the outcomes in which one member of the set is a case and all other members are controls is

$$K\omega_C(\theta_1 + \theta_2 + \theta_3 + \cdots)$$

so that the conditional probability that subject 1 is the case is:

$$\frac{K\omega_C\theta_1}{K\omega_C(\theta_1 + \theta_2 + \theta_3 + \cdots)} = \frac{\theta_1}{\theta_1 + \theta_2 + \theta_3 + \cdots}.$$

The contribution of one set to the log likelihood is, therefore,

$$\log\left(\theta_{(\text{for case})} \Big/ \sum_{\text{Case-control set}} \theta\right).$$

The total log likelihood is obtained by adding the contributions for all case-control sets.

From the form of this log likelihood it is clear that the conditional approach does not allow estimation of multiplicative effects of variables used in matching. Since all subjects in the set share the same value for such a variable its multiplicative effect will cancel out in the ratio of θ for the case to the sum of all θ's in the case-control set. However, interaction terms involving matching variables *can* be fitted. For example, for a case-control study in which sex was one of the matching variables, the sex effect cannot be estimated but the parameters for interaction between sex and exposure can be, because they will not occur in all of the θ's from the same case-control set.

29.4 Sets containing more than one case ★

The conditional argument can be generalized quite easily to allow for case-control sets containing more than one case, although the computation of the log likelihood may become rather lengthy. The idea is illustrated for a set containing two cases and one control. Fig. 29.2 shows the probability tree for case/control status of a set of three subjects. In three of the eight possible outcomes there are two cases and one control. The probabilities for these branches are written to the right of the figure, again using the

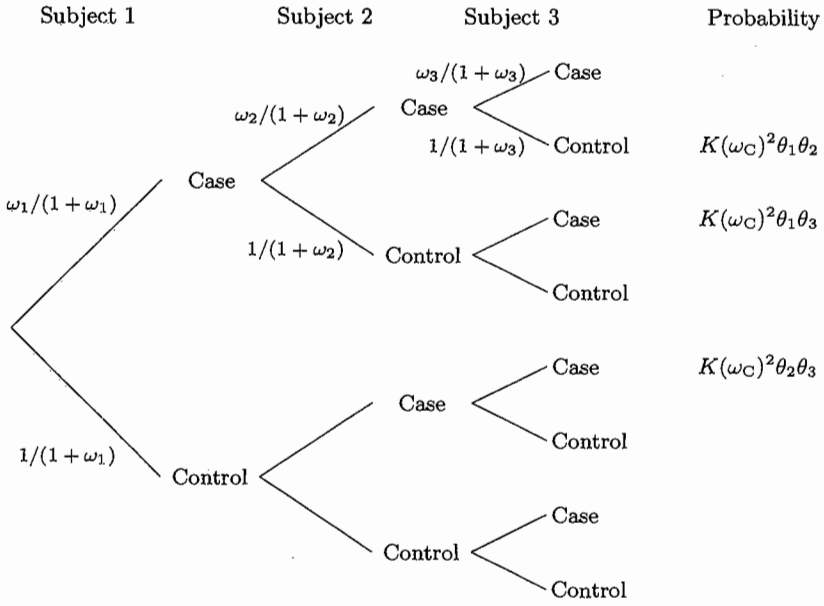

Fig. 29.2. Sets with two cases and one control.

abbreviation

$$K = \frac{1}{1+\omega_1} \times \frac{1}{1+\omega_2} \times \frac{1}{1+\omega_3}.$$

Conditional on the observed outcome being one of the three with two cases and one control the probability that the cases are subjects 1 and 2 is

$$\frac{K(\omega_C)^2\theta_1\theta_2}{K(\omega_C)^2\theta_1\theta_2 + K(\omega_C)^2\theta_1\theta_3 + K(\omega_C)^2\theta_2\theta_3} = \frac{\theta_1\theta_2}{\theta_1\theta_2 + \theta_1\theta_3 + \theta_2\theta_3}.$$

The log of this conditional probability is the contribution of the set to the log likelihood.

It is easy to see how this argument can be extended to deal with any number of cases and controls in a set. For example, for sets of size 6 containing 3 cases, the conditional probability that subjects 1, 2, and 3 are the cases is

$$\frac{\theta_1\theta_2\theta_3}{\theta_1\theta_2\theta_3 + \theta_1\theta_2\theta_4 + \theta_1\theta_2\theta_5 + \cdots}.$$

The denominator contains a term for each of the 20 ways of selecting three subjects from 6, and does not depend on the way the subjects have been numbered.

Solutions to the exercises

29.1 The values of θ for the four subjects are:

		Multiplicative effects		
Subject	Corner	Age	Exposure	θ
1	1.0		×5.0	5.0
2	1.0	×1.5		1.5
3	1.0		×5.0	5.0
4	1.0	×3.0	×5.0	15.0

Subject 1 is the case in the first set and subject 3 is the case in the second set. The log likelihood contributions are, therefore

$$\log\left(\frac{5.0}{5.0+1.5}\right) + \log\left(\frac{5.0}{5.0+15.0}\right) = -0.262 - 1.386.$$

Preamble: C&H motivate this chapter by noting that with individually matched case-control studies, one cannot add a separate 'intercept' for each matched set or 'stratum'. The best example of the danger of doing this (and thus overfitting) is the example of matched pairs: JH's notes for Ch 19 contains the relevant excerpt from Breslow and Day Vol. I section 7.1. p 251 where they use a worked example to show that the 'unconditional' (and close to saturated) model yields a $\hat{\beta}$ value that is twice the value of the one obtained when the individual intercepts are conditioned out.

As is seen in the Oscar Predictions article by Pardoe (see the Resources for conditional logistic regression) conditional logistic models are also useful for predictions: they are not limited to 'case-control' and other 'outcome-based' sampling schemes. But the likelihood can be viewed as having been constructed '*after the fact*': it involves the probability of observing what we *did* (*already*) observe. So it has a certain 'in retrospect' aspect to it. In the case of the Oscar data, we can do the 5 probability calculations (one per nominee, each a function of the β and the particular nominee's covariates) ahead of time, but we need to wait until the winner is declared before we select the one associated with that winner as the actual likelihood contribution.

29.1 The logistic model

The '***corner***' terminology was introduced in the (excellent) general chapter 22 (introduction to regression models). It starts at the 'point of departure' or 'corner', where $x_1 = 0, x_2 = 0, \ldots$ and places the intercept there, then works outwards from this corner as it goes to non-zero values of the x's.

The point made in section 29.1 is that we could have a different intercept for each stratum or matched set, but as we know, it is dangerous to have so many fitted parameters when the amounts of data are small. So section 29.2 conditions out these intercepts.

Incidentally, in Fig 29.1 in the theoretical development, one can replace the words "case" and 'control" by "winner" and "non-winner" without loss of generality.

29.2 Conditional likelihood for 1:1 matched sets

Here again, we do not have to limit ourselves to *case-control* pairs. Imagine twins born to an HIV infected mother. Prospectively, what are the chances they will become HIV positive? and does it depend on which is born first and thus spends more time in the birth canal?

Or think of the prospective vasectomy-MI example.

Or think about the one winner in each US presidential election, where for each of the 2 candidates, X_1 might be age, X_2 height, X_3 millions of dollars spent by the candidate's party, etc.

In the diagram in the notes on Chapter 19, we have already derived and shown the binomial likelihood (and the fitting via GLM) for the 1:k matched sets situation, but where there was only one covariate (e.g. breast cancer screening). For matched pairs, think of just one 'diagonal' , i.e. containing the discordant pairs.

The only difference here in 29.2 is that the covariate is no longer necessarily a scalar. To make it more general, think of $\theta_1 = e^{\boldsymbol{\beta}\mathbf{x}_1}$ & $\theta_2 = e^{\boldsymbol{\beta}\mathbf{x}_2}$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors.

The probability that subject 1 (subject 2) is the case is

$$\frac{\theta_1}{\theta_1+\theta_2} \quad \left(\frac{\theta_2}{\theta_1+\theta_2}\right).$$

You will recognize these as two Bernoulli's (only one of which can come to pass) where

$$n = 1; \quad \text{and either } \pi = \frac{\theta_1}{\theta_1+\theta_2} \text{ or } \pi = \frac{\theta_2}{\theta_1+\theta_2}$$

depending on which subject you focus on: after the fact, you will know which one represents the case.

29.3 Conditional likelihood for 1:m matched sets

You can think of each θ as the (relative) number of tickets that that person holds in a lottery or contest, where there can only be one winner (if the context is the Oscars) or one loser (in the 4 women per matched set in the breast cancer screening example)

In 'riskset' or 'candidate set' i, with candidates $j = 1, \ldots, n_j$ and associated covariate vectors x_{ij} (with subscript i suppressed but implicit)

$$\theta_j \propto \exp[\beta x_j]$$

Thus the likelihood contribution from the riskset is

$$\text{Likelihood}(\beta) = \frac{\exp[\beta x_{case}]}{\sum_j \exp[\beta x_j]}$$

so the log-likelihood contribution is

$$\text{Log Likelihood}(\beta) = LL(\beta) = \beta x_{case} - \log\left[\sum_j \exp[\beta x_j]\right].$$

Take the easiest case, where x_j and β are scalars. The first derivative is

$$LL'(\beta) = x_{case} - \frac{\sum_j x_j \exp[\beta x_j]}{\sum_j \exp[\beta x_j]} = x_{case} - \overline{x}_{weighted},$$

where $\overline{x}_{weighted}$ is a weighed mean of $\{x_1, \ldots, x_n\}$, with weights $w_1 = \exp[\beta x_1]$ to $w_n = \exp[\beta x_n]$.

The second derivative is

$$LL''(\beta) = -\left[\frac{\sum_j x_j^2 w_j}{\sum_j w_j} - \left\{\frac{\sum_j x_j w_j}{\sum_j w_j}\right\}^2\right] = -Var[x]_{weighted}.$$

The form makes intuitive sense: the larger the spread of $\{x_1, \ldots, x_n\}$, the larger the information about β.

Estimation of β

By setting $LL'(\beta) = 0$, we get the **estimating equation**

$$\sum_{sets} x_{case} = \sum_{sets} \overline{x}_{weighted},$$

but since β is involved in a complex way in the right hand side, there is no obvious way to isolate it. (Contrast this with the estimating equation in the case of the 'one β' binomial likelihood obtained by conditioning on the sum of two (stratum-specific) Poisson random variables when the two denominators are known– the one where the first iteration leads to the Mante-Haenszel estimator, and where subsequent ratios, used as estimators of the rata ratio, involve the reciprocal of $[pt_0 + RR \times pt_1]$).

Newton-Raphson iteration

When JH last taugh this material, in 2011, his students had not yet introduced him to the `optimize` and `optim` functions. So he relied on the Newton-Raphson algorithm to find the $\hat{\beta}_{ML}$. It doesn't hurt to know this algorithm, as one is much more in control with it than one is with the `optimize` and `optim` functions.

In the case of a scalar β, we have

$$\beta_{new} = \beta_{prev} - \left.\frac{\sum_{sets} LL'(\beta)}{\sum_{sets} LL''(\beta)}\right|_{\beta=\beta_{prev}}.$$

In the case of a vector $\underline{\beta}$ of length p, so that $LL'(\underline{\beta})$ is a vector of length p and $LL''(\underline{\beta})$ is a square (and symmetric) matrix of size $p \times p$, we have

$$\underline{\beta}_{new} = \underline{\beta}_{prev} - \left[\sum_{sets} LL''(\underline{\beta})\right]^{-1} \sum_{sets} LL'(\underline{\beta})\Big|_{\underline{\beta}=\underline{\beta}_{prev}}.$$

At convergence, we can use $-\left[\sum_{sets} LL''(\underline{\beta})\right]^{-1}$ as the variance-covariance matrix for $\hat{\underline{\beta}}_{ML}$.

Supplementary Exercise 29.1

The above derivation of $LL''(\beta)$ in the case of 1:m matched sets omitted some steps. Show the derivation step by step.

Supplementary Exercise 29.2

Refer again to the 1:3 matched sets in the study of breast cancer mortality and screening; results are tabulated in C&H Table 19.2).

i. Using the log-likelihood set out in the middle of C&H page 295, set up the overall log-likelihood (from all 46 matched sets in Table 19.2) in an `R` function, and use `optimize` to find the ML of the Rate Ratio.

   Hint: if the 4 observations are in the order: case, control$_1$, control$_2$ and control$_3$, you can think of the 4 θ's in a matched set where the case and the 1st and 3rd controls were screened as

   $$e^{\beta\times 1},\ e^{\beta\times 1},\ e^{\beta\times 0},\ e^{\beta\times 1},$$

   and the 4 θ's in a set where only and the 2nd control was screened as

   $$e^{\beta\times 0},\ e^{\beta\times 0},\ e^{\beta\times 1},\ e^{\beta\times 0}.$$

ii. Using the Newton-Raphson procedure with the same log-likelihood, find the MLE of the Rate Ratio, along with an estimate of the variance of the log of the estimate.

iii. What contributions to the log-likelihood are made by the 11 matched sets in the bottom left of Table 19.2? the 1 set on the top right?

Supplementary Exercise 29.3

In section 15.3, C&H show the form of the 'almost ML' (M-H) estimator of a common rate ratio when the population-time denominators are known. They also show how one could use this to calculate new weights, and to continue to re-weight until the process converges to $\hat{\theta}_{ML}$. In his notes on 15.3, JH shows another way to derive the final (fixed point) version.

Consider the setup in section 19.4, and – in the notes – the diagram with the diagonals, where the 1:m matched data can be seen as binomials with different offsets i.e., different probabilities modulated by the margins N_1 and N_0.

Starting from the single estimating equation linking the expected binomial splits with the expected splits, derive an estimating equation for θ.

Hint: For example, for the data in the 3 rows in Table 19.3, this is equivalent to starting with $g + e + c = \hat{g} + \hat{e} + \hat{c}$, substituting in $\hat{\theta}_{ML}$ into each of the 3 fitted counts, and re-arranging terms until $\hat{\theta}_{ML}$ is alone on the left had side. (Of course, it is also on the right hand side, but one can start with a trial value and iteratively re-weight until one reaches equilibrium.)

More generally, avoiding Table 19.3 notation, focus on the number of sets in each diagonal in JH's diagram; write y for the number of sets in the upper left of a diagonal strip, and n for the total number of sets in the the strip (e.g., $y = \{1, 4, 3\},\ n = \{11, 16, 7\}$, while $N_1 = \{1, 2, 3\},\ N_0 = \{3, 2, 1\}$.)
Set $\sum y = \sum n \times \frac{N_1 \hat{\theta}_{ML}}{N_1 \hat{\theta}_{ML} + N_0}$, then isolate $\hat{\theta}_{ML}$ to the left hand side.

Supplementary Exercise 29.4

What, if anything, in your R code for 29.2 would need to be changed if you were to fit a conditional logistic regression to find what weights (β's) should be given to the variables collected by Pardoe so as to predict which woman would win the Oscar award for best actress?

Supplementary Exercise 29.5

Refer to the Walker article, R code, and paired data (in 'wide' & 'tall' formats) on the effect of vasectomy on the risk of a myocardial infarction (MI). (Resources Condn'l Regr'n.) Cf. also Hanley J. Survival analysis; risk sets; case control studies: a unified view of some epidemiologic data-analyses (Regression Models and Lifetables).

Use the R code provided to (i) by trial and error balance the 3 estimating equations with respect to $\hat{\beta}_{ML}$ (ii) carry out the Newton-Raphson iteration and arrive at the same $\hat{\beta}_{ML}$. Check your answers (point estimate and variance) against those produced by the `clogit` and `coxph` functions in the `R` `survival` package.

Supplementary Exercise 29.6

Refer to the articles and data on the role of stimulation (rocking) in the delay of onset of crying in the newborn infant.

i. Fit a (stratified-by-day) proportional hazards model using various ways of handling the 'ties'.
See `http://www.medicine.mcgill.ca/epidemiology/hanley/c681/cox/TiesCoxModelR.txt`.

ii. Which method does `clogit` use? Verify this by doing the likelihood calculation 'from scratch' (there are just a few days where it is an issue, and the likelihood involves just a scalar parameter).

iii. The experiment was carried out for 18 days between 25th May and 24th August. The article mentions the temperature for some of the days. Use an imputed (interpolated) value for each of the others and assess the effect of adding temperature to the model.

Supplementary Exercise 29.7

An Analysis of Contaminated Well Water and Health Effects in Woburn, Massachusetts

S. W. LAGAKOS, B. J. WESSEN, and M. ZELEN*

In 1979, two of the eight municipal wells servicing Woburn, Massachusetts, were discovered to be contaminated with several chlorinated organics. Shortly afterwards, the town was found to have an elevated rate of childhood leukemia. Using recent information about the space–time distribution of water from the two contaminated wells, we find positive statistical associations between access to this water and the incidence rates of childhood leukemia, perinatal deaths (1970–1982), two of five categories of congenital anomalies, and two of nine categories of childhood disorders. We find no associations with spontaneous abortions, low birth weight, or the other categories of congenital anomalies and childhood disorders. This article discussed these results and other features of the data relevant to their interpretation.

KEY WORDS: Environmental exposure; Health survey; Observational study; Proportional hazards model; Time-dependent covariate.

1. INTRODUCTION

near Horn Pond in southwest Woburn (Fig. 1), were tested and found to meet both state and federal drinking-water standards.

Independently, during site excavations in July 1979 for an industrial complex located north of wells G and H (Fig. 1), large pits of buried animal hides and chemical wastes were discovered. A nearby abandoned lagoon was found to be heavily contaminated with lead, arsenic, and other metals. Subsequently, the groundwater under eastern Woburn was sampled at 61 test wells and found to contain 48 EPA priority pollutants and raised levels of 22 metals (Ecology and Environment, Inc. 1982).

The closing of the two municipal wells and the discovery of the abandoned waste sites occurred at about the same time as the Love Canal incident and alerted Woburn residents to possible health hazards. One resident contacted the Centers for Disease Control (CDC), asking if cancer rates were elevated in Woburn. A citizens' group formed and, in late 1979, produced a list of children diagnosed with leukemia.

JASA '86; [full text in link2]. This study led to the lawsuit that formed the basis for the movie *A Civil Action* (`http://www.imdb.com/title/tt0120633/`).

Table 2 on next page is from the same report, as are sections 3.1 & 4.1. See also Hanley J. Survival analysis; risk sets; case control studies: a unified view of some epidemiologic data-analyses (Resources for Regression Models and Lifetables).

Table 1. Annual G and H Exposure Scores by Zone

Year	1960–1969 Zones 1	2	3	4
1960–1963	0	0	0	0
1964	.23	.05	0	0
1965	.08	.08	.08	.08
1966	.95	.70	.25	.12
1967	.51	.47	.32	.17
1968	.72	.34	0	0
1969	.75	.40	0	0
Year	**1970–1982 Zones A**	**B**	**C**	**D**
1970	.54	.27	.03	0
1971	.46	.38	.08	0
1972	.29	.29	.14	0
1973	0	0	0	0
1974	.37	.32	.25	.14
1975	.55	.44	.13	.01
1976	.49	.38	.09	0
1977	.94	.52	.04	.01
1978	1.00	.88	.61	.05
1979	.39	.39	.29	0
1980–1982	0	0	0	0

NOTE: Table entries give estimated fraction of residential water supply derived from wells G and H by year and residential zone. Refer to Figure 1 for zonal definitions. Exposure scores are zero for Zones 5 and E in all years.

other data and assigned to each pregnancy the annual exposure score corresponding to the mother's residence in the year the pregnancy ended. We also determined for each child an exposure "history," consisting of his or her set of annual exposure scores, beginning from the first year of Woburn residency. For example, an exposure of .49 would be assigned to a pregnancy ending in 1976 for a mother residing in Zone A. Similarly, a child born in 1967 and residing in the intersection of Zones 1 and B for the first 4 years of life would generate cumulative exposures of .51, 1.23, 1.98, and 2.25 during this period. If a child changed residences, we arbitrarily defined his or her exposure score for that year to be the score corresponding to the former residence.

3. STATISTICAL METHODS

3.1 Childhood Leukemia

Based on national rates (SEER 1981), the 20 childhood leukemia cases observed between 1964 and 1983 are significantly higher than expected ($E = 9.1$, $P = .001$). To determine whether the space–time distribution of these cases within Woburn is correlated with water from wells G and H, we used the failure time regression model (Cox 1972)

$$h\{t \mid x(t), y\} = h_y(t)\exp\{\alpha x(t)\}, \tag{1}$$

where $x(t)$ is some expression of G and H exposure history from birth to age t, y is the year of birth, $h\{t \mid x(t), y\}$ is the leukemia risk (hazard function) at age t for an individual born in year y and with exposure $x(t)$, and $h_y(t)$ is the baseline Woburn risk at age t for an unexposed person born in year y. With this model, the relative risk at age t for someone with exposure $x(t)$, relative to an unexposed individual born in the same year, is $\exp\{\alpha x(t)\}$. The hypothesis of no association is given by $\alpha = 0$.

We used two exposure metrics $x(t)$: (a) cumulative G and H exposure from birth until age t and (b) a binary indicator of whether there had been any G and H exposure by age t. With the first of these measures, risk increases steadily with cumulative exposure. With the latter, an individual's hazard function jumps from $h_y(t)$ to $h_y(t)\exp(\alpha)$ upon exposure. Partial-likelihood-based tests of $\alpha = 0$ from this "none versus some" exposure model are closely related to the variation of the log-rank test proposed by Mantel and Byar (1974) (see also Aitkin, Laird, and Francis 1983; Crowley and Hu 1977). Misspecification of the form of $x(t)$ with either test results in a loss of efficiency but not in a distortion of size.

The partial likelihood score test for $\alpha = 0$ can be expressed in the form (see Kalbfleisch and Prentice 1980)

$$\sum (X_i - E_i)/(\sum V_i)^{1/2}, \tag{2}$$

where the sum is over the leukemia cases and X_i is the observed value of $x(t)$ for the ith case at t_i, the age of diagnosis. The quantities E_i and V_i are the average and variance of the $x(t_i)$ for the "risk set" of children born in the same year as the ith case and not diagnosed with leukemia before age t_i and represent the null mean and variance of X_i, conditional on this risk set. When $\alpha = 0$, the distribution of (2) is approximately $N(0, 1)$. A closed-form approximation to the maximum likelihood estimator (MLE) of α is given by $\sum (X_i - E_i)/\sum V_i$. This will closely approximate the MLE for α close to 0 but may be conservative for large $|\alpha|$.

In our situation we had the exposure histories for all of the leukemia cases. We had only the exposures, however, for those noncases identified in the sample survey. Accordingly, E_i and V_i could not be computed directly and were estimated. One approach is to adopt a form of risk-set sampling and estimate E_i and V_i from the survey data (see Breslow, Lubin, Marek, and Langholtz 1983; Cox and Oakes 1984; Liddell, McDonald, and Thomas 1977; Prentice 1985); that is, for each case we identified all surveyed children who were born in the same year and were residents at the same time as the case and then computed the average and variance of their $x(t)$ values for the period of residency of the case. The results of this approach are presented in detail in Section 4.1. Alternatively, since each individual's exposure history is uniquely determined by his or her residence history, E_i and V_i can be estimated from the population distributions of the G and H exposure zones. Details of this approach are given in the Appendix.

When $x(t)$ is positively associated with the risk of leukemia, the number of leukemia cases that are statistically "explained" by the association is

$$\Sigma_1 [h(t_i \mid X_i, y) - h_y(t_i)]/h(t_i \mid X_i, y) = \Sigma_1 [1 - \exp(-\alpha X_i)],$$

where the sum is over cases (see National Research Council 1985).

Table 2. Observed and Expected Exposures to Wells G and H for 20 Childhood Leukemia Cases

Case	Year of diagnosis	Year of birth	Period of residency	Observed cumulative exposure	Size of risk set sample	Expected cumulative exposure (var)	Proportion of risk set exposed
1	1966	1959	1959–1966	1.26	218	.31 (.26)	.33
2	1969	1957	1968–1969	0	290	.34 (.36)	.26
3	1969	1964	1969	.75	265	.17 (.10)	.25
4	1972	1965	1965–1972	4.30	182	.90 (2.23)	.36
5	1972	1968	1968–1972	2.76	183	.58 (.88)	.32
6	1973	1970	1970–1973	.94	170	.20 (.20)	.19
7	1974	1965	1968–1974	0	213	.56 (1.04)	.29
8	1975	1964	1965–1975	0	239	.99 (2.78)	.38
9	1975	1975	1975	0	115	.09 (.03)	.25
10	1976	1963	1963–1976	.37	219	1.18 (3.87)	.40
11	1976	1972	1972–1976	0	132	.24 (.32)	.18
12	1978	1963	1963–1978	7.88	219	1.41 (6.23)	.40
13	1979	1969	1969–1979	2.41	164	.73 (2.56)	.31
14	1980	1966	1966–1980	0	199	1.38 (6.00)	.39
15	1981	1968	1968–1981	0	187	1.14 (4.20)	.35
16	1982	1979	1979–1982	.39	154	.08 (.02)	.23
17	1983	1974	1974–77, 1980–83	0	84	.25 (.45)	.23
18	1982	1981	1981–1983	0	—	0 (0)	0
19	1983	1980	1980–1982	0	—	0 (0)	0
20	1983	1980	1981–1983	0	—	0 (0)	0
Totals				21.06		10.55 (31.52)	5.12
Score test statistic:						1.87	2.08
Significance level:						$P = .03$	$P = .02$

NOTE: Risk set for a case consists of children born in the same year as the case and who were residents of Woburn when the case was. Variance of proportion, say p, of risk set exposed equals $p(1 - p)$. Cases 18–20 do not contribute to the test statistic because birth occurred after closure of wells G and H.

4. RESULTS 4.1 Childhood Leukemia Using either the cumulative ($P = .03, \hat{\alpha} = .33$) or none versus some ($P = 0.02, \hat{\alpha} = 1.11$) exposure metrics, there is a positive association between G and H exposure and the incidence rate of childhood leukemia. Table 2 gives information on the 20 cases and their contributions to the score tests of $\alpha = 0$. For example, case 13 was born in 1969 and resided in Woburn until being diagnosed in 1979, at which time his cumulative G and H exposure was 2.41. The corresponding 164 surveyed children who were born in Woburn in 1969 and resided there until 1979 had an average cumulative G and H exposure of 0.73, and 31% of these were exposed. Overall, given the years of birth and periods of Woburn residence of the 20 cases, the expected number exposed to wells G and H when $\alpha = 0$ is $\sum_i E_i = 5.1$, compared with 9 observed, and the sum of their expected cumulative exposures is 10.6, compared with an observed number of 21.1. The alternative approach (see the Appendix) based on estimating E_i and V_i from the regional distribution of the population gave very similar results for both the cumulative ($\sum E_i = 11.0, P = .04, \alpha = .29$) and none versus some($\sum E_i = 4.9, P = .02, \alpha = 1.22$) exposure metrics. Although both G and H exposure metrics are associated with leukemia risk, there are two few cases to be confident of which, if either, best describes this relationship. Furthermore, it does not logically follow that Woburn's entire leukemia excess, based on national rates, is explainable by these associations. Indeed, the cumulative and none versus some metrics of G and H exposure statistically explain about 4 and 6 leukemia cases, respectively, whereas national rates suggest a townwide excess of about 11 cases between 1964 and 1983 (Sec. 3.1). We return to this point in Section 6.

i. Using a Mantel-Haenszel test and a Mantel-Haenszel summary rate ratio, and each row as a 'stratum' or riskset, derive a P-value and an $\hat{\alpha}$ corresponding to those underlined in line 2 of section 4.1. Then Check them against the results of fitting a conditional logistic regression.

ii. How much would point and interval estimates of α change if instead of risksets of the sizes used, they had used risksets of (say) size 11 (10 plus case)?