C&H motivate this chapter by noting that 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 those overfitting) is the example of matched pairs: see Breslow and Day Vol. I section 71. p 251 for a worked example where 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, 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 (each a function of β) ahead of time, 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

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-contro*l pairs. Imagine twins born to an HIV infected mother. 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?

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.

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

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

so the log-likelihood contribution is

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

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

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

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

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

Supplementary Exercise 29.1 The above derivation of $LL''(\beta)$ omitted some steps. Show the derivation step by step.

Supplementary Exercise 29.2 Refer (under Resources) to the `R` code used to select a subset of variables for a selected subset (female performers) of the Oscars.csv file provided by Pardoe.

i. Focus first on your choice of *one* of the extracted variables {Age, FN, FP, FPrN, Gf1, Gf2}. Find the corresponding (*scalar*) $\hat{\beta}_{ML}$ from first principles, using (i) trial and error to balance the estimating equation (ii) Newton-Raphson iteration. Check your answers (point estimate and variance) against those produced by the `clogit` and `coxph` functions in the `R survival` package.

ii. Scale up the Newton-Raphson procedure to find the $\hat{\beta}_{ML}$ *vector* associated with the 6 extracted variables {Age, FN, FP, FPrN, Gf1, Gf2}. Again, check your answers against those produced by the `clogit` and `coxph` functions.

iii. Explain how Pardoe dealt with the fact that the two variables Gf1 and Gf2 have no meaning before 1943. If he has changed the values he assigned to these two variables in the early competitions to some other (non-zero) value, would it have changed the fitted $\hat{\beta}_{ML}$?

Supplementary Exercise 29.3 Refer to the Walker article, R code, and data (in 'wide' and 'tall' formats) on the effect of vasectomy on the risk of a myocardial infacrction (MI).

i. 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.

ii. When each 'risk set' or 'matched set' consists of just 2 candidates, it is possible to use regular (unconditional) logistic regression to fit the model (cf Brelsow and Day Vol I p253). To see this, consider a riskset where the $\underline{x}$ vectors associated with the 2 candidates are arbitrarily labeled $\underline{x}_1$ and $\underline{x}_2$. Consider first a riskset where the event happened to candidate '1'. Then the likelihood contribution is

$$\frac{exp[x_1\beta]}{exp[x_1\beta]+exp[x_2\beta]} = \frac{exp[(x_1-x_2)\beta]}{exp[(x_1-x_2)]+1}.$$

Consider first a riskset where the event happened to candidate '2'. Then the likelihood contribution is

$$\frac{exp[x_2\beta]}{exp[x_1\beta]+exp[x_2\beta]} = \frac{exp[(x_2-x_1)\beta]}{exp[(x_2-x_1)]+1}.$$

These have the same form as the contributions from realizations of *Bernoulli* responses. Verify that we could instead fit the conditional logistic model by regressing the 36 y's on the $\underline{d}$'s in a non-intercept logistic regression model, where for each observation (each riskset), either

(a) setting the Bernoulli response y to '1' and using as a predictor vector $\underline{d} = \underline{x}_{case} - \underline{x}_{non.case}$ or

(b) setting the Bernoulli response y to '1' if candidate 1 represents the case or 0 if candidate 2 does, and using as a predictor vector $\underline{d} = \underline{x}_1 - \underline{x}_2$.

iii. Compare the fitted coefficients of the conditional logistic regression model with those from an unconditional model where the pair matching is ignored and the observations are treated as 72 independent (but not identical!) Bernoulli observations.

Supplementary Exercise 29.4 Refere to the aricles 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.biostat.mcgill.ca/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 value for each of the others and assess the effect of adding temperature to the model.

Supplementary Exercise 29.5 The following table is from the article Analysis of Contaminated Well Water and Health Effects in Woburn, Mass [Lagakos, Wessen, Zelen; JASA '86] and involves one of the lawsuits in the movie *A Civil Action* (`http://www.imdb.com/title/tt0120633/`).

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.

i. Use conditional logistic regression to replicate the calculations involving the binary exposure metric.

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

Vasectomy and Health

Results From a Large Cohort Study

Frank J. Massey, Jr, PhD; Gerald S. Bernstein, PhD, MD; William M. O'Fallon, PhD; Leonard M. Schuman, MD, MS; Anne H. Coulson, Ruth Crozier, MA; Jack S. Mandel, PhD; Robert B. Benjamin, MD; Heinz W. Berendes, MD, MHS; Potter C. Chang, PhD; Roger Detels, MD, MS; Richard F. Emslander, MD; James Korelitz, PhD; Leonard T. Kurland, MD, DrPH; Irwin H. Lepow, MD, PhD; Douglas D. McGregor, MD, DPhil; Robert N. Nakamura, PhD; Jose Quiroga, MD; Stanwood Schmidt, MD; Gary H. Spivey, MD, MPH; Timothy Sullivan, MD

• In this historical cohort study we identified, located, and, if living, interviewed 10,590 vasectomized men from four cities, along with a paired neighborhood control for each. The times between procedure data and interview or death ranged from under one to 41 years, with median equal to 7.9 years and with 2,318 pairs having ten or more years of follow-up. Participant reports of diseases or conditions that might possibly be related to vasectomy through an immunopathological mechanism were validated by direct contact with physicians and review of medical records. Results of this study do not support the suggestions of immunopathological consequences of vasectomy within the period of follow-up. Except for epididymitis-orchitis, the incidence of diseases for vasectomized men was similar or lower than for their paired controls.

(*JAMA* 1984;252:1023-1029)

ALTHOUGH vasectomy has been considered a safe method of birth control, questions have been raised about the possibility of adverse effects related to an immune response.[1] One half to two thirds of vasectomized men are reported to experience the development of antisperm antibodies,[1] which may persist for ten years.[2] Men with antisperm antibodies may be at increased risk for the development of immunologically mediated diseases. Furthermore, immune complex orchitis,[3] glomerulonephritis,[3] and exacerbated atherosclerosis[4,5] have been demonstrated in vasectomized animals. This historical cohort study of 10,590 pairs of men was initiated to determine whether vasectomized men experience an increase in morbidity or mortality from any of a wide range of diseases.

See also p 1005.

SUBJECTS AND METHODS

The study population was derived from existing clinic and physicians' records of men who had been vasectomized for contraceptive reasons in four cities, Minneapolis and Rochester, Minn, and Los Angeles, and Eureka, Calif. All of the surgeries were performed before Jan 1, 1976, most of them after 1965.

For each vasectomized man, a "matching" nonvasectomized man was sought according to a prescribed protocol (F.J.M., et al, unpublished data, 1984). This was to be a man of approximately the same age, the same race, and marital status (ever married, or never married) and living in the same neighborhood as the vasectomized man at the time of the vasectomy. The date of vasectomy served as a baseline date for subsequent health events for both members of the pair.

Potential participants were traced to their current locations and, if they agreed, were interviewed once during the period 1977 to 1982. Deceased vasectomized men were included in the study and if, for any vasectomized man, the first matching man had died, he was included as the match. If a potential match was found to be ineligible or could not be located, he was replaced by the next eligible man, living or dead, whereas, if a potential match refused to participate, he was replaced with the next living eligible man. No surrogate interviews were performed and only data from death certificates were available for deceased subjects.

Trained interviewers, who were not informed of the vasectomy status of their assigned interviewees, administered a comprehensive health questionnaire. In addition to a large number of demographic and health-related questions, the questionnaire included questions as to whether the participants had been informed of any diagnosis from a comprehensive list of 98 diseases among which were 54 diseases identified as of special interest in relation to possible effects of vasectomy.

To enhance the potential for detection of any vasectomy effect, in the analysis diseases and conditions of possible immunologic origin were grouped according to eight postulated mechanisms. The eight groups and diseases included in each are given in Table 1.

Attempts were made to validate each report of the occurrence of any of the 54 diseases of special interest after the age of 18 years. Ninety-seven percent of the participants gave written permission for such validation. For each report of each of the 54 diseases, a special form was sent to the named diagnosing physician. Each form included a request for information regarding the fact and date of the reported diagnosis and a check-off list of usual and possible diagnostic criteria. To detect occurrences of diseases that the participant had not reported, the participant's

From the School of Public Health, University of California at Los Angeles (Drs Massey, Chang, Detels, Quiroga, and Spivey and Ms Coulson); the Department of Medical Statistics and Epidemiology (Drs O'Fallon and Kurland), the Department of Internal Medicine (Dr Emslander), the Mayo Clinic, Rochester, Minn; the Department of Obstetrics and Gynecology, University of Southern California, Los Angeles (Drs Bernstein, Nakamura, and Korelitz); the Department of Urology, School of Medicine, University of California, San Francisco (Dr Schmidt); the School of Public Health, University of Minnesota (Drs Schuman and Mandel), the Department of Surgery (Dr Benjamin), St Louis Park Medical Center, Minneapolis; the Department of Internal Medicine, Health Science Center, Dallas (Dr Sullivan); the James A. Baker Institute for Animal Health, Cornell University, Ithaca, NY (Dr McGregor); the Sterling-Winthrop Research Institute, Rensselaer, NY (Dr Lepow); and the Center for Population Research (Ms Crozier), the Epidemiology and Biometry Research Program (Dr Berendes), National Institute of Child Health and Human Development, Bethesda, Md.

Reprint requests to School of Public Health, University of California, Los Angeles, CA 90024 (Dr Massey).

Table 2.—Number of Deaths by Cause Among 10,590 Matched Pairs

Cause, *ICDA Codes**	Vasectomized Men	Matched Men	Total
Malignant neoplasms (140/207)	44	88	132
Ischemic heart disease (410/414)	63	103	166
Stroke (430/438)	5	12	17
Other circulatory (390/458 except 430/438)	13	18	31
Digestive system (520/577)	5	14	19
Other natural causes	9	16	25
Accidents, poisonings, and violence			
Suicide	24	28	52
Homicide	6	4	10
Other accidents, poisonings, violence	43	42	85
Total Deaths	212	325	537

*ICDA indicates International Classification of Diseases, Adapted.

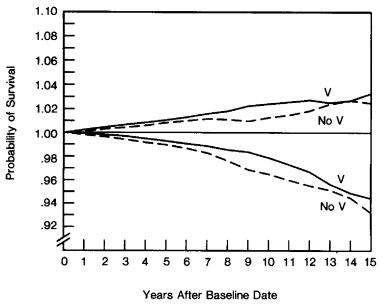

Fig 1.—Estimated probability of postbaseline survival from baseline date for vasectomized (V) and nonvasectomized (NoV) subjects; Kaplan-Meier method and survival as proportion of expected survival, 1970 US white male population.

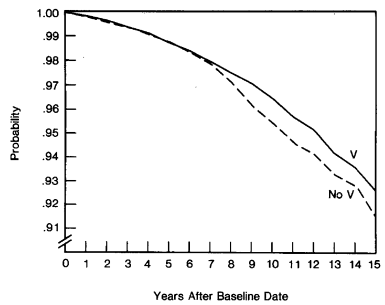

Fig 2.—Estimated probability of postbaseline survival free of any disease or death from cardiovascular core (Table 1, group 8) for vasectomized (V) and nonvasectomized (NoV) subjects.

Table 4.—Stratified Cox-Covariate Analyses*

Diseases, or Groups	Age, z†	Vasectomy, z†
Respiratory diseases		
Asthma	.34	−0.18
Heart or circulatory diseases		
Heart attack	5.11	−1.19
Angina pectoris	4.13	−0.32
Heart attack or angina pectoris	5.96	−0.52
Stroke	3.15	−0.88
Thrombophlebitis	3.09	−0.23
Liver diseases		
Hepatitis	−1.79	1.98
Glandular disorders		
Diabetes	1.62	−2.14
Bone or joint diseases		
Gout	0.79	−0.85
Cancer		
All except skin	5.19	−1.25
Genital or urinary disorders		
Epididymitis-orchitis		
Within 12 mo of baseline	−0.29	6.58
Total postbaseline	−0.92	8.33
Impotence or other sexual difficulty	2.55	1.19
Group 1, IgE-mediated disease		
Core	−0.02	−1.04
Group 7, blood coagulation defects		
Core	2.90	−1.00
Core + ring I	2.88	−1.09
Group 8, atherosclerotic disease		
Core	8.47	−0.65
Core + ring I	7.08	−0.91

*Subjects were stratified into 20 strata according to age (five levels) and residence (four locations) at baseline date. Age groups included 29 years or younger, 30 through 34 years, 35 through 39 years, 40 through 44 years, and 45 years or older. Locations included Rochester and Minneapolis-St Paul, Minn; Los Angeles; and Eureka, Calif.

†Ratio of estimated coefficient to its SE.

for death and for the eight groups of diseases. For the groups, pairs were included only if neither member had a prebaseline occurrence and at least one member had a postbaseline occurrence of one or more of the diseases or conditions in the groups. For each individual test a value of z greater than +1.65 would be labeled significant. However, the multiple tests performed in this study make this limit of +1.65 unreasonably small. It can be seen in Table 3 that this point is moot, since of the 39 tests only the z value for epididymitis-orchitis exceeds this level.

Without censoring at the earlier interview or death, the total number of man-years between the baseline date and the time of interview or death differed only slightly for vasectomized men (86,201 man-years) and their nonvasectomized matches (90,342 man-years). The postbaseline incidence rates per 10,000 man-years for each cohort of men and for the individual diseases occurring at least 15 times in the total cohort are also given in Table 3.

In Fig 1, survival curves are presented giving the estimated probability of survival (Kaplan-Meier method) from the baseline date for