Environmental Health Perspectives
Vol. 63, pp. 225-234, 1985

Relative Risk Regression Analysis of Epidemiologic Data

by Ross L. Prentice*

Relative risk regression methods are described. These methods provide a unified approach to a range of data analysis problems in environmental risk assessment and in the study of disease risk factors more generally. Relative risk regression methods are most readily viewed as an outgrowth of Cox's regression and life model. They can also be viewed as a regression generalization of more classical epidemiologic procedures, such as that due to Mantel and Haenszel.

In the context of an epidemiologic cohort study, relative risk regression methods extend conventional survival data methods and binary response (e.g., logistic) regression models by taking explicit account of the time to disease occurrence while allowing arbitrary baseline disease rates, general censorship, and time-varying risk factors. This latter feature is particularly relevant to many environmental risk assessment problems wherein one wishes to relate disease rates at a particular point in time to aspects of a preceding risk factor history. Relative risk regression methods also adapt readily to time-matched case-control studies and to certain less standard designs.

The uses of relative risk regression methods are illustrated and the state of development of these procedures is discussed. It is argued that asymptotic partial likelihood estimation techniques are now well developed in the important special case in which the disease rates of interest have interpretations as counting process intensity functions. Estimation of relative risks processes corresponding to disease rates falling outside this class has, however, received limited attention. The general area of relative risk regression model criticism has, as yet, not been thoroughly studied, though a number of statistical groups are studying such features as tests of fit, residuals, diagnostics and graphical procedures. Most such studies have been restricted to exponential form relative risks as have simulation studies of relative risk estimation procedures with moderate numbers of disease events.

Introduction

One of the most important developments in biostatistics in recent years has been the evolution of regression methods for "failure" time data. In epidemiology, failure may refer to the diagnosis of a certain disease or to death from the disease. Primary interest typically centers around the relationship between individual characteristics or exposures and subsequent disease incidence or mortality.

In a cohort study, a group of subjects is selected from a population of interest and followed forward in time for disease occurrence. Both baseline characteristics or exposures, and characteristics or exposures measured during follow-up, may be of interest as disease risk factors. Such information will be referred to as the subjects' covariate history.

A cohort study is often too long-term and expensive to be feasible particularly for exploratory studies of rare diseases. A case-control design involves the monitoring of a large population for disease occurrence followed by a retrospective ascertainment of covariate histories. Such ascertainment takes place both for a representative sample of cases of disease and for a suitably selected disease-free, or control, group. Case-control design strategies often involve some degree of matching of controls to cases in respect to potential "confounding" variables that may otherwise obscure the relationship between covariates of primary interest and disease occurrence.

The ideas of case-control sampling can also be useful in the context of a cohort study. Specifically, cases of disease arising in a cohort may be compared to a subset of the disease free group in the cohort in order to avoid the assembly of covariate histories on the entire cohort. Such an approach is useful, for example, in the exploitation of a serum bank, since biochemical or viral analysis of stored sera on every cohort member may be prohibitively expensive. More generally a "synthetic" case-control analysis of a large cohort data set may be considered strictly for computational reasons.

A hybrid "case-cohort" design in which covariate histories are assembled only for a preselected subcohort and for cases developing disease may give rise to further cost saving in the context of certain types of cohort studies.

*Fred Hutchinson Cancer Research Center, 1124 Columbia St., Seattle, WA 98104, and Department of Biostatistics, University of Washington, Seattle WA 98195.

Regression Analysis of Cohort Data

Regression Models

Disease occurrence data in a cohort study takes the form of a random time variable T for each subject. Typically T will be defined as time from entry into the cohort until disease occurrence, though other time specifications, such as age at disease occurrence, may be more natural in some applications. The time variate T will usually be subject to right censorship, as the subject may be without disease at the cut-off time for data analysis or may be lost to follow-up. Suppose initially that each subject has a fixed covariate vector z describing baseline characteristics or exposures under study, along with auxiliary data, for example, on potential confounding factors. The probability distribution for an absolutely continuous T can be equivalently described by its density, its survivor or distribution function, or by its (instantaneous) disease rate function

$$\lambda(t;z) = \lim_{\Delta t \to 0} \mathrm{pr}(t \leq T < t + \Delta t \mid T \geq t, z)/\Delta t$$

The disease rate, or hazard rate, function is a convenient representation for modeling purposes since it is natural to think in terms of disease rates and variations in disease rates over the follow-up course of a cohort study. Conventional parametric models, such as exponential and Weibull regression models, specify a hazard rate function of the form

$$\lambda(t;z) = \lambda_0(t) r(x\beta) \qquad (1)$$

where $\lambda_0(\cdot)$ and $r(\cdot)$ are fixed functions, $\boldsymbol{x} = x(z)$ is a row vector consisting of functions of z, and β is a corresponding column vector to be estimated. An exponential regression model is characterized by a $\lambda_0(t) \equiv \lambda$ while the $\lambda_0(\cdot)$ function is a power function of time in a Weibull regression model. Since the ratio of hazard functions at any two z-values is independent of t, the class (1) is sometimes referred to as the proportional hazards model. For uniqueness one requires $r(0) = 1$ so that $r(x\beta) = \lambda(t;z)/\lambda(t;z_0)$, where z_0 is a "standard" covariate vector giving rise to $x(z_0) = 0$. Since $r(x\beta)$ is the ratio of the failure rate at a general covariate vector to that at a standard vector it is often referred to as the relative risk function. Very often the relative risk function will be taken to be of exponential form, $r(\cdot) = \exp(\cdot)$, but other forms such as $r(\cdot) = 1 + (\cdot)$ may be more useful in some applications.

In many epidemiologic risk factor problems estimation of the relative risk is of primary interest, while the baseline disease rate function $\lambda_0(t) = \lambda(t;z_0)$ can be thought of as a nuisance parameter. A major advance in the theory of failure time regression took place when Cox (*1*) discovered that estimation of the regression parameter β could conveniently take place without placing any restrictions on the baseline hazard function $\lambda_0(\cdot)$. Based on a cohort giving rise to distinct failure times at $t_1, \ldots, t_d$ on subjects with respective regression vectors $x_1, \ldots, x_d$, Cox argued that standard asymptotic likelihood formulae could be applied to the function

$$L(\beta) = \prod_{i=1}^{d} [r\{x_i\beta\}/\sum_{l \in R(t_i)} r\{x_l\beta\}] = \prod_{i=1}^{d} L_i(\beta) \qquad (2)$$

where $R(t)$ denotes the set of subjects at risk for disease at t^-, for estimation of the relative risk parameter β. Under independent failure times and independent censorship (see below) the i-th factor in Eq. (2) is precisely the probability that failure occurs on the subject with regression vector $\boldsymbol{x}_i$, given the risk set $R(t_i)$ and given that exactly one failure is observed at t_i. Since the factors in Eq. (2) are dependent, special justification is required to show that Eq. (2) could be manipulated as an ordinary likelihood function, at least as far as asymptotic inference is concerned. Kalbfleisch and Prentice (*2*) showed $L(\beta)$ to have a marginal likelihood interpretation. Cox (*3*) introduced the notion of partial likelihood which not only encompasses Eq. (2) but also a range of important related functions arising from generalizations of the class of models (1). The fact that Eq. (2) is a partial likelihood function implies, very generally, that the score statistic

$$U(\beta) = \partial \log L(\beta)/\partial\beta = \sum_{i=1}^{d} \partial \log L_i(\beta)/\partial\beta = \sum_{i=1}^{d} U_i(\beta)$$

is such that each $U_i(\beta)$ has mean 0 and conditional variance estimated by $-\partial^2 \log L_i(\beta)/\partial\beta^2$, and that score statistic components $U_i(\beta)$ and $U_j(\beta)$ are uncorrelated, $i \neq j$. The partial likelihood structure then sets the stage for central limit theory to show $n^{1/2}(\hat{\beta} - \beta)$ to converge in distribution to a normal variate with mean vector zero and with variance matrix estimated by $nI^{-1}[\hat{\beta}] = -n\{\partial^2 \log L(\hat{\beta})/\partial\hat{\beta}^2\}^{-1}$, where n is the cohort size and $\hat{\beta}$ is the maximum partial likelihood estimate defined by $U(\hat{\beta}) = 0$. Formal asymptotic convergence results were developed somewhat later, notably by Tsiatis (*4*).

Efron (*5*) and Oakes (*6*) showed that it is not possible to improve on the efficiency of $\hat{\beta}$ provided $\lambda_0(\cdot)$ is completely unrestricted, and, equally important, that generally good efficiency properties obtain relative to the maximum likelihood estimates from parametric submodels of (1), even relative to parametric models that specify $\lambda_0(\cdot)$ up to a single scale parameter.

With arbitrary $\lambda_0(\cdot)$, the sole restriction in the model (1) is the relative risk specification $r(x\beta)$. The requirement that this relative risk be independent of follow-up time may be unnecessarily restrictive in many applications; in fact, the change over time in the relative risk associated with a certain characteristic or exposure may be of considerable interest in some settings. For example, one may be interested in latent periods and other aspects of the temporal pattern of cancer relative risk over time in the follow-up of cohorts exposed to ionizing radiation or other carcinogens. The model (1) is readily relaxed to allow a dependence of relative risk on time

by setting

$$\lambda(t;z) = \lambda_0(t)r\{x(t)\beta\} \quad (3)$$

where the modeled regression vector $x(t)$ now may consist not only of functions of z but also of product terms between functions of z and t. For example, with a single binary z, $r(\cdot) = \exp(\cdot)$ and $x(t) = (z, z \log t)$ the relative risk $\lambda(t;z=1)/\lambda(t;z=0)$ is $e^{\beta_1}t^{\beta_2}$ which is constant, monotone increasing, or monotone decreasing according to whether the coefficient β_2 is zero, positive, or negative. Based on Eq. (3), a partial likelihood function is readily developed that differs only from Eq. (2) through the replacement of x_i and x_l in the i-th factor of Eq. (2) by $x_i(t_i)$ and $x_l(t_i)$, respectively, for $i = 1, \ldots, d$.

To this point the regression model (3) presumes the parametric modeling of a relative risk function that includes not only the characteristics or exposures of primary interest, but also the auxiliary variables in z that may have been included, for example, to control confounding. Epidemiologic tradition, dating from the seminal paper by Mantel and Haenszel (*7*), very much concentrates on the use of stratification to control confounding or other potential biases.

The model (3) may be generalized to permit stratification by writing

$$\lambda(t,z) = \lambda_{0s}(t)\ r\{x(t)\beta_s\} \quad (4)$$

where the population is divided into q strata, $s \in [1, \ldots, q]$ on the basis of z values, and baseline disease rates $\lambda_{0s}(\cdot)$ are allowed to differ arbitrarily among strata. Note also that the regression parameter can be allowed to vary among strata. A partial likelihood function for $\beta = (\beta_1, \ldots, \beta_q)$ is readily developed as

$$L(\beta) = \prod_{s=1}^{q} \prod_{i=1}^{d_s} [r\{x_{si}(t_{si})\beta_s\}/\sum_{l \in R_s(t_{si})} r\{x_l(t_{si})\beta_s\}] \quad (5)$$

where $t_{sl}, \ldots, t_{sd_s}$ denote the distinct disease incidence times in stratum s and $R_s(t)$ denotes the set of subjects at risk in stratum s at t^-. A convenient approximation (*8*) is available to accommodate tied disease times within a stratum. Note also that stratum assignments may be time-dependent, that is $s = s(t,z)$, as a subject may move from one stratum to another during the course of follow-up.

Model (4) allows the data analyst the choice of stratification or regression modeling for the control of confounding factors. It therefore allows one to avoid excessive stratification that sometimes poses a problem in direct application of the Mantel-Haenszel technique, and also avoids the unnecessary restrictions or unwieldy regression models that may arise if Eq. (4) were used without stratification. In short, Eq. (4) allows one to extract the best from traditional epidemiologic methods and modern failure time data methods. In large cohorts with relatively rare disease occurrence there is evidently little efficiency loss through a detailed stratification on key confounding variables. Some further study of this topic would be worthwhile.

The regression model (4) presumes a fixed baseline regression vector z. An important aspect of a number of large scale epidemiologic cohort studies, however, is the periodic recording of risk factor and confounding factor levels during the course of follow-up. Denote by $z(u)$ a covariate measurement pertaining to follow-up time u and by $Z(t) = [z(u); u < t]$ the entire covariate history for a subject prior to time t. The disease rate at time t may be defined as

$$\lambda\{t;Z(t)\} = \lim_{\Delta t \to 0} \mathrm{pr}\{t \leq T < t + \Delta t \mid T \geq t, Z(t)\}/\Delta t$$

and a relative risk regression model

$$\lambda\{t;Z(t)\} = \lambda_{0s}(t)\ r\{x(t)\beta_s\} \quad (6)$$

may be defined, where $\lambda_{0s}(\cdot)$ is a baseline disease rate for stratum s, $x(t) = x[t,Z(t)]$ is a row regression p-vector that specifies the dependence of disease rate on risk factor histories under study, such that $x(t) = 0$ corresponds to a standard risk factor history, and by convention $r(0) = 1$. Regression models in the class (6) provide a flexible framework for a broad range of analyses to relate risk factor levels and changes in risk factor levels to subsequent disease incidence. A partial likelihood function for the estimation of $\beta = (\beta_1, \ldots, \beta_q)$ is once again given by Eq. (5).

Illustrations

There are many examples of the use of Eq. (4) in the literature. For example, Prentice et al. (*9*) apply Eq. (4) to a cohort of over 18,000 mice receiving a single time exposure to gamma radiation. The time-dependent feature of the regression variable in Eq. (4) was used to show that, for most cancer sites, the relative risk associated with a specific radiation dose drops off markedly as the animal's age.

For an illustration involving periodically measured covariate values consider a cohort of nearly 20,000 residents of Hiroshima and Nagasaki followed by the Radiation Effects Research Foundation. Prentice et al. (*10*) use data from this cohort to study the relationship between serial blood pressure measurements and subsequent cardiovascular disease incidence. Systolic and diastolic blood pressure along with a number of other cardiovascular disease risk factors and potential confounding factors were measured during the course of biennial examinations, beginning in 1958. The analyses described (*10*) make use of data on 16,711 subjects examined at least once during the time period 1958-74, including 108 incident cases of cerebral hemorrhage, 469 incident cases of cerebral infarction, and 218 incident cases of coronary heart disease. Specific objectives of their analysis concerned the relative importance of systolic and diastolic blood pressure as risk indicators for the three major cardiovascular disease categories just

mentioned, and the relative importance of blood pressure levels from two or more biennial exam periods before a risk period, given the blood pressure measurements from the most recent examination period. The application of relative risk regression methods described (*10*) used model (6) with t defined as the examination cycle (i.e., $t = 1$ in 1958-60, $t = 2$ in 1960-62, . . . with 32 strata defined on the basis of sex and 16 five-year age-at-baseline categories. The modeled regression vector $x(t)$ was taken to consist of systolic and diastolic blood pressure levels in examination cycles $1, 2, \ldots, t-1$ or functions thereof. Naturally, in order that $x(t)$ be defined, it is necessary that certain preceding examination have been attended and that the desired blood pressure measurements have been taken. In order to accommodate missing covariate data it is necessary to assume that the set of subjects at risk in examination cycle t with covariate history $Z(t)$ are represented by the subset for whom the corresponding $x(t)$ value can be specified. This assumption is subsumed in the independent censorship process described below. In terms of the partial likelihood function (5), the risk sets $R_s(t)$ consist only of those subjects under active follow-up in stratum s for whom the modeled regression vector $x(t)$ can be derived from available covariate data.

Table 1 shows the results of relative risk regression analyses with $r(\cdot) = \exp(\cdot)$, $x(t) = [\mathrm{SBP}(t-1), \mathrm{DBP}(t-1)]$ the systolic and diastolic blood pressure measurements in examination cycle $t-1$, and with common regression parameters across strata ($\beta_s \equiv \beta$). Note that previous cycle diastolic blood pressure is the important disease risk predictor for cerebral hemorrhage, while the corresponding systolic blood pressure is the more important predictor for cerebral infarction and for coronary heart disease. This observation has clinical implications and provides insight into the three disease processes.

Table 2 gives results of analyses in which a sequence of blood pressure measurements are related to subsequent disease incidence. The regression vector is now defined as $x(t) = [\mathrm{DBP}(t-1), \mathrm{DBP}(t-2), \mathrm{DBP}(t-3)]$ for cerebral hemorrhage and $x(t)$ equal to the corresponding SBP values from the three preceding cycles for cerebral infarction and coronary heart disease. Note that for a subject to contribute to the risk set in examination cycle t, all three previous biennial examinations need to have been attended. From Table 2 one can note that the most recent systolic blood pressure measurement is highly predictive of cerebral infarction risk, while the next most recent makes some additional contribution to risk prediction. With coronary heart disease, on the other hand, a recent elevated systolic blood pressure measurement is not predictive, or is possibly even negatively predictive, of risk given the levels of SBP in the two preceding cycles. One possible explanation for this result would be that hypertensive medication brings about blood pressure control without a corresponding reduction in coronary heart disease risk. The analysis for cerebral hemorrhage indicates that both elevated diastolic blood pressure and the duration of elevation are strong risk predictors.

Most applications to date of relative risk regression methods have presumed the exponential relative risk form $r(\cdot) = \exp(\cdot)$. Thomas (*11*) and Prentice et al. (*12*) use the linear form $r(\cdot) = 1 + (\cdot)$ to examine the joint dependence of certain cancer relative risks on radiation exposure and other factors.

Table 3, from Prentice et al. (*12*), is based on data from 40,498 subjects in a larger cohort monitored by the Radiation Effects Research Foundation. These subjects were surveyed at least once in the time period 1964-70 in respect to cigarette smoking habits and had available (T65) total body radiation dose estimates. In this analysis T is defined to be years since the subjects first survey participation and $Z(t) = [Z_1(t), Z_2(t)]$ is (t) defined to consist of radiation exposure information $Z_1(t)$ and cigarette smoking data (cigarettes per day and duration of smoking) $Z_2(t)$. The analysis also involved 128 fixed strata defined on the basis of age at radiation exposure (16 five-year classes), city, sex, and survey date (before or after the end of 1966). Table 3 shows results of fitting both exponential form $r(u) = \exp(u)$ and linear form $r(u) = 1 + u$ relative risk models with $x(t)$ defined to include linear and quadratic terms in $T65$ total dose estimate (truncated at 600 Rads), indicator variables for four cigarettes per day categories, and a single term involving both exposures defined as the product of $T65$ dose (truncated) and a cigarette per day variate that takes values 0 for nonsmokers, and values 1 to 4 for the four cigarette per day categories indicated in Table 3. The results given in Table 3 are based on 1570 cancer deaths excluding hematologic cancers (which are apparently not smoking related) and excluding short-term smokers with smoking durations of between 5 and 20 years. Smokers of less than 5 years duration were pooled with nonsmokers. The coefficient of the product term [($T65$ dose/100) × cig/day category]

Table 1. Relative risk regression of cardiovascular disease incidence in relation to previous examination cycle systolic and diastolic blood pressure measurements. The analyses stratify on age and sex.

Regression variable	Cerebral hemorrhage $\hat{\beta}(\times 10^2)$[a]	Cerebral infarction $\hat{\beta}(\times 10^2)$	Coronary heart disease $\hat{\beta}(\times 10^2)$
SBP(t-1)	0.58	1.77	1.15
	(0.30)[b]	(<0.0001)	(0.003)
DBP(t-1)	5.48	0.46	−0.46
	(<0.0001)	(0.36)	(0.56)
Cases	92	406	187

[a] $\hat{\beta}$ values are maximum partial likelihood estimates.
[b] Asymptotic significance levels for testing $\beta = 0$ are given in parentheses.

Table 2. Relative risk regression of cardiovascular disease incidence in relation to blood pressure measurements from the three preceding examination cycles. The analyses stratify on age and sex.

Regression variable	Cerebral hemorrhage $\hat{\beta}(\times 10^2)$[a]	Cerebral infarction $\hat{\beta}(\times 10^2)$	Coronary heart disease $\hat{\beta}(\times 10^2)$
SBP($t-1$)		1.13	−1.06
		(0.001)	(0.06)
DBP($t-1$)	3.23		
	(0.01)[b]		
SBP($t-2$)		0.80	1.46
		(0.03)	(0.007)
DBP($t-2$)	−1.07		
	(0.45)		
SBP($t-3$)		0.35	0.64
		(0.30)	(0.22)
DBP($t-3$)	4.77		
	(<0.0001)		
Cases	48	207	97

[a] $\hat{\beta}$ values are maximum pratial likelihood estimates.
[b] Asymptotic significance levels for testing $\beta = 0$ are given in parentheses.

is of particular interest. The significantly negative coefficient in the exponential form regression indicates that the relative risk corresponding to a joint exposure to radiation and cigarette smoke is less than the product of relative risks for the individual exposures. For example, the estimated relative risk for a nonsmoker exposed to 100 rads ($T65$) of radiation is exp{0.237 - 0.009} = 1.25, the estimated relative risk for a long-term 20 cigarette per day smoker with no radiation exposure is exp{0.565} = 1.76, while the estimated relative risk for a long-term 20 cigarette per day smoker with 100 rads of radiation exposure is exp{0.237 - 0.009 + 0.565 = 0.067(3)} = 1.81. This last number can be compared with the estimate (1.25)(1.76) = 2.20 which would apply under a multiplicative relative risk model. In good agreement, the linear form relative risk model gives estimates of 1 + 0.292 - 0.001 = 1.29 for a nonsmoker exposed to 100 rads, 1 + 0.774 = 1.77 for a long-term 20 cigarette per day smoker unexposed to radiation, and 1 + 0.292 - 0.001 + 0.774 - 0.094(3) = 1.78 for the long-term 20 cigarette per day smoker with an estimated 100 rads of exposure. This latter number may be compared with a relative risk estimate of 1 + 0.292 − 0.001 + 0.774 = 2.06 which would apply under an additive relative risk model. Table 3 thus implies that the relative risk for all nonhematologic cancer among individuals exposed to both radiation and cigarette smoke is less than a multiplicative model would imply and possibly less than additive as well. When a more thorough account of age at radiation exposure is taken there, however, ceases to be evidence against an additive relative risk model, but evidence for submultiplicativity remains. Such analyses provide useful insights into the carcinogenic mechanism in addition to their obvious public health implications.

Distribution Theory

Rigorous distribution theory for the maximum partial likelihood estimator and corresponding baseline disease

Table 3. Relative risk regression analyses of cigarette smoking and radiation exposure in relation to all nonhematologic cancer mortality. Ex-smokers and smokers with duration of smoking between 5 and 20 years are excluded, while short-term smokers (<5 years) are pooled with nonsmokers.

Regression variable	Exponential form RR model	Linear form RR model
$T65$ dose/100	0.237[a]	0.292
	(0.004)[b]	(0.04)
($T65$ dose/100)^2	−0.009	−0.001
	(0.59)	(0.99)
About 5 cig/day	0.179	0.226
	(0.16)	(0.16)
About 10 cig/day	0.438	0.562
	(<0.0001)	(<0.0001)
About 20 cig/day	0.565	0.774
	(<0.0001)	(<0.0001)
About 30 cig/day	0.785	1.213
	(<0.0001)	(<0.0001)
($T65$ dose/100) × cig/day category	−0.067	−0.094
	(0.006)	(0.05)
Maximized log likelihood	−10106.008	−10105.898
Cases	1570	1570

[a] Maximum partial likelihood estimate.
[b] Significance level for testing coefficient equal to zero.

rate estimators is given in Andersen and Gill (*13*). Their work was limited to the exponential relative risk form, a restriction removed by Prentice and Self (*14*). An overview of these developments will be given here.

It is convenient to change notation slightly and to assume a single stratum for notational ease. Denote the n subjects in the cohort by $i = 1, \ldots, n$. For the i-th subject, define the counting process $N_i(t)$ to take value zero up to disease occurrence for subject i and value one thereafter. Let the censoring process $Y_i(t)$ take value one if the i-th subject is at risk at the time t and value zero otherwise. Let $z_i(t)$ denote a covariate vector for the i-th subject at time t. Denote by

$$F_t^i = \{N_i(u), Y_i(u), z_i(u); u \leq t\}$$

the counting, censoring and covariate data for the i-th subject up to and including time t, and by $F_t = [F_t^1, \ldots, F_t^n]$ the collection of such information on the cohort. Assuming $[F_t, t \geq 0]$ to form a right continuous family of σ-algebras, one can define the disease rate process corresponding to the i-th subject at time t via

$$\lambda_i(t;F_{t-}) = \lim_{s \uparrow t} \lim_{\Delta t \downarrow 0} \mathrm{pr}[N_i(s + \Delta t) - N_i(s) = 1 \mid F_s](\Delta t)^{-1}$$

Two basic assumptions underlie the regression analyses described above. An independent failure time assumption among distinct study subjects requires

$$\lambda_i(t;F_{t-}) = \lambda_i(t;F_{t-}^i), \quad \text{all } (i,t) \tag{7}$$

indicating that the disease rate at time t for the i-th subject does not depend on data recorded for other study subjects. An independent censorship assumption requires further that

$$\lambda_i(t;F_{t-}^i) = Y_i(t)\, \lambda\{t;Z_i(t)\}, \text{ all } (i,t) \tag{8}$$

indicating that at times t at which the i-th subject is at risk $\{Y_i(t) = 1\}$ the disease rate is independent of the subject's prior censoring history, where $Z_i(t) = \{z_i(u), u < t\}$. This assumption requires the set of subjects at risk at a given $\{t,Z(t)\}$ to be representative of the subpopulation having these same $\{t,Z(t)\}$ values. The relative risk regression modeling assumption now enters upon setting

$$\lambda\{t;Z(t)\} = \lambda_0(t\, r\{x_i(t)\beta\}, \text{ all } (i,t) \tag{9}$$

Together Eqs. (7), (8), and (9) imply

$$\lambda_i(t;F_{t-}) = Y_i(t)\lambda_0(t)r\{x_i(t)\beta\} \tag{10}$$

The probability that subject $\{i\}$ develops disease at t_i given F_{t-_i} and given exactly one disease occurrence at t_i, is easily calculated as

$$\begin{aligned} &\mathrm{pr}\{i \text{ develops disease} \mid \text{failure at } t_i \text{ and } F_{t_i}\} \\ &= \lambda_i(t_i \mid F_{t_i})/\sum_{l=1}^{n} \lambda_l\{t_i \mid F_{t_i}\} \\ &= Y_i(t)\, r\{x_i(t_i)\beta\}/\sum_{l=1}^{n} Y_l(t)\, r\{x_l(t_i)\beta\} \end{aligned}$$

and, as before, a partial likelihood function for β is given by

$$L(\beta) = \prod_{i=1}^{n} [r\{x_i(t_i)\beta\}/\sum_{l=1}^{n} Y_l(t) r\{x_l(t_i)\beta\}]^{\delta_i}$$

where t_i is the observed follow-up time for subject i and δ_i indicates whether ($\delta_i = 1$) or not ($\delta_i = 0$) subject i was observed to develop disease. In stochastic integral notation one can write

$$\log L(\beta) = \sum_{i=1}^{n} \int_0^1 [\log r\{x_i(t)\beta\} - \log \sum_{l=1}^{n} Y_l(t) r\{x_l(t)\beta\}] dN_i(t)$$

where a finite follow-up period has been assumed.

The reason for introducing counting process and stochastic integral notation in this context is to make use of the counting process decomposition

$$N_i(t) = \Lambda_i(t) + M_i(t) \qquad i = 1, \ldots, n$$

where M_i is a locally square integrable martingale and, under slight regularity (*15*), the cumulative intensity process Λ_i relates to the above disease rate process via

$$\Lambda_i(t) = \int_0^t \lambda_i(u;F_{u-})du$$

The disease rate process which has been modeled via Eq. (10) as a relative risk regression model then has a representation as a counting process intensity. This representation allows convergence results for stochastic integrals over martingales to be applied in order to develop asymptotic convergence results for the maximum partial likelihood estimate and related quantities. In that convergence results for stochastic integrals with respect to martingales require the integrand to be a "predictable" process it is natural to require the processes appearing in Eq. (10), namely the censoring process Y_i and the regression process x_i, to have the sample paths that are left continuous with right hand limits.

The principal results to arise from applying martin-

gale convergence theory are: (*i*) $n^{-1/2}\,\partial \log L(\beta)/\partial\beta$ converges in distribution to a normal variate with mean zero and with variance matrix consistently estimated by $\hat{\Sigma} = -n^{-1}\,\partial^2 \log L\,\hat{\beta})/\partial\beta^2$; (*ii*) $n^{1/2}(\hat{\beta} - \beta)$ converges in distribution to a normal variate with mean zero and with variance matrix consistently estimated by $\hat{\Sigma}^{-1}$; and (*iii*) $n^{1/2}(\hat{\Lambda} - \Lambda_0)$ converges to a certain Gaussian process, where $\hat{\Lambda}$ is a natural estimator of the cumulative baseline disease rate

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du$$

Without going into detail, sufficient conditions for these convergences include a finite follow-up period, the asymptotic stability of certain processes arising in log $L(\beta)$ and its first and second derivatives, a Lindeberg condition, certain asymptotic regularity conditions and, in order to accommodate regression forms other than $r(\cdot) = \exp(\cdot)$, a regression positivity condition and a condition to assure the asymptotic stability of the observed information matrix. In spite of this rather lengthy list these conditions are collectively quite unrestrictive. For example, it is not necessary that $[N_i, Y_i, z_i]$ be independent and identically distributed. Recently Gill (*16*) has given an informal and intuitive presentation of this martingale approach.

An interesting technical point in this use of stochastic covariates relates to the fact that the disease rate process being modeled, namely $\lambda_i[t;F_{t-}]$, conditions on the subject's entire preceding covariate history. Risk factor associations of interest may, however, involve the relationship between disease rate and a subset of the preceding covariate history. For example, Table 1 above is concerned with cardiovascular disease rates in relation to previous blood pressure measurements, but only blood pressure measurements recorded in the immediately preceding examination cycle. An application of the asymptotic results just mentioned to Table 1 would then implicitly require one to assume disease rates to be independent of earlier blood pressure measurements, given the most recent measurements; an assumption not substantiated by Table 2. To address this issue, Self and Prentice, in a submitted manuscript, have generalized the above results to allow aspects of preceding covariate history to be excluded from the conditioning at the division points of a time axis partition. The relative risk parameter is then chosen to maximize a pseudo-likelihood function that is the product of partial likelihoods over the elements of the time axis partition. The maximum pseudo-likelihood function is identical to that which would be obtained by specifying an oversimplified intensity process model that involves only selected aspects of the preceding covariate history. An adjustment is required, however, to give a consistent variance estimator for this maximum pseudo-likelihood estimator.

Generalizations and Current Status of Relative Risk Regression Methods

The counting process formulation described above encompasses multivariate failure time data. Such a feature may be useful, for example, in studying the epidemiology of epileptic seizures or asthmatic attacks. Prentice et al. (*17*) and Andersen and Gill (*13*) consider relative risk regression models of the type

$$\lambda_i(t;F_{t-}) = \lambda_{0s}(t)r\{x(t)\beta_s\} \tag{11}$$

which merely continue the intensity process modeling for the *i*-th subject beyond the first failure time to the times of second and subsequent failures. Note that in Eq. (11) F_{t-} will include the counting process histories, including all preceding random failure times, for each subject and that the stratification $s = s(t)$ and regression variable may be defined to reflect aspects of the subject's preceding failure time information. For example, the subject may be required to move to the next stratum whenever the subject experiences a failure. Model (11) directly gives rise to a partial likelihood function to which the asymptotic results previously cited apply. A second class of multivariate relative risk regression models (*17*) can be written:

$$\lambda_i(t;F_{t-}) = \lambda_{0s}(t - t_i^*)r\{x(t)\beta_s\} \tag{12}$$

where t_i^* is the most recent random failure time on subject *i* prior to time *t*. This model also naturally gives rise to a partial likelihood function provided the stratification is fine enough to require the subject to enter a new stratum each time the subject experiences a failure. Formal asymptotic results for such estimation have been given in certain special cases (*18*).

Competing risk generalizations of relative risk regression models have also been described (*19*). Specifically, if *m* distinct disease categories may arise in a follow-up study, a relative risk regression model

$$\lambda\{t,j;Z(t)\} = \lambda_{0j}(t)\; r\{x(t)\beta_j$$

may be specified for the rate of disease *j* occurrence, for selected values of $j \in [1, 2, \ldots, m]$. Straightforward partial likelihood estimation of the disease-*j* relative risk regression parameter β_j proceeds by regarding disease occurrences of types other than *j* as censored.

Some work (*20*) has also taken place to allow relative risk regression parameter estimation and testing in the presence of random measurement errors in the covariate processes, a topic of obvious practical importance in epidemiologic research.

In the context of occupational mortality data studies, Breslow et al. (*21*) have considered the use of external

mortality rate data, for example from vital statistical records, in order to partially specify the baseline disease rates $\lambda_{0s}(\cdot)$. In most applications such usage typically turns out to provide little benefit in respect to estimation efficiency, provided the baseline rates are allowed to differ from the external rates by a scale factor. Such external rates are, of course, indispensible if the cohort is essentially homogeneous in respect to the covariate histories of interest. See Breslow (*22*) for a discussion of relative risk regression estimation in these circumstances.

To date there has been rather limited study of the sample sizes and data configurations necessary to ensure a good approximation by the asymptotic distributions mentioned above. Johnson et al. (*23*) describe some simulation results pertinent to a fixed regression vector and exponential form relative risk function.

A full regression approach, of course, requires not only suitable model fitting and estimation procedures, but also a range of procedures for model criticism. In general the area of model criticism is at a rather early stage of development for relative risk regression methods. Some relevant works include proposals in respect to test of fit (*24*), residuals (*25–28*), regression diagnostics (*29*), and choice of relative risk form (*11*).

Relative Risk Regression for Time-Matched Case-Control Studies

Suppose now that a large population is being monitored for disease occurrence, perhaps by means of a cancer registry or by a mortality index. It would often be impractical to enumerate and collect covariate data on such a large cohort, and furthermore since disease rates are likely to be low the data on many of the individuals who do not develop disease in some defined "follow-up" period will be largely redundant. The case-control design provides a valuable and much used alternative to the cohort design in such circumstances. A time-matched case-control study would proceed by matching each case that arises in some defined case accession period to one or more control subjects who are without disease at the time of case ascertainment. Here time would usually refer to age, although other specifications, including calendar time may be preferable in some applications. The cases ascertained by the disease surveillance system should be representative of the cases arising in the population in respect to their prior covariate histories, and controls selected should be representative of the sub-population who are without disease at the "time" of control ascertainment. Controls may also be matched to cases in respect to other potential confounding factors in which case the controls corresponding to a specific case need only to be representative of the disease free group in that stratum at the time of case occurrence. Upon selection the covariate histories $Z(t)$ are ascertained retrospectively for cases and controls, usually by personal interview. Here t refers to the time of case occurrence. A major methodologic concern relates to the ability to retrospectively construct accurate covariate histories, and to do so equally for cases and controls (recall bias), and the ability to sample randomly from case and control populations (selection bias). Assuming these concerns are met a relative risk regression model (6) is readily applied to case-control data (*30,31*). Specifically a suitable likelihood function is again given by Eq. (5), where $t_{s1}, \ldots t_{sd_s}$ are the times of case ascertainment in the s-th stratum and $R_s(t_{si})$ consists only of the case occurring at t_{si} along with its corresponding time- and stratum-matched controls. The (s,i) factor of Eq. (5) can be derived as the conditional probability that covariate history $Z_{si}(t_{si})$, giving rise to the regression vector $x_{si}(t_{si})$, corresponds to the diseased individual, given the set of covariate histories $[Z_l(t_{si}),\ l \in R_s(t_{si})]$ and the fact that $R_s(t_{si})$ includes exactly one case. This assertion requires an independent disease times assumption. Such an assumption furthermore implies that the contributions to Eq. (5) at distinct (s,i) are statistically independent, since distinct individuals are involved at each (s,i), so that Eq. (5) has a conditional likelihood interpretation. It follows that standard asymptotic likelihood methods can be expected to apply to Eq. (5), under time-matched case-control sampling, under mild conditions (*32*). Note that, under model (6), covariate histories need be assembled only to the point of permitting $x(t)$ to be specified at the time of occurrence for a case, or at the time of the corresponding case occurrence for a matched control.

Synthetic Case-Control and Case-Cohort Designs

Consider again the cohort study discussed above. Partial likelihood estimation based on Eq. (5) can be computationally intensive especially with large cohorts and time-dependent regression variables. Consequently a number of authors (*31,33–37*) have suggested the imposition of case-control sampling on the cohort for computational reasons. This idea involves replacement of the denominator in each (s,i) factor of Eq. (5) by a summation over a set that includes only the subject developing disease at t_{si} disease and a comparison group randomly selected from $R_s(t_{si})$. In many situations selection of as few as five "controls" per case will yield regression parameter estimates of high efficiency (e.g., 80% or more) compared to a full cohort analysis, though Breslow et al. (*21*) indicate that twenty or more controls per case may be necessary to ensure good efficiency in the presence of large relative risks and unbalanced regression variable distributions.

The synthetic case-control approach is a useful aid to the data analyst in the exploration of a large cohort data set. Not only might risk sets in Eq. (5) involving several thousand subjects be replaced by sets involving only 10 or 20 subjects, but also only a single fixed regression vector $x(t)$ needs to be stored for each subject selected.

Equally important, the synthetic case-control ap-

proach gives the possibility of considerable cost saving in relative risk estimation in circumstances wherein assembly of key covariate data requires expensive synthesis of specimens or other raw materials that have been collected and stored during the course of a cohort study. For example, a number of prominent cohort studies and disease prevention trials have developed blood serum banks on large numbers of participating subjects. The use of these serum samples, for example to relate biochemical factors to subsequent disease incidence, may, however, be prohibitively expensive. The synthetic case-control design allows efficient relative risk estimation based on serum analyses for cases and a small number of time-matched controls. A full cohort analysis on the other hand would typically involve a much larger number of serum analyses.

The synthetic case-control design does not, however, appear to be the most efficient approach to this type of estimation problem. In particular, a given subject could properly serve as a control for a number of cases arising at times during the subject's risk period. The synthetic case-control approach, however, rather arbitrarily links a specific control subject to a single case. Prentice has proposed (*38*) a case-cohort design to avoid this limitation. In such a design a subcohort is randomly selected from the entire cohort to serve as comparison group for all cases arising during follow-up. The sampling can be relaxed to allow different sampling fractions among baseline defined strata.

Estimation can then be based on Eq. (5) with the risk set $R_s(t_{si})$ replaced by a set that consists only of the case occurring at t_{si} and the subcohort risk set at t^-_{si}. It follows that covariate histories need be assembled only for cases and subcohort members. With the risk sets modified as just mentioned standard asymptotic likelihood formulae can evidently be applied to Eq. (5) with a modification to the score statistic variance to accommodate a correlation among score statistic contributions within a stratum. Specifically, the score statistic contribution at t_{si} will typically be weakly correlated with score statistic contributions at t_{sj}, $j<i$ whenever the disease occurrence at t_{si} arises outside the selected cohort.

This work was supported by grants GM-24472, GM-28314 and CA-34847 from the National Institutes of Health. Parts of this manuscript are identical to subsets of a manuscript by R. L. Prentice and V. T. Farewell, which appeared in the Proceedings of Second IMACS Symposium on Biomedical Systems Modeling, North Holland, Amsterdam.

REFERENCES

1. Cox, D. R. Regression models and life tables (with discussion). J. Roy. Statist. Soc. B34: 187–220 (1972).
2. Kalbfleisch, J. D., and Prentice, R. L. Marginal likelihoods based on Cox's regression and life model. Biometrika 60: 267–278 (1973).
3. Cox, D. R. Partial likelihood. Biometrika 62: 269–276 (1975).
4. Tsiatis, A. A. A large sample study of Cox's regression model. Ann. Statist. 9: 93–108 (1981).
5. Efron, B. Efficiency of Cox's likelihood function for censored data. J. Am. Statist. Assoc. 72: 557–565 (1977).
6. Oakes, D. The asymptotic information in censored data. Biometrika 64: 441–448 (1977).
7. Mantel, N., and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 22: 719–748 (1959).
8. Breslow, N. E. Covariance analysis of censored survival data. Biometrics 30: 89–99 (1974).
9. Prentice, R. L., Peterson, A. V., and Marek, P. Dose mortality relationship in RFM mice following 136 Cs gamma irradiation. Radiat. Res. 90: 57–76 (1982).
10. Prentice, R. L., Shimizu, Y., Lin, C. H., Peterson, A. V., Kato, H., Mason, M. W., and Szatrowski, T. P. Serial blood pressure measurements and cardiovascular diseasae in a Japanese cohort. Am. J. Epid. 116: 1–28 (1982).
11. Thomas, D. C. General relative risk models for survival time and matched case-control analysis. Biometrics 37: 673–686 (1981).
12. Prentice, R. L., Yoshimoto, Y., and Mason, M. W. Relationship of cigarette smoking and radiation exposure to cancer mortality in Hiroshima and Nagasaki. J. Natl. Cancer Inst. 70: 611–622 (1983).
13. Andersen, P. K., and Gill, R. D. Cox's regression model for counting processes: a large sample study. Ann. Statist. 10: 1100–1120 (1982).
14. Prentice, R. L., and Self, S. G. Asymptotic distribution theory for Cox-type regression models with general relative risk form. Ann. Statist. 11: 804–813 (1983).
15. Aalen, O. O. Nonparametric inference for a family of counting processes. Ann. Statist. 6: 701–726 (1978).
16. Gill, R. Understanding Cox's regression model: a martingale approach. J. Am. Statist. Assoc. 79: 441–447 (1984).
17. Prentice, R. L., Williams, B. J., and Peterson, A. V. On the regression analysis of multivariate failure time data. Biometrika 68: 373–379 (1981).
18. Voelkel, J. G., and Crowley, J. Nonparametric inference for a class of semi-Markov processes with censored observations. Ann. Statist. 12: 142–160 (1984).
19. Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. The analysis of failure times in the presence of competing risks. Biometrics 34: 541–554 (1978).
20. Prentice, R. L. Covariate measurement errors and parameter estimation in Cox's failure time regression model. Biometrika 69: 331–342 (1982).
21. Breslow, N. E., Lubin, J. H., Marek, P., and Langholz, B. Multiplicative models and cohort analysis. J. Am. Statist. Assoc. 78: 1–12 (1983).
22. Breslow, N. E. Some statistical models useful in the study of occupational mortality. In: Environmental Health: Quantitative Methods (A. Whittemore, Ed.), SIAM, Philadelphia, 1978, pp. 88–103.
23. Johnson, M. E., Tolley, H. D., Bryson, M. C., and Goldman, A. S. Covariate analysis of survival data: a small-sample study of Cox's model. Biometrics 38: 685–698 (1982).
24. Schoenfeld, D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. Biometrika 67: 145–153 (1980).
25. Kalbfleisch, J. D., and Prentice, R. L. The Statistical Analysis of Failure Time Data. Wiley, New York, 1980.
26. Lagakos, S. W. The graphical evaluation of explanatory variables in proportional hazards regression models. Biometrika 68: 93–98 (1981).
27. Schoenfeld, D. Partial residuals for the proportional hazards regression model. Biometrika 69: 239–242 (1982).
28. Anderson, P. K. Testing goodness-of-fit of Cox's regression and life model. Biometrics 38: 67–78 (1982).
29. Storer, B., and Crowley, J. A diagnostic for Cox regression and general conditional likelihoods. J. Am. Statist. Assoc. 80: 139–147 (1985).
30. Prentice, R. L., and Breslow, N. E. Retrospective studies and failure time models. Biometrika 65: 153–158 (1978).
31. Liddell, F. D. K., McDonald, J. C., and Thomas, D. C. Methods of cohort analysis: appraisal by application to asbestos mining. J. Roy. Statist. Soc. A140: 469–491 (1977).
32. Andersen, E. B. Asymptotic properties of conditional maximum

likelihood estimators. J. Roy. Statist. Soc. B32: 283–301 (1970).
33. Mantel, N. Synthetic retrospective studies and related topics. Biometrics 29: 479–486 (1973).
34. Breslow, N. E., and Patton, J. Case-control analysis of cohort studies. In: Energy and Health (N. E. Breslow and A. S. Whittemore, Eds.), SIAM, Philadelphia, 1979, pp. 226–242.
35. Whittemore, A. S., and McMillan, A. Analyzing occupational cohort data: Application to U.S. uranium miners. In: Environmental Epidemiology: Risk Assessment (R. L. Prentice and A. S. Whittemore, Eds.), SIAM, Philadelphia, 1982, pp. 65–81.
36. Oakes, D. Survival times: aspects of partial likelihood. Int. Statist. Rev. 49: 235–264 (1981).
37. Lubin, H. J., and Gail, M. H. Biased selection of controls for case-control analyses of cohort studies. Biometrics 40: 63–75 (1984).
38. Prentice, R. L. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika, in press.