Modern Statistical Methods in Chronic Disease Epidemiology

Edited by
Suresh H. Moolgavkar
& Ross L. Prentice

Modern Statistical Methods in Chronic Disease Epidemiology

EDITED BY

Suresh H. Moolgavkar

Ross L. Prentice

Proceedings of a Conference sponsored by
SIAM Institute for Mathematics and Society
and supported by the
Department of Energy

A Wiley-Interscience Publication
JOHN WILEY & SONS
New York Chichester Brisbane Toronto Singapore

SIMS

The SIAM Institute for Mathematics and Society was established in 1973 by the Society for Industrial and Applied Mathematics. Its purpose is to develop, promote, support, and maintain research in the application of mathematics in the study and solution of social problems. To this end, SIMS conducts conferences relevant to its objectives, a transplant program wherein mathematicians are "transplanted" for two years into university interdisciplinary centers to work as members of a team on societal problems, and university research and education studies on statistics and environmental factors in health.

Library of Congress Cataloging in Publication Data:

Modern statistical methods in chronic disease epidemiology.

"A Wiley-Interscience publication."
Conference held in 1985 in Alta, Utah.
1. Chronic diseases—Research—Statistical methods—Congresses. 2. Epidemiology—Statistical methods—Congresses. I. Moolgavkar, Suresh H. II. Prentice, Ross L. III. SIAM Institute for Mathematics and Society. [DNLM: 1. Epidemiologic Methods—congresses. 2. Statistics—congresses. WA 950 M689 1985]
RA644.5.M63 1986 614.4 86-1597
ISBN 0-471-83904-3

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To the memory of Mark Kac,
enthusiastic founder and ardent supporter of SIMS,
this volume is affectionately dedicated

Foreword

In 1974 SIMS initiated a series of five-day Research Application Conferences (RAC's) at Alta, Utah, for the purpose of probing in depth societal fields in light of their receptivity to mathematical and statistical analysis. The first nine conferences addressed ecosystems, epidemiology, energy, environmental health, time series and ecological processes, energy and health, energy conversion and fluid mechanics, environmental epidemiology: risk assessment, and atomic bomb survivor data: utilization and analysis.

These *Proceedings* are a result of the tenth conference "Modern Statistical Methods in Chronic Disease Epidemiology" which was held in 1985. Forty speakers and observers contributed their expertise in such disciplines as biometry, environmental medicine, epidemiology, genetics, mathematics, and statistics. Topics addressed were: issues in matching and covariate adjustment, choice of primary time variate and evolutionary covariates, design and analysis of prevention trials, problems involving auxiliary and incomplete covariate data, confidence region and model criticism, absolute and relative risk methods, methods in genetic epidemiology, models for carcinogenesis and cancer progression, and multivariate failure time methods.

Suresh H. Moolgavkar and Ross L. Prentice, both of the Fred Hutchinson Cancer Research Center (Seattle) and the University of Washington (Seattle) co-chaired the Conference. Donald R. Snow of Brigham Young University served as Assistant Conference Director.

The Conference was supported by the Department of Energy, Human Health and Assessments Division, Office of Health and Environmental Research, Office of Energy Research.

D.L. Thomsen, Jr.
President

August 1985

Preface

The last quarter century, since the publication of Mantel and Haenszel's pioneering paper in 1959, has seen a veritable explosion of statistical methodology in chronic disease epidemiology. The central methodologic issues revolve around environmental and genetic risk assessment, and risk extrapolation. The tenth Research Application Conference held under the auspices of SIMS brought together experts from around the world to discuss the theory and applications of statistical methods in chronic disease epidemiology. This volume represents the proceedings of that conference.

Relative risk regression models provide flexible and powerful tools for the analysis of epidemiologic data. These models have been the objects of intense study in the past several years, and it seems reasonable to predict that relative risk regression methods will become a, or perhaps, the, central analytical tool in chronic disease epidemiology. Thus, a major emphasis of the conference was on relative risk regression, and various papers in this volume deal with time-dependent covariates, new study designs, multivariate failure time data, methods of model criticism, parameter transformations for optimal inference, and issues in matching, covariate adjustment, and incomplete and missing covariate information.

The relative risk regression models in current use are generalizations of a semi-parametric model for survival data analysis proposed by Cox in 1972. In the original model of Cox, the relative risk function was $\exp(\beta^t z)$, where β is a vector of parameters and z is a vector of covariates. Estimation of β proceeds via maximization of a partial likelihood. The original model has been generalized in two main directions. First, relative risk functions other than the exponential are being increasingly used. Second, the covariates are allowed to evolve over time. The large sample properties of such generalized models are now fairly well understood and elegant proofs using martingale theory are avaiable.

Nevertheless, the use of time-dependent covariates raises some technical problems. An approach to some of these is described in the paper by Andersen. The use of relative risk functions other than the exponential leads to problems in small to moderate sized samples. The use of parameter transformations to alleviate some of these problems is considered in the paper by Moolgavkar and Venzon.

Epidemiologic studies are largely observational in nature, and particular care needs to be exercised in their design. Often, the cost of processing information on a large number of study subjects is an important consideration. The issues arising in various designs for cohort studies are discussed in the paper by Prentice et al. A consequence of the observational nature of epidemiologic studies is that covariate information is sometimes missing and often measured with error. The impact and accommodation of covariates that are

measured with error are discussed in the paper by Whittemore and Grosser. The impact, on various aspects of the data analysis, of the complete omission of certain 'balanced' covariates is discussed by Gail. Careful selection of controls is of crucial importance in epidemiologic studies. Often controls with the appropriate characteristics are difficult to find. Partial matching is discussed in the paper by Greenland.

There has recently been interest in the analysis of failure time data in which the response in subgroups of individuals may be correlated. This situation may arise, for example, in twin studies. The papers by Oakes and by Self and Prentice discuss the issues that arise in multivariate failure time data.

An important area of research is the development of methods of model criticism for the relative risk regression models used for the analysis of epidemiologic studies. While much work still remains to be done, some approaches are discussed in the papers by Lustbader and Davis et al.

Often, time measurements other than elapsed time on study, may be of importance in the analysis of cohort data. A "real time" approach to the analysis of cohort data, which does not require the rezeroing of time as subjects enter the cohort, is advocated in the paper by Arjas. Finally, papers by Breslow and Thomas discuss the fitting of certain non-standard relative and absolute risk models to epidemiologic data.

It is becoming increasingly clear that most chronic diseases are a complex interplay of heredity and environment. Genetic epidemiologists have devised powerful and flexible statistical tools for the analysis of pedigree and linkage data. Unfortunately, there is a paucity of dialogue between scientists whose primary interest is the environment and those whose primary interest is heredity. Such a dialogue could only benefit both groups. Papers on pedigree and path analysis by Elston and Rice, respectively, are valuable contributions to such interchange.

Often, the risk to human populations from exposures to low levels of various agents must be inferred from the results of experiments in which animals have been exposed to very high levels of the agent in question. Various statistical methods have been devised for such "low-dose extrapolation", and at least some of these methods are based on biologically derived models. A satisfactory solution to the extrapolation problem is not presently at hand. However, the paper by Krewski et al addresses the issue and describes some models currently in use.

Finally, in some cancers, early detection appears to improve prognosis. For example, screening for cervical cancer is now widespread. The statistical issues involved in large scale screening

programs are the subject of the paper by Day and Walters.

The conference was characterized by excellent presentations and stimulating discussions. We feel that this collection of papers represents a timely and provocative discourse on some of the central statistical issues in chronic disease epidemiology.

Suresh H. Moolgavkar
Ross L. Prentice

Seattle, 1985

SECTION 1
Aspects of the Validity and of the Design of Epidemiologic Studies

A central issue in the interpretation of observational studies concerns the possible omission of important explanatory factors or covariates. Mitchell Gail considers the omission of a covariate that, by design, is not associated with the primary exposure variable in a cohort setting. His paper demonstrates that, except in certain special cases, such omission can be expected to bias parameter estimates that relate exposure level to disease occurrence. Even if such bias does not occur, standard tests of the absence of exposure effect may well be invalid.

Alice Whittemore and Stella Grosser consider problems that arise when the exposure variables of interest are measured inaccurately. This is a topic of obvious importance in observational studies, and one that has received limited attention in the context of odds ratio or relative risk methods that are commonly used in epidemiologic research. One wonders, for example, the extent to which an apparent lack of consistency of results relating intake levels of selected nutrients to the incidence of major chronic diseases may be attributed by important and highly correlated measurement errors among the estimated nutrient intakes. Whittemore and Grosser provide a quite general approach to the use of information on the covariate error distribution into regression analysis, along with several illustrations.

Literature on the role of matching in case-control studies is quite extensive. Issues of practicality and ease of conduct may compete with those of simplicity and efficiency of data analysis. The paper by Sander Greenland notes that much of the benefit of a fully matched, but possibly awkward, design can be retained by certain logistically advantageous 'partially matched' designs.

The time-matched case-control design has been advocated by a number of authors for relative risk regression in the context of a large cohort study. Possible 'case-control within cohort' designs are described in the paper by Prentice, Self and Mason, as is a case-cohort design. This latter design appears to have some efficiency advantages relative to corresponding case-control designs and is particularly useful in situations (e.g., large scale prevention trials) where it is useful to be able to identify the 'comparison group' prior to cohort follow-up.

Adjusting for Covariates That Have the Same Distribution in Exposed and Unexposed Cohorts

*Mitchell H. Gail**

Abstract. We examine the effects of omitting a balanced covariate, namely a covariate, X, that has the same distribution among exposed and unexposed subjects, from regression analyses of cohort data. Except for models with linear or multiplicative regressions of the response variable on exposure and X, omission of a balanced covariate yields biased estimates of treatment effect. Moreover, even in cases where bias is not introduced, omitting X can lead to hypothesis tests for no exposure effect that have supranominal size, if the Fisher information is used to estimate required variances. A robust variance estimate is recommended, instead, which leads to tests of nominal size, but omitting X can still lead to substantial power loss. These ideas are discussed in relation to the following models (Table 1) for epidemiological cohort studies: normal linear, exponential multiplicative, exponential reciprocal, Bernoulli logistic, Bernoulli additive, Bernoulli multiplicative, Poisson multiplicative, Cox model, and the proportional hazards model for paired survival data.

INTRODUCTION

The ready availability of computing facilities allows epidemiologists to perform a variety of linear and non-linear regression analyses in order to estimate the effects of exposure, to detect effect modification, and to adjust for potential confounding variables [3,5,17,20,23,31]. Which potential confounders or effect modifiers to include in a regression model is problematic, and this issue is a topic of continuing research and discussion [8, 10,22]. We shall consider the implications of discarding a covariate X that has the same distribution in the exposed and unexposed cohorts.

Recent publications show that estimates of treatment effect may or may not be changed by ignoring such a covariate, depending on the nature of the response measurement. For example, if the dis-

*Mitchell H. Gail, National Cancer Institute, Bethesda, Maryland 20892

tribution of X is the same in each of two exposure cohorts, a "valid" (i.e., asymptotically unbiased) estimate of relative risk may be obtained without adjustment on X, whereas adjustment on X is required to obtain a "valid" estimate of the odds ratio of disease [2,15,28]. The purpose of this paper is to describe more generally what happens to inference about exposure when a balanced covariate is omitted from the model.

The emphasis will be on cohort designs, which permit one to study a wide variety of possible response measures, rather than on case-control designs, which typically yield only the relative odds. The prospective risk models we use apply directly to cohort data, and much of the required theory has already been developed for randomized clinical trials in which the distribution of X is known to be independent of treatment, T [12,13].

In a cohort study we assume that the response variable, Y, for an individual with exposure T and covariate X, has a conditional density $f(Y|T,X)$. The likelihood is the product of such densities over all study participants. Clearly, no aspect of the inference on exposure will be altered by omitting X if $f(Y|T,X)=f(Y|T)$, namely if Y is conditionally independent of X, given T. We call this requirement NC1 (non-confounding specification 1). This is the strongest requirement for X to be a non-confounder. It is equivalent to model $\mathfrak{E}$ in Samuels [26], who considers dichotomous responses. Other less stringent criteria of non-confounding may be utilized. For example, suppose $f(Y|T,X)$ is normal with conditional expectation $E(Y|T,X)=\mu+T\alpha+X\beta$ and conditional variance σ^2. Then, if $\beta\neq 0$, NC1 does not hold. Yet if X and T are independent, standard results in linear regression analysis show that estimates $\hat{\alpha}^*$ of α^* in the false model with X omitted, namely $E(Y|T,X)=\mu^*+T\alpha^*$, will converge to the true treatment effect, α, for large samples. In epidemiological parlance, the estimate $\hat{\alpha}^*$ is "valid". We define non-confounder criterion NC2 to be the condition that estimates of the treatment effect with X omitted converge to the true treatment effect α, and we apply this notion to a variety of response models. Yet another possible definition of non-confounder, NC3, is the condition that model-based score tests for no treatment effect retain nominal size when X is omitted. For most models this is not so, though the problem can be circumvented by replacing the Fisher information with a robust estimate of the variance of the score. Clearly NC1 implies both NC2 and NC3. However, we shall discuss models that satisfy NC2 but neither NC1 nor NC3, and models, like the logistic, that satisfy NC3 but neither NC1 nor NC2.

There are several epidemiological settings in which the covariate X and exposure indicator T might be statistically independent, or at least uncorrelated. The unexposed cohort (T=-1) might be chosen to have the same distribution of X as the exposed cohort (T=1). This could be accomplished by pair matching exposed and unexposed individuals on X or by randomly sampling unexposed individuals whose X values

fall into various categories with probabilities defined by the conditional density $f(X|T=1)$. This latter procedure is called "frequency matching" [29]. Recently, Rosenbaum and Rubin [24,25] defined the "propensity score", $e(X)\equiv P(T=1|X)$, where X includes all possible covariates, and they showed that any component of X, such as X, is conditionally independent of T given e(X). Thus within strata defined by e(X), X and T are independent. Our results might also be of interest to the data analyst who has just compared the distributions of a number of covariates in the exposed and unexposed groups and has identified several covariates that are uncorrelated with T. Less commonly, an epidemiologist might have access to data from a randomized experiment. For example, Boice et al [1] studied the long term risk of leukemia in patients who had previously been randomly assigned to receive the alkylating agent, Semustine, for treatment of gastrointestinal cancer. The exposure to Semustine had been assigned at random, guaranteeing the independence of T and X. We shall use the phrases "X is independent of T" and "X is a balanced covariate" interchangeably. Although the results we present on bias and the NC2 criterion hold in each of these settings, our comments on hypothesis testing, power, "variance deflation" and the NC3 criterion pertain only to the last three situations, where it is reasonable to suppose that individuals are selected by simple random sampling from a defined population. As emphasized by Weinberg [29], frequency matching and pair matching induce variances that correspond to stratified random sampling, for which our results on the NC3 criterion are not directly applicable.

RISK MODELS AND A SUMMARY OF RESULTS FOR BALANCED COVARIATES

We imagine two cohorts of individuals, one unexposed (T=-1) and one exposed (T=1). For simplicity we assume equal numbers in each cohort. We observe an individual in exposure group T with covariate X and subsequently measure his response Y. The response Y may be a quantitative measurement like blood pressure, or a categorical event, like whether or not he survived a fixed time interval. We suppose that the expectation of Y depends on T and X according to the regression model

$$E(Y|T,X) = h(\mu + T\alpha + X\beta) \qquad (2.1)$$

where h is a twice differentiable function. We assume that X is a scalar covariate, independent of T. Without loss of generality, we center X so that $E(T)=E(X)=0$. The response Y is assumed to depend on T and X only through the argument

$$\eta = \mu + T\alpha + X\beta \quad . \qquad (2.2)$$

Equation (2.1) defines what is meant by "treatment effect", α. Note that this model does not include an interaction term for effect modification. Thus each subject has the same treatment effect, regardless of X. Model (2.1) is oversimplified in one important respect; it ignores the possible influence of other covariates $\underline{X}$ that may or may not be independent of T. As Fisher and Patil [10] demon-

strate, confounders should be evaluated jointly, but to do so requires specification of joint probability distributions and is beyond the scope of this paper. However, it is straightforward to extend results on a scalar X to a vector of covariates, all of which are independent of T [12,13].

We shall concentrate primarily on members of the exponential family

$$f(Y|T,X;\mu,\alpha,\beta)=f(Y|\eta)=\exp\{K(\emptyset)[Y\gamma(\eta)-g\{\gamma(\eta)\}+r(Y)]+\psi(Y,\emptyset)\} \quad (2.3)$$

where $\gamma(\eta)$ is the "natural parameter" linking X and T to Y, $g'\{\gamma(\eta)\} = E(Y|T,X) = h(\eta)$, and $K(\emptyset)$ is a positive scale factor that usually equals 1.0 in our models.

Some commonly used models are listed in Table 1. All but the Cox and paired survival models, which are defined in the next section, fall within the framework outlined at (2.3). In the next section, we shall discuss each of the models in Table 1 in relation to the following general results, most of which are taken from Gail, Wieand and Piantadosi [12] and from Gail, Tan and Piantadosi [13].

TABLE 1

Some Models Used in Epidemiological Cohort Studies[†]

	$\gamma(\eta)$	$h(\gamma)$	Asymptotic Bias	Variance Deflation
Normal Linear	η	η	0	yes
Uncensored Exponential Multiplicative	$-e^{\eta}$	$e^{-\eta}$	0	yes
Uncensored Exponential Reciprocal	$-\eta$	η^{-1}	$\lvert\alpha\rvert<\lvert\alpha^*\rvert$	yes
Bernoulli Logistic	η	$e^{\eta}(1+e^{\eta})^{-1}$	$\lvert\alpha^*\rvert<\lvert\alpha\rvert$	no
Bernoulli Additive	$\log\{\eta/(1-\eta)\}$	η	0	no
Bernoulli Multiplicative	$\log\{e^{\eta}/(1-e^{\eta})\}$	e^{η}	0	no
Poisson Multiplicative	η	e^{η}	0	yes
Cox Model*	-	-	$\lvert\alpha^*\rvert<\lvert\alpha\rvert$	no
Paired Survival*	-	-	$\lvert\alpha^*\rvert<\lvert\alpha\rvert$	no

[†]The scale factor $K(\emptyset)$ equals one in all these models except the normal, for which $\{K(\emptyset)\}^{-1} = \sigma^2$, the conditional variance. The term "bias" describes whether the estimate, $\hat{\alpha}^*$, with X omitted converges to the true treatment effect, α.

*The theory requires some modification because conditional or partial likelihoods are used.

We estimate the treatment effect α under the correct model, $\eta=\mu+T\alpha+X\beta$, and under the false model, $\eta^*=\mu^*+T\alpha^*$ with X omitted. We shall be interested in the relationship in large samples between "estimates" $\hat{\alpha}$ and $\hat{\alpha}^*$, and the quantities α and α^*, respectively. We say $\hat{\alpha}^*$ is a "valid" estimate if it is asymptotically unbiased, namely $\alpha^*=\alpha$. The estimators $\hat{\alpha}$ and $\hat{\alpha}^*$ are maximum likelihood estimates for models like (2.3) or maximum conditional or partial likelihood estimates for paired survival data or the Cox model. The condition that $\hat{\alpha}^*$ be "valid" is equivalent to the non-confounding specifications, NC2. The main results on bias, for independent X and T, are summarized as follows:

1. If $\alpha=0$, then $\alpha^*=0$, no matter what model is used.
2. Condition NC2 holds for uncensored data if and only if $h(\eta)=\eta$ or $h(\eta)=\exp(\eta)$. In other words, only additive and multiplicative regression models yield valid estimates $\hat{\alpha}^*$ when X is omitted.
3. For the family (2.3), the approximate asymptotic bias is given by
$$\alpha^*-\alpha = (Q/4)\{h''(\mu+\alpha)/h'(\mu+\alpha)-h''(\mu-\alpha)/h'(\mu-\alpha)\} \quad (2.4)$$
where $Q=\beta^2 \mathrm{var}(X)$.
4. For randomly censored survival data with hazard proportional to $\exp(\eta)$, parametric models with known nuisance hazards yield conservative estimates, $|\alpha^*|<|\alpha|$, as does the Cox partial likelihood analysis.

We now consider hypothesis testing. Under the complete model, the one-sided score test for $\alpha=0$ is

$$U(n\hat{V})^{-1/2} > c \quad ,$$

where
$$U = \sum T\gamma'(\hat{\eta}_o)\{Y-h(\hat{\eta}_o)\} \quad , \qquad (2.5)$$
$$\hat{V} = \{nK(\emptyset)\}^{-1} \sum \gamma'(\hat{\eta}_o)h'(\hat{\eta}_o) \quad ,$$

summations are over the n subjects under study, and $\hat{\eta}_o=\hat{\mu}_o+X\hat{\beta}_o$ is the maximum likelihood estimate of η under the hypothesis $\alpha=0$. The variance estimate $\hat{V}$ is n^{-1} times the observed information calculated from the second derivative of the log-likelihood.

Under the false model with X omitted, (2.5) is replaced by

$$U^*(n\hat{V}^*)^{-1/2} > c \qquad (2.6)$$

where $\hat{\eta}_o^*=\hat{\mu}^*$. As discussed in [13], the model-based variance estimate $\hat{V}^* = \{nK(\emptyset)\}^{-1} \sum \gamma'(\hat{\eta}_o^*)h'(\hat{\eta}_o^*)$ is inconsistent, and asymptotically it differs from the true variance of $U^* n^{-1/2}$ by a "variance deflation factor", $k\neq 1$. Thus, omitting X may lead to an anticonservative significance test if the model-based variance estimate $\hat{V}^*$