BIOMETRICS 55, 252–257
March 1999

SHORTER COMMUNICATIONS

EDITOR:
LOUISE M. RYAN

Poisson Regression with Missing Durations of Exposure

D. Y. Lin,[1,2,*] P. Arbogast,[1,2] D. S. Siscovick,[2,3] and R. N. Lemaitre[2]

[1]Department of Biostatistics, Box 357232, University of Washington, Seattle, Washington 98195, U.S.A.
[2]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, 1730 Minor Avenue, Suite 1360, Seattle, Washington 98101, U.S.A.
[3]Department of Epidemiology, Box 357236, University of Washington, Seattle, Washington 98195, U.S.A.
*email: danyu@biostat.washington.edu

SUMMARY. In this paper, we develop Poisson-type regression methods that require the durations of exposure be measured only on a possibly nonrandom subset of the cohort members. These methods can be used to make inferences about the incidence density during exposure as well as the ratio of incidence densities during exposure versus not during exposure. Numerical studies demonstrate that the proposed methods yield reliable results in practical settings. We describe an application to a population-based case-control study assessing the transient increase in the risk of primary cardiac arrest during leisure-time physical activity.

KEY WORDS: Case-control study; Cohort study; Epidemiology; Incidence density; Relative risk; Retrospective study.

1. Introduction

Incidence density is a fundamental measure in epidemiology. It is defined as the number of new episodes of illness per unit (say 1000 years) of person-time of exposure. (Incidence density is also called incidence rate, though the latter is used in broader contexts.) Subjects are often exposed to the risk factor of interest for varying amounts of time rather than over the entire study period. Nonetheless, if the durations of exposure are measured on all cohort members, then Poisson regression can be used to estimate the incidence density and to assess its associations with covariates (Breslow and Day, 1987, Chapter 4). In many applications, especially in large-scale retrospective studies of rare diseases, it is unfortunately infeasible to measure the durations of exposure on all cohort members. The objective of this paper is to develop Poisson-type regression methods for situations in which the durations of exposure are measured only on a subset of cohort members.

This work was motivated by a population-based case-control study, to be referred to as CABS (Siscovick et al., 1995). From reports filled out by paramedics, all cases of out-of-hospital primary cardiac arrest (PCA) attended by paramedics were identified in the King County of the State of Washington during the period of October 1988 through July 1994. The study excluded the subjects with a history of clinically recognized heart disease or life-threatening comorbidity and further restricted the subjects to married residents aged 25 to 74 years (Siscovick et al., 1995). Control subjects matched for age (within 7 years) and gender were selected from the community by random-digit dialing. Detailed information about the cases and controls was collected through in-person interviews with the spouses of the study subjects. One objective of this study was to estimate and compare the incidence densities of PCA during and not during leisure-time physical activity (LTPA). Because the amounts of time spent in LTPA were measured only on the subjects in the case-control sample, which was not a random sample from the study population, it would not even be possible to estimate the overall incidence density (without covariate adjustment) using the existing Poisson methodology.

In the next section, we develop some simple methods for Poisson-type regression that only require that the durations of exposure be available on a possibly nonrandom subset of all cohort members. These methods allow one to make inferences about the incidence density during exposure as well as the ratio of the incidence densities during exposure versus not during exposure. In Section 3, we demonstrate through numerical studies that the proposed methods yield reliable results in practical applications. In Section 4, we illustrate the proposed methods with the aforementioned CABS study. Some concluding remarks are provided in Section 5.

2. Methods

Let T be the duration of exposure, D be the number of episodes of illness during exposure, and $\mathbf{Z}$ be a set of covariates including one. The Poisson regression model specifies that, conditional on T and $\mathbf{Z}$, the number of episodes D has a Poisson distribution with mean $Te^{\boldsymbol{\beta}'\mathbf{Z}}$, i.e.,

$$\mathrm{E}(D \mid T, \mathbf{Z}) = Te^{\boldsymbol{\beta}'\mathbf{Z}}, \tag{2.1}$$

where $\boldsymbol{\beta}$ is a set of unknown regression parameters. This model reduces to the exponential model if the event is terminal so that D is either zero or one.

Suppose that we have a cohort of n subjects that is a random sample from an infinite population. For $i = 1, \ldots, n$, let $(D_i, T_i, \mathbf{Z}_i)$ be the values of $(D, T, \mathbf{Z})$ on the ith subject. If the data $(D_i, T_i, \mathbf{Z}_i)$ $(i = 1, \ldots, n)$ are fully observed, then the likelihood score function for $\boldsymbol{\beta}$ is

$$\mathbf{U}^F(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(D_i - T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\right) \mathbf{Z}_i.$$

The solution to $\mathbf{U}^F(\boldsymbol{\beta}) = \mathbf{0}$, denoted by $\hat{\boldsymbol{\beta}}^F$, is consistent and asymptotically normal with covariance matrix estimator $-\{\partial \mathbf{U}^F(\hat{\boldsymbol{\beta}}^F)/\partial \boldsymbol{\beta}\}^{-1}$ (Breslow and Day, 1987, Chapter 4).

In epidemiological studies, it is often difficult to collect information on the duration of exposure for every cohort member. Suppose that the durations of exposure are only measured on the subjects in a subcohort of size $\tilde{n}$ $(\tilde{n} < n)$. The subcohort needs not be a random sample from the whole cohort, but the selection probabilities are assumed to be known. Let ξ_i indicate, by the values one versus zero, whether or not the ith subject is included in the subcohort so that his duration of exposure T_i is measured. We allow ξ_i to depend on D_i, T_i, and $\mathbf{Z}_i$. Denote

$$\pi_i = \Pr(\xi_i = 1 \mid D_i, T_i, \mathbf{Z}_i). \tag{2.2}$$

Clearly, $\tilde{n} = \Sigma_{i=1}^{n} \xi_i$. We assume that $\pi_i > 0$ for all i.

For studies that do not use a cohort design, the definitions of cohort and subcohort require some attention. In particular, for a population-based case-control study such as CABS, we regard the finite study population as a random sample from an infinite superpopulation, i.e., the joint distribution of $(D, T, \mathbf{Z})$, while regarding the case-control sample as the subcohort. If all the cases in the study population are selected into the case-control sample, then the cohort of interest is the whole study population; if half of the cases in the study population are randomly selected into the case-control sample, then the cohort of interest is a random half of the whole study population. With these definitions, the ξ_i's equal one for the subjects in the case-control sample and equal zero for the controls not selected; the π_i's are one for the cases and are typically very small for the controls.

Given the incomplete data, we propose estimating $\boldsymbol{\beta}$ by the following estimating function:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(D_i - \frac{\xi_i}{\pi_i} T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\right) \mathbf{Z}_i. \tag{2.3}$$

The function (2.3) involves the durations of exposure from the subcohort members only, but their contributions are weighted inversely by their selection probabilities. It follows from (2.2) and (2.1) that $\mathrm{E}\{\mathbf{U}(\boldsymbol{\beta})\} = \mathrm{EE}\{\mathbf{U}(\boldsymbol{\beta}) \mid D_i, T_i, \mathbf{Z}_i; i = 1, \ldots, n\} = \mathrm{E}\left[\mathrm{E}\{\mathbf{U}^F(\boldsymbol{\beta}) \mid T_i, \mathbf{Z}_i; i = 1, \ldots, n\}\right] = \mathbf{0}$. Thus, $\mathbf{U}(\boldsymbol{\beta})$ is an unbiased estimating function. Denote the solution to $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$ by $\hat{\boldsymbol{\beta}}$.

Write $\hat{\mathbf{A}}(\boldsymbol{\beta}) = -n^{-1}\partial \mathbf{U}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$. Then

$$\hat{\mathbf{A}}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i} \mathbf{Z}_i \mathbf{Z}_i',$$

which is positive semidefinite. By the law of large numbers, $\hat{\mathbf{A}}(\boldsymbol{\beta})$ converges in probability to $\mathbf{A}(\boldsymbol{\beta}) \equiv \mathrm{E}(Te^{\boldsymbol{\beta}'\mathbf{Z}}\mathbf{Z}\mathbf{Z}')$. Assume that $\mathbf{A}(\boldsymbol{\beta})$ is nonsingular and consequently positive definite, which implies that $\hat{\mathbf{A}}(\boldsymbol{\beta})$ is positive definite at least for large n. Therefore, $\hat{\boldsymbol{\beta}}$ is unique at least for large n and converges in probability to $\boldsymbol{\beta}$ as $n \to \infty$.

In order to obtain the asymptotic distribution for $\hat{\boldsymbol{\beta}}$, we need to derive the asymptotic distribution for $\mathbf{U}(\boldsymbol{\beta})$. The latter depends on how the subcohort is chosen. We consider two common sampling schemes, (I) independent Bernoulli sampling and (II) stratified simple random sampling. Under the first sampling scheme, the subjects are selected into the subcohort independently so that the ξ_i's are independent Bernoulli random variables with success probabilities π_i's; the subcohort size $\tilde{n}$ is random. Under the second sampling scheme, the cohort may be stratified by some variables such as case-control status, age, and gender, and a simple random sample of a fixed size is drawn without replacement within each stratum. The latter scheme is routinely used in case-control studies, including CABS; the unstratified simple random sampling is a special case. The key feature of this sampling scheme is that, whether there is one or multiple strata, the ξ_i's are no longer independent due to the fixed stratum sizes in the subcohort.

Under independent Bernoulli sampling, $\mathbf{U}(\boldsymbol{\beta})$ is a sum of n zero-mean random vectors. Thus, by the multivariate central limit theorem, $n^{-1/2}\mathbf{U}(\boldsymbol{\beta})$ is asymptotically zero-mean normal with covariance matrix

$$\mathbf{B}_I(\boldsymbol{\beta}) \equiv \mathrm{E}\left(D - \frac{\xi}{\pi} Te^{\boldsymbol{\beta}'\mathbf{Z}}\right)^2 \mathbf{Z}\mathbf{Z}'.$$

It follows from the law of large numbers and the consistency of $\hat{\boldsymbol{\beta}}$ that $\mathbf{B}_I(\boldsymbol{\beta})$ can be consistently estimated by $\hat{\mathbf{B}}_I(\hat{\boldsymbol{\beta}})$, where

$$\hat{\mathbf{B}}_I(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \left(D_i - \frac{\xi_i}{\pi_i} T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\right)^2 \mathbf{Z}_i \mathbf{Z}_i'.$$

The derivation of the asymptotic distribution for $\mathbf{U}(\boldsymbol{\beta})$ under stratified simple random sampling is quite delicate due to the dependence of the ξ_i's (it is given in Appendix 1). To describe this distribution, we need to introduce some notation. Suppose that the subjects are classified into J strata,

$S_1, \ldots, S_J$, the jth stratum having n_j and $\tilde{n}_j$ subjects in the whole cohort and subcohort, respectively. Let $f_j = \tilde{n}_j/n_j$. Then $n^{-1/2}\mathbf{U}(\boldsymbol{\beta})$ is asymptotically zero-mean normal with a covariance matrix that can be consistently estimated by $\hat{\mathbf{B}}_{II}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{B}}_I(\hat{\boldsymbol{\beta}}) - \boldsymbol{\delta}(\hat{\boldsymbol{\beta}})$, where

$$\boldsymbol{\delta}(\boldsymbol{\beta}) = n^{-1}\sum_{j=1}^{J}\frac{1-f_j}{\tilde{n}_j f_j^2}\left(\sum_{i\in S_j}\xi_i T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i\right) \times \left(\sum_{i\in S_j}\xi_i T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i\right)'.$$

Given the asymptotic normality of $n^{-1/2}\mathbf{U}(\boldsymbol{\beta})$, we can use the Taylor series expansion to show that $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically zero-mean normal with a covariance matrix that can be consistently estimated by $\hat{\mathbf{V}} \equiv \hat{\mathbf{A}}^{-1}(\hat{\boldsymbol{\beta}})\hat{\mathbf{B}}(\hat{\boldsymbol{\beta}})\hat{\mathbf{A}}^{-1}(\hat{\boldsymbol{\beta}})$, where $\hat{\mathbf{B}}(\boldsymbol{\beta})$ corresponds to $\hat{\mathbf{B}}_I(\boldsymbol{\beta})$ and $\hat{\mathbf{B}}_{II}(\boldsymbol{\beta})$ under sampling schemes I and II, respectively. Since $\boldsymbol{\delta}(\boldsymbol{\beta})$ is nonnegative, the variance of $\hat{\boldsymbol{\beta}}$ is smaller under scheme II than under scheme I.

Let us examine the special case of estimating the overall incidence density without adjusting for any covariate, i.e., $Z = 1$. Write $\lambda = e^{\beta}$ and $\hat{\lambda} = e^{\hat{\beta}}$. Clearly, $U(\beta) = 0$ has the explicit solution

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} D_i}{\sum_{i=1}^{n}(\xi_i/\pi_i)T_i},$$

which is an intuitive estimator for λ in view of (2.1). In the presence of missing durations of exposure, the variance of $\hat{\lambda}$ is not obvious but follows easily from the general theory developed above. Specifically, a consistent variance estimator for $\log\hat{\lambda}$, i.e., $\hat{\beta}$, is $n^{-1}\hat{B}(\hat{\beta})/\hat{A}^2(\hat{\beta})$, where $\hat{A} = \Sigma_{i=1}^{n} D_i$ and $\hat{B}$ depends on the sampling scheme.

In certain applications, including CABS, one is interested in comparing the incidence densities during exposure versus during nonexposure. To this end, suppose that the incidence density during nonexposure can also be formulated by a Poisson regression model. Let $\bar{T}$ be the duration of nonexposure, and let $\bar{D}$ be the number of episodes of illness during nonexposure. Then

$$\mathrm{E}(\bar{D} \mid \bar{T}, \mathbf{Z}) = \bar{T}e^{\boldsymbol{\beta}_u'\mathbf{Z}}, \tag{2.4}$$

where $\boldsymbol{\beta}_u$ is a set of unknown regression parameters.

For $i = 1, \ldots, n$, let $\bar{T}_i$ and $\bar{D}_i$ be the values of $\bar{T}$ and $\bar{D}$ on the ith subject. Since $\bar{T}_i = \tau - T_i$, where τ is the length of the study period, the value of $\bar{T}_i$ is known as long as T_i is. In analogy to (2.3), we estimate $\boldsymbol{\beta}_u$ by the estimating function

$$\mathbf{U}_u(\boldsymbol{\beta}_u) = \sum_{i=1}^{n}\left(\bar{D}_i - \frac{\xi_i}{\pi_i}\bar{T}_i e^{\boldsymbol{\beta}_u'\mathbf{Z}_i}\right)\mathbf{Z}_i.$$

Denote the solution to $\mathbf{U}_u(\boldsymbol{\beta}_u) = \mathbf{0}$ by $\hat{\boldsymbol{\beta}}_u$. The random vector $n^{1/2}(\hat{\boldsymbol{\beta}}_u - \boldsymbol{\beta}_u)$ is asymptotically zero-mean normal with a covariance matrix that can be consistently estimated by $\hat{\mathbf{V}}_u \equiv \hat{\mathbf{A}}_u^{-1}(\hat{\boldsymbol{\beta}}_u)\hat{\mathbf{B}}_u(\hat{\boldsymbol{\beta}}_u)\hat{\mathbf{A}}_u^{-1}(\hat{\boldsymbol{\beta}}_u)$, where $\hat{\mathbf{A}}_u$ and $\hat{\mathbf{B}}_u$ are analogous to $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$.

Since we are interested in comparing $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_u$, it is necessary to derive the asymptotic distribution for $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_u$. This derivation is complicated by the fact that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_u$ are calculated from the same set of subjects and are therefore correlated. In Appendix 2, we prove that $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ and $n^{1/2}(\hat{\boldsymbol{\beta}}_u-\boldsymbol{\beta}_u)$ are asymptotically joint normal and the covariance matrix between these two random vectors can be consistently estimated by $\hat{\mathbf{C}} \equiv \hat{\mathbf{A}}^{-1}(\hat{\boldsymbol{\beta}})\hat{\mathbf{B}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}_u)\hat{\mathbf{A}}_u^{-1}(\hat{\boldsymbol{\beta}}_u)$, where $\hat{\mathbf{B}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}_u)$ equals $\hat{\mathbf{B}}_I(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}_u)$ and $\hat{\mathbf{B}}_I(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}_u) - \boldsymbol{\delta}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}_u)$ under sampling schemes I and II, respectively, and

$$\hat{\mathbf{B}}_I(\boldsymbol{\beta},\boldsymbol{\beta}_u) = n^{-1}\sum_{i=1}^{n}\left(D_i - \frac{\xi_i}{\pi_i}T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\right) \times \left(\bar{D}_i - \frac{\xi_i}{\pi_i}\bar{T}_i e^{\boldsymbol{\beta}_u'\mathbf{Z}_i}\right)\mathbf{Z}_i\mathbf{Z}_i',$$

$$\boldsymbol{\delta}(\boldsymbol{\beta},\boldsymbol{\beta}_u) = n^{-1}\sum_{j=1}^{J}\frac{1-f_j}{\tilde{n}_j f_j^2}\left(\sum_{i\in S_j}\xi_i T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i\right) \times \left(\sum_{i\in S_j}\xi_i \bar{T}_i e^{\boldsymbol{\beta}_u'\mathbf{Z}_i}\mathbf{Z}_i\right)'.$$

Thus, $n^{1/2}\{(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_u)-(\boldsymbol{\beta}-\boldsymbol{\beta}_u)\}$ is asymptotically zero-mean normal with a covariance matrix that can be consistently estimated by $\hat{\mathbf{V}} + \hat{\mathbf{V}}_u - \hat{\mathbf{C}} - \hat{\mathbf{C}}'$.

The definition of $\mathbf{U}(\boldsymbol{\beta})$ implicitly assumes that $\mathbf{Z}_i$ is completely measured if $D_i > 0$ or T_i is known. This assumption is trivially satisfied if one is only interested in estimating the overall incidence density, in which case $\mathbf{Z}_i = 1$ for all i. This assumption is also satisfied in case-control studies that collect information on the duration of exposure and covariates from all the subjects in the case-control sample. To relax this assumption, we let ϕ_i indicate, by the values one versus zero, whether or not $\mathbf{Z}_i$ is measured and redefine ξ_i as the indicator for whether or not both T_i and $\mathbf{Z}_i$ are available. Also, define $\rho_i = \Pr(\phi_i = 1 \mid D_i, T_i, \mathbf{Z}_i)$. Then we can generalize (2.3) as follows:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(\frac{\phi_i}{\rho_i}D_i - \frac{\xi_i}{\pi_i}T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\right)\mathbf{Z}_i. \tag{2.5}$$

The resulting estimator, again denoted by $\hat{\boldsymbol{\beta}}$, remains consistent and asymptotically normal with an estimated covariance matrix of the form $\hat{\mathbf{A}}^{-1}(\hat{\boldsymbol{\beta}})\hat{\mathbf{B}}(\hat{\boldsymbol{\beta}})\hat{\mathbf{A}}^{-1}(\hat{\boldsymbol{\beta}})$. The exact form of $\hat{\mathbf{B}}$ depends on the sampling scheme. If (ξ_i, ϕ_i) $(i = 1, \ldots, n)$ are independent, then

$$\hat{\mathbf{B}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\left(\frac{\phi_i}{\rho_i}D_i - \frac{\xi_i}{\pi_i}T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\right)^2\mathbf{Z}_i\mathbf{Z}_i'.$$

By interpreting ϕ_i as the indicator of whether or not both $\mathbf{Z}_i$ and D_i are known, we may use (2.5) to make inferences about $\boldsymbol{\beta}$ even if the cases are selected with unequal probabilities. Similar extensions can be made for drawing inferences about $\boldsymbol{\beta}_u$ and $\boldsymbol{\beta} - \boldsymbol{\beta}_u$.

3. Numerical Studies

We conducted extensive simulation studies to evaluate the asymptotic approximations used in the previous section and found that the approximations are very accurate for practical use. For example, in one simulation experiment mimicking the population-based case-control study, we considered model

Table 1
Risk of primary cardiac arrest during leisure-time physical activity

	Incidence densities[a]		
	During LTPA	Not during LTPA	Incidence ratio[b]
Without covariate adjustment	13.8 (10.0, 19.1)	1.4 (1.3, 1.6)	9.6 (6.8, 13.6)
With covariate adjustment[c]	17.1 (9.4, 31.0)	1.5 (1.1, 2.0)	11.6 (6.1, 22.4)

Note: The 95% confidence intervals are shown in parentheses.

[a] The number of PCA per 10^8 person-hours of exposure.

[b] The incidence density during LTPA divided by the incidence density not during LTPA.

[c] The incidence densities shown pertain to a male nonsmoker aged 60 with high-school education and free of diabetes and hypertension.

(2.1) in which $Z_1 = 1$, Z_2 is a Bernoulli variable with 0.5 success probability, and Z_3 is another Bernoulli variable with 0.4 success probability if $Z_2 = 0$ and with 0.6 success probability if $Z_2 = 1$; we set $\boldsymbol{\beta} = (-5, 0.5, 0.5)'$. We let T be uniformly distributed over $[0, 0.1]$ and generated 100,000 subjects from this model. We obtained 1000 such samples. On average, there was 58 cases per sample. In each sample, we selected all the cases; the controls were so selected that the numbers of controls with $Z_2 = 0$ and with $Z_2 = 1$ were the same as the corresponding numbers of cases. With such small numbers of cases and controls, i.e., less than 60, the asymptotic approximations were found to be adequate. Specifically, the average of the $\hat{\boldsymbol{\beta}}$ estimates over the 1000 simulation samples was $(-5.02, 0.50, 0.51)'$, which is remarkably close to the true $\boldsymbol{\beta}$; the sampling standard error of $\hat{\boldsymbol{\beta}}$ was $(0.40, 0.36, 0.43)'$, while the average of the standard error estimates was $(0.39, 0.35, 0.43)'$.

4. Applications to the CABS Study

We now apply the proposed methods to the CABS study. The case-control sample contains data on 340 cases and 568 controls. Out of the 340 cases, 54 occurred during LTPA. As mentioned previously, this study was stratified on age group and gender and excluded unmarried subjects as well as those who had prior clinical heart disease and major comorbidity. We used the 1990 census data for King County to determine the number of married people in each age group × gender stratum and multiplied it by 0.78, the response rate for the cases, and by 0.90, an estimated proportion of the population without prior clinical heart disease and major comorbidity. The resulting numbers were then compared to the corresponding numbers of controls in the case-control sample to calculate the selection probabilities for the controls.

By applying the methods of Section 2 to the case-control data on the number of PCA, duration of LTPA, and covariates, along with the aforementioned selection probabilities, we obtained the results of Table 1. There was roughly a 10-fold increase in the incidence of PCA during LTPA relative to not during LTPA, with or without the adjustment of covariates including age, gender, smoking, education, diabetes, and hypertension. Results not included in Table 1 showed that males, older subjects, smokers, and subjects with diabetes or hypertension were significantly associated with increased risks of PCA both during and not during LTPA. In addition, applying the standard logistic regression method to the case-control data, we found that the relative risk of PCA for the subjects who engaged in LTPA relative to those who did not engage in LTPA was 0.34, with a 95% confidence interval of $(0.19, 0.61)$, after adjusting for the aforementioned covariates. These results support the hypothesis that the risk of PCA is transiently increased during LTPA, although LTPA is associated with an overall reduction in the risk of PCA.

5. Discussion

The formulation in this paper allows multiple episodes of illness from the same subject. Although we used the Poisson framework, our derivations only made use of the mean structures for D and $\bar{D}$ given in (2.1) and (2.4). Thus, the proposed methods are valid even if the repeated episodes of illness from the same subject are correlated.

The methods described in Section 2 pertain to the incidence density over the entire study period. If one is interested in the incidence densities over finer time intervals, say K subintervals of the study period, then it is natural to formulate the incidence density in each of the K intervals with a model of form (2.1). For parsimoniousness, we impose a common set of slope parameters over the K intervals while allowing a separate intercept for each interval. We then combine the estimating functions of the form (2.3) from the K intervals to make inferences about the regression parameters.

The proposed estimators weight the individuals' contributions by the inverses of their selection probabilities. This is a very old idea in survey sampling attributed to Horvitz and Thompson (see Cochran, 1977, pp. 259–261). Recent years have seen innovative applications of this idea to various other missing-data problems. In particular, weighted likelihood estimators have been developed for case-cohort and two-stage case-control designs (Prentice, 1986; Kalbfleisch and Lawless, 1988; Flanders and Greenland, 1991; Breslow and Holubkov, 1997). Those designs deal with the problem of missing covariate data, whereas this work is mainly concerned with incomplete measurements on the durations of exposures. Furthermore, we have shown how to make formal inferences about the ratio of incidence densities during exposure versus not during exposure. Such methods are not currently available even if the durations of exposure are measured on all the cohort members.

Acknowledgements

The authors are grateful to the editor and associate editor for their rapid and careful reviews of the paper. This research was supported by the National Institutes of Health.

Résumé

Nous développons, dans cet article, des méthodes de régression de type poisson qui nécessitent que les durées d'exposition soient mesurées seulement sur un sous-ensemble probablement non aléatoire des membres de la cohorte. Ces méthodes peuvent être utilisées pour faire de l'inférence sur la densité d'incidence durant l'exposition ainsi que sur les ratios des

densités d'incidence pendant la période d'exposition par rapport à la période de non exposition. Des études numériques démontrent que les méthodes proposées fournissent des résultats fiables dans des situations pratiques. Nous décrivons une application à partir d'une étude cas-témoin évaluant l'accroissement passager du risque d'un premier arrêt cardiaque pendant les activités de loisir.

References

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research,* Volume 2, *The Design and Analysis of Cohort Studies.* Lyon: International Agency for Research on Cancer.

Breslow, N. E. and Holubkov, R. (1997). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic analysis of two-stage data. *Statistics in Medicine* **16,** 103–116.

Cochran, W. G. (1977). *Sampling Techniques.* New York: Wiley.

Flanders, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* **10,** 739–747.

Kalbfleisch, J. D. and Lawless, J. F. (1988). Likelihood analysis of multi-stage models for disease incidence and mortality. *Statistics in Medicine* **7,** 149–160.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73,** 1–11.

Siscovick, D. S., Raghunathan, T. E., King, I., et al. (1995). Dietary intake and cell membrane levels of long-chain n-3 polyunsaturated fatty acids and the risk of primary cardiac arrest. *Journal of the American Medical Association* **274,** 1363–1367.

Received December 1997. *Revised February* 1998. *Accepted March* 1998.

Appendix 1

Asymptotic Distribution of $n^{-1/2}\mathbf{U}(\boldsymbol{\beta})$ *Under Stratified Simple Random Sampling*

Clearly, $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{U}^F(\boldsymbol{\beta}) - \mathbf{U}^D(\boldsymbol{\beta})$, where

$$\mathbf{U}^D(\boldsymbol{\beta}) = \sum_{i=1}^{n} \pi_i^{-1}(\xi_i - \pi_i) T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i = \sum_{j=1}^{J}\sum_{i\in S_j} f_j^{-1}(\xi_i - f_j) T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i.$$

By the multivariate central limit theorem, $n^{-1/2}\mathbf{U}^F(\boldsymbol{\beta})$ converges in distribution to a zero-mean normal random vector with covariance matrix $\mathbf{B}^F(\boldsymbol{\beta}) \equiv \mathrm{E}\{(D - Te^{\boldsymbol{\beta}'\mathbf{Z}})^2\mathbf{Z}\mathbf{Z}'\}$. By extending Hajek's central limit theorem on sampling without replacement from a finite population as stated in Cochran (1977, pp. 39–40), we can show that, conditional on $(D_i, T_i, \mathbf{Z}_i)$ $(i = 1, \ldots, n)$, $n^{-1/2}\mathbf{U}^D(\boldsymbol{\beta})$ converges in distribution to a zero-mean normal random vector with covariance matrix

$$\mathbf{B}^D(\boldsymbol{\beta}) \equiv \lim_{n\to\infty} \sum_{j=1}^{J} \frac{(\tilde{n}_j/n)(1-f_j)}{f_j^2} \times \left\{ \frac{\sum_{i\in S_j} T_i^2 e^{2\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i\mathbf{Z}_i'}{n_j} - \left(\frac{\sum_{i\in S_j} T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i}{n_j}\right) \times \left(\frac{\sum_{i\in S_j} T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i'}{n_j}\right) \right\}.$$

The convergence of the distribution also holds unconditionally because, as a limit, $\mathbf{B}^D(\boldsymbol{\beta})$ is a deterministic matrix that does not depend on the actual values of $(D_i, T_i, \mathbf{Z}_i)$ $(i = 1, \ldots, n)$. Furthermore, (2.2) implies that $n^{-1/2}\mathbf{U}^F(\boldsymbol{\beta})$ and $n^{-1/2}\mathbf{U}^D(\boldsymbol{\beta})$ are uncorrelated and therefore asymptotically independent. Hence, $n^{-1/2}\mathbf{U}(\boldsymbol{\beta})$ converges in distribution to a zero-mean normal random vector with covariance matrix $\mathbf{B}(\boldsymbol{\beta}) \equiv \mathbf{B}^F(\boldsymbol{\beta}) + \mathbf{B}^D(\boldsymbol{\beta})$. We can estimate $\mathbf{B}^F(\boldsymbol{\beta})$ consistently by $\hat{\mathbf{B}}^F(\boldsymbol{\beta}) \equiv n^{-1}\Sigma_{i=1}^{n}(D_i^2 - 2\pi_i^{-1}\xi_i D_i T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i} + \pi_i^{-1}\xi_i T_i^2 e^{2\boldsymbol{\beta}'\mathbf{Z}_i})\mathbf{Z}_i\mathbf{Z}_i'$. In addition, we can estimate $\mathbf{B}^D(\boldsymbol{\beta})$ consistently by

$$\hat{\mathbf{B}}^D(\boldsymbol{\beta}) \equiv \sum_{j=1}^{J} \frac{(\tilde{n}_j/n)(1-f_j)}{f_j^2} \times \left\{ \frac{\sum_{i\in S_j} \xi_i T_i^2 e^{2\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i\mathbf{Z}_i'}{\tilde{n}_j} - \left(\frac{\sum_{i\in S_j} \xi_i T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i}{\tilde{n}_j}\right) \times \left(\frac{\sum_{i\in S_j} \xi_i T_i e^{\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i'}{\tilde{n}_j}\right) \right\},$$

which can be written as $n^{-1}\Sigma_{i=1}^{n}(\pi_i^{-2}\xi_i - \pi_i^{-1}\xi_i)T_i^2 e^{2\boldsymbol{\beta}'\mathbf{Z}_i} \times \mathbf{Z}_i\mathbf{Z}_i' - \boldsymbol{\delta}(\boldsymbol{\beta})$. It is easy to see that $\hat{\mathbf{B}}^F(\boldsymbol{\beta}) + \hat{\mathbf{B}}^D(\boldsymbol{\beta}) = \hat{\mathbf{B}}_I(\boldsymbol{\beta}) - \boldsymbol{\delta}(\boldsymbol{\beta})$. Therefore, $\mathbf{B}(\boldsymbol{\beta})$ can be consistently estimated by $\hat{\mathbf{B}}_I(\hat{\boldsymbol{\beta}}) - \boldsymbol{\delta}(\hat{\boldsymbol{\beta}})$.

Appendix 2

Asymptotic Joint Distribution for $\hat{\boldsymbol{\beta}}$ *and* $\hat{\boldsymbol{\beta}}_u$

Let $\mathbf{A}_u(\boldsymbol{\beta}_u)$ and $\mathbf{B}_u(\boldsymbol{\beta}_u)$ be the limits of $\hat{\mathbf{A}}_u(\boldsymbol{\beta}_u)$ and $\hat{\mathbf{B}}_u(\boldsymbol{\beta}_u)$. By Taylor series expansions,

$$n^{1/2}\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\beta}}_u - \boldsymbol{\beta}_u \end{bmatrix} = n^{-1/2}\begin{bmatrix} \mathbf{A}^{-1}(\boldsymbol{\beta})\mathbf{U}(\boldsymbol{\beta}) \\ \mathbf{A}_u^{-1}(\boldsymbol{\beta}_u)\mathbf{U}_u(\boldsymbol{\beta}_u) \end{bmatrix} + o_p(1).$$

It then follows from the arguments of Section 2 and Appendix 1 that $n^{1/2}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})', (\hat{\boldsymbol{\beta}}_u-\boldsymbol{\beta}_u)']'$ converges in distribution to a zero-mean normal random vector with joint covariance matrix

$$\begin{bmatrix} \mathbf{A}^{-1}(\boldsymbol{\beta})\mathbf{B}(\boldsymbol{\beta})\mathbf{A}^{-1}(\boldsymbol{\beta}) & \mathbf{A}^{-1}(\boldsymbol{\beta})\mathbf{B}(\boldsymbol{\beta},\boldsymbol{\beta}_u)\mathbf{A}_u^{-1}(\boldsymbol{\beta}_u) \\ \mathbf{A}_u^{-1}(\boldsymbol{\beta}_u)\mathbf{B}'(\boldsymbol{\beta},\boldsymbol{\beta}_u)\mathbf{A}^{-1}(\boldsymbol{\beta}) & \mathbf{A}_u^{-1}(\boldsymbol{\beta}_u)\mathbf{B}_u(\boldsymbol{\beta}_u)\mathbf{A}_u^{-1}(\boldsymbol{\beta}_u) \end{bmatrix},$$

where $\mathbf{B}(\boldsymbol{\beta},\boldsymbol{\beta}_u)$ is the limiting covariance matrix between $n^{-1/2}\mathbf{U}(\boldsymbol{\beta})$ and $n^{-1/2}\mathbf{U}_u(\boldsymbol{\beta}_u)$. For independent Benoulli sampling, $\mathbf{B}(\boldsymbol{\beta},\boldsymbol{\beta}_u) = \mathrm{E}\{(D-\pi^{-1}\xi Te^{\boldsymbol{\beta}'\mathbf{Z}})(\bar{D}-\pi^{-1}\xi\bar{T}e^{\boldsymbol{\beta}_u'\mathbf{Z}})\mathbf{Z}\mathbf{Z}'\}$; for stratified simple random sampling, $\mathbf{B}(\boldsymbol{\beta},\boldsymbol{\beta}_u) = \mathbf{B}^F(\boldsymbol{\beta},\boldsymbol{\beta}_u) + \mathbf{B}^D(\boldsymbol{\beta},\boldsymbol{\beta}_u)$, where $\mathbf{B}^F(\boldsymbol{\beta},\boldsymbol{\beta}_u) = \mathrm{E}\{(D-Te^{\boldsymbol{\beta}'\mathbf{Z}})(\bar{D}-\bar{T}e^{\boldsymbol{\beta}_u'\mathbf{Z}}) \times \mathbf{Z}\mathbf{Z}'\}$ and

$$\mathbf{B}^D(\boldsymbol{\beta},\boldsymbol{\beta}_u) = \lim_{n\to\infty}\sum_{j=1}^{J}\frac{(\tilde{n}_j/n)(1-f_j)}{f_j^2}$$

$$\times\left\{\frac{\sum_{i\in S_j}T_i\bar{T}_ie^{(\boldsymbol{\beta}+\boldsymbol{\beta}_u)'\mathbf{Z}_i}\mathbf{Z}_i\mathbf{Z}_i'}{n_j} - \left(\frac{\sum_{i\in S_j}T_ie^{\boldsymbol{\beta}'\mathbf{Z}_i}\mathbf{Z}_i}{n_j}\right) \times\left(\frac{\sum_{i\in S_j}\bar{T}_ie^{\boldsymbol{\beta}_u'\mathbf{Z}_i}\mathbf{Z}_i'}{n_j}\right)\right\}.$$

By replacing the unknown quantities in $\mathbf{B}(\boldsymbol{\beta},\boldsymbol{\beta}_u)$ with appropriate sample estimators in the same manner as with the case of $\mathbf{B}(\boldsymbol{\beta})$, we can show that the resulting covariance matrix estimator is indeed $\hat{\mathbf{B}}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}_u)$ described in Section 2.