

## Analysis of a Two-Stage Case-Control Study with Cluster Sampling of Controls: Application to Nonmelanoma Skin Cancer

Thomas R. Fears\* and Mitchell H. Gail

Biostatistics Branch, National Cancer Institute,  
Executive Plaza South, Room 8040, 6120 Executive Boulevard, MSC 7244,  
Rockville, Maryland 20892-7244, U.S.A.

\* email: fearst@exchange.nih.gov

**SUMMARY.** We present a pseudolikelihood approach for analyzing a two-stage population-based case-control study with cluster sampling of controls. These methods were developed to analyze data from a study of nonmelanoma skin cancer (NMSC). This study was designed to evaluate the role of ultraviolet radiation (UVB) on NMSC risk while adjusting for age group, which is known for all subjects, and for other individual-level factors, such as susceptibility to sunburn, which are known only for participants in the case-control study. The methods presented yield estimates of relative and absolute risk, with standard errors, while accounting naturally for the two-stage sampling of the cohort and cluster sampling of controls.

**KEY WORDS:** Case-control study; Population-based case-control study; Pseudolikelihood; Skin cancer; Survey sampling; Two-phase case-control study; Ultraviolet radiation.

### 1. Introduction

In this paper, we analyze population-based, two-stage case-control data with cluster sampling of controls. The methods we present were motivated by the following study of the effects of ultraviolet radiation on the risk of nonmelanoma skin cancer (NMSC).

Short wavelength ultraviolet radiation (UVB) is known to cause nonmelanoma skin cancer. An ozone layer surrounds the earth and filters out most short wavelength ultraviolet radiation (UVB) before it can reach the earth's surface. The thickness of the layer and the effective angle of incidence vary with latitude in such a way that the amount of UVB reaching the earth's surface also varies with latitude. The farther a location is from the equator the smaller the total amount of damaging UVB that reaches that location.

In order to study the effects of UVB on the risk of nonmelanoma skin cancer while controlling for other factors like skin color, the National Cancer Institute (NCI) and the Environmental Protection Agency (EPA) conducted a population-based case-control study in nine separate regions of the United States. Since nonmelanoma skin cancer is often treated without hospitalization, case ascertainment is difficult. Therefore, a special study was required to estimate the incidence of NMSC at each of these locations. For 1 year at each location, all dermatologists and physicians treating skin cancer, as well as all pathology laboratories and hospitals, were contacted to ascertain nonmelanoma skin cancer cases at that location. The level of UVB at each location was also measured (Scotto et al., 1988), and the regression of log cancer

incidence on  $\ln(\text{UVB})$  and age group was used to study the relationship between nonmelanoma incidence, age group, and annual UVB insolation (Scotto, Fears, and Fraumeni, 1981; Fears and Scotto, 1983).

In addition to the incidence surveys, case-control interview surveys were carried out at each of the nine locations to define individual risk factors in greater detail (Scotto, Fears, and Fraumeni, 1982). At each of the nine locations, the telephone interview survey designs were similar. Information was obtained on epidemiologic items of interest, including the individual's natural susceptibility, i.e., questions about eye color, hair color, ancestry or ethnic group, and skin complexion, and the individual's environmental susceptibility, i.e., questions about residence mobility, occupations, outdoor exposure habits, and exposure to materials that harm or protect the skin. A simple random sample without replacement (SRS) of 450 cases was selected from the cases aged 20-74 ascertained in the incidence study, and an additional 50 cases (a further SRS) were selected from the younger cases aged 20-49. Physician permission was required before selected patients were contacted, resulting in low response rates at some locations.

To obtain the controls, a general population survey (GPS) was conducted at each location. At each location, at least 500 households were sampled using a two-stage random digit dialing (RDD) cluster sample (Waksburg, 1978). In the first stage of each RDD sample, 100 clusters of 100 telephone numbers were selected based on the first five digits of seven-digit telephone numbers available for use at a location. Clusters were selected with probability proportional to the number of

households in the cluster. Telephone numbers within a selected cluster were then sampled until five households were identified. An attempt was made to interview all adults aged 20-74 in each selected household.

If the RDD sample is self-weighting, the sample means are unbiased estimates. Even so, the association of telephone exchanges with neighborhoods is a source of intraclass correlation. Such correlations tend to increase variance compared with SRS because clusters usually have less information. The extent of the increase in the variance of logistic parameter estimates (see Section 3) is diluted, to some extent, by the fact that cases are obtained by simple random sampling (Graubard, Fears, and Gail, 1989).

In addition to the general RDD samples, supplemental samples in the RDD clusters focused only on adults aged 65-74. A second supplemental sample of those aged 65-74 was obtained from an available Health Care Financing Administration (HCFA) file, and it was regarded as an SRS from the general population aged 65-74. The control sample thus consisted of a two-stage cluster sample with two supplemental samples, one RDD and the other SRS, taken to oversample the oldest age group. The numbers of interviews, after post-stratification by age group, are given in Table 1 for both cases and general population samples and sampling methodology.

This study can be regarded as a two-stage case-control design. In the first stage, all individuals in the nine regions are classified by disease status (incident case within the 1-year study period or noncase), gender, 5-year age group, and UVB exposure level (each region had its own exposure level). In the second stage, cases are sampled using SRS and controls are sampled using RDD. Second-stage sampling fractions for the case-control study depended on region, age stratum, and disease status. The case-control data yielded additional information on factors that influence the risk of NMSC. Because the entire population is classified in the first stage, this is a population-based case-control design and, consequently, we can estimate absolute exposure-specific risk as well as relative risk.

Published methods for analyzing two-stage case-control data by White (1982), Breslow and Cain (1988), and Flanders and Greenland (1991) assume that controls are obtained

in the second stage by simple random sampling. Graubard et al. (1989) developed modifications of the standard Mantel-Haenszel and Wolfe-Haldane methods for one-stage case-control studies with cluster sampling of controls. Modifications of logistic regression analysis that account for cluster sampling of controls have also been developed for one-stage case-control designs (Graubard and Gail, 1992; Graubard, Gail, and Brogan, unpublished manuscript), but these methods do not address the special features of the sampling employed in the NMSC study. In this paper, we employ a logistic model and extend the pseudolikelihood approach used by Flanders and Greenland to account for the complexity of the control sampling scheme; unlike Flanders and Greenland, we must account for cluster sampling of controls. The pseudolikelihood method allows us to exploit the population-based element in the sampling and to estimate absolute risk as well as relative risk. In the example we present, the cluster sampling is not self-weighting because some households have more than one telephone number. We use proper weights to obtain unbiased estimates of parameter effects. Ignoring the effects of clustering, but allowing (incorrectly) for the fact that cases and controls were selected by simple random sampling, yielded very slight underestimates of variance in this example. Nonetheless, we would always recommend appropriate adjustments for clustering, which may have larger effects on variance in other applications.

## 2. The Model and Pseudolikelihood Inference

### 2.1 The Model and Pseudolikelihood Approach

There are nine locations ( $l = 1, 2, \dots, 9$ ), and within each location, we consider three age strata ( $s = 1, 2, 3$ ), i.e., ages [20, 49], [50, 64], and [65, 74]. We will use  $\underline{X}$  to denote an individual-level covariate vector of dimension  $r_1 \times 1$  measured for all subjects at baseline. Each component of  $\underline{X}$  takes on only a finite number of values; thus,  $\underline{X}$  is a discrete vector random variable. A component of  $\underline{X}$  may take on many values for individuals within a category defined by location and age, as is the case for age, or take on the same value for all individuals in the stratum, as is the case for UVB insolation. We will use  $\underline{T}$  to denote a discrete covariate vector of dimension  $r_2 \times 1$  measured only in the interviews of selected

Table 1  
The population characteristics and number of interviews by age stratum and sampling method

Location	UVB units	Population size age 20-74	Total number of cases age 20-74	General population samples				
				Age 20-64 RDD	Age 65-74		Case samples	
					RDD	HCFA SRS	Age 20-49 SRS	Age 50-74 SRS
San Francisco/Oakland	150	1,727,340	4994	805	120	62	72	196
Minneapolis/St. Paul	104	1,117,245	2309	925	170		109	332
Detroit	101	2,008,798	3514	625	128	65	103	271
New Mexico	197	669,383	2510	980	141	61	106	315
New Orleans	186	453,458	2212	595	138	62	46	204
Seattle	95	716,988	1612	532	135		86	256
Utah	136	723,156	2654	687	165	84	99	247
Atlanta	153	775,844	3365	615	142	61	127	271
San Diego	175	1,063,730	5169	514	151		38	229
Total		9,255,942	28,339	6278	1290	395	786	2321

cases and controls. The covariates  $\mathbf{X}$  and  $\mathbf{T}$  are treated separately because information about  $\mathbf{X}$  variables is available for all individuals in the study populations, while information about  $\mathbf{T}$  variables is available only for those in the case-control samples. The number of individuals at location  $l$  in age stratum  $s$  with covariate levels  $\mathbf{x}$  and  $\mathbf{t}$  is  $M(l, s, \mathbf{x}, \mathbf{t})$ , and the number of such cases observed during the study period is  $D(l, s, \mathbf{x}, \mathbf{t})$ .

Each individual in a finite population is assumed to become diseased or not ( $Y=1$  or  $0$ ) according to a causal logistic risk model,

$$\Pr(Y | \mathbf{X}, \mathbf{T}) = \frac{\exp\{Y(\alpha + \underline{\beta}'\mathbf{X} + \underline{\gamma}'\mathbf{T})\}}{\{1 + \exp(\alpha + \underline{\beta}'\mathbf{X} + \underline{\gamma}'\mathbf{T})\}} \quad (2.1)$$

If each individual had known values of  $l, s, \mathbf{x}$ , and  $\mathbf{t}$ , the population log likelihood would be

$$L = \sum \{ D(l, s, \mathbf{x}, \mathbf{t})(\alpha + \underline{\beta}'\mathbf{x} + \underline{\gamma}'\mathbf{t}) - M(l, s, \mathbf{x}, \mathbf{t}) \ln [1 + \exp(\alpha + \underline{\beta}'\mathbf{x} + \underline{\gamma}'\mathbf{t})] \},$$

where the summation is over all levels of  $l, s, \mathbf{x}$ , and  $\mathbf{t}$ . If we use a  $+$  to indicate summation on the corresponding index, the resulting  $(1 + r_1 + r_2)$  estimating equations are

$$D(+, +, +, +) - \sum_{\mathbf{x}, \mathbf{t}} M(+, +, \mathbf{x}, \mathbf{t}) p(\alpha, \underline{\beta}, \underline{\gamma}, \mathbf{x}, \mathbf{t}) = 0 \quad (2.2)$$

$$\sum_{\mathbf{x}} \mathbf{x} D(+, +, \mathbf{x}, +) - \sum_{\mathbf{x}, \mathbf{t}} \mathbf{x} M(+, +, \mathbf{x}, \mathbf{t}) p(\alpha, \underline{\beta}, \underline{\gamma}, \mathbf{x}, \mathbf{t}) = 0$$

$$\sum_{\mathbf{t}} \mathbf{t} D(+, +, +, \mathbf{t}) - \sum_{\mathbf{x}, \mathbf{t}} \mathbf{t} M(+, +, \mathbf{x}, \mathbf{t}) p(\alpha, \underline{\beta}, \underline{\gamma}, \mathbf{x}, \mathbf{t}) = 0,$$

where

$$p(\alpha, \underline{\beta}, \underline{\gamma}, \mathbf{x}, \mathbf{t}) = \exp(\alpha + \underline{\beta}'\mathbf{x} + \underline{\gamma}'\mathbf{t}) / \{1 + \exp(\alpha + \underline{\beta}'\mathbf{x} + \underline{\gamma}'\mathbf{t})\}.$$

If  $D(+, +, \mathbf{x}, \mathbf{t})$  and  $M(+, +, \mathbf{x}, \mathbf{t})$  were known, these equations could be solved iteratively for  $\alpha, \hat{\underline{\beta}}$ , and  $\hat{\underline{\gamma}}$ . The total number of cases,  $D(+, +, +, +)$ , is known because the stratum subtotals  $D(l, s, \mathbf{x}, +)$  are known; also the quantities  $M(l, s, \mathbf{x}, +)$  are known. The terms  $D(+, +, +, \mathbf{t})$  and  $M(+, +, \mathbf{x}, \mathbf{t})$  are not known, however, because we have measured  $\mathbf{t}$  only on sampled members of the population. We obtain pseudolikelihood estimates (Gong and Samaniego, 1981) for  $\alpha, \underline{\beta}$ , and  $\underline{\gamma}$  by substituting consistent estimates (see Section 3.2) for  $\hat{D}(+, +, +, \mathbf{t})$  and  $\hat{M}(+, +, \mathbf{x}, \mathbf{t})$  in these estimating equations. For informal assessment of goodness-of-fit, we also compute a pseudo log likelihood by substituting these estimates of  $\alpha, \underline{\beta}, \underline{\gamma}, D(+, +, +, \mathbf{t})$ , and  $M(+, +, \mathbf{x}, \mathbf{t})$  into the log likelihood above.

It should be noted that the overall exposure level of UVB and location are totally confounded. Thus, it would not be feasible to allow for separate intercepts for each location in equation (2.1) and also to estimate the effects of UVB, which correspond to components of  $\underline{\beta}$ . Some of the effects attributed to UVB in this model might represent other location-specific factors not otherwise represented in  $\mathbf{X}$  or  $\mathbf{T}$ . If, however, model (2.1) is correct, the analyses presented will yield proper estimates not only of relative risk parameters  $\underline{\beta}$  and  $\underline{\gamma}$  but also

of the intercept  $\alpha$ . Recall that  $\alpha$  is the ln odds of disease for persons at levels  $\mathbf{X} = \mathbf{T} = 0$ ; thus,  $\alpha$  can be used to estimate baseline risk. Hence, absolute, as well as relative risk, can be estimated in this population-based study.

### 2.2 Consistent Estimates of Terms Needed in the Pseudolikelihood

Classical sampling theory can be used to obtain the consistent estimates necessary for the pseudolikelihood equations. First, an estimate for  $M(l, s, \mathbf{x}, \mathbf{t})$  is developed. The basic general population samples and one supplemental sample for the age group 65-74 are RDD samples, which are probability samples of household telephone numbers. Individuals living in a household with one telephone number have the same probability of selection, but those with two household telephone numbers are twice as likely to be selected. Weighting the latter individuals with a weight of one half allows the expectation of stratum-specific counts to be obtained. If  $m(l, s, \mathbf{x}, \mathbf{t})$  is the weighted number of individuals with confounder values  $\mathbf{x}$  and  $\mathbf{t}$  in the general population sample (GPS) at location  $l$  in stratum  $s$ , then

$$E\{m(l, s, \mathbf{x}, \mathbf{t})\} = m(l, s, +, +)M(l, s, \mathbf{x}, \mathbf{t})/M(l, s, +, +).$$

From sampling theory, the ratio estimator,

$$\hat{M}(l, s, \mathbf{x}, \mathbf{t}) = M(l, s, +, +)m(l, s, \mathbf{x}, \mathbf{t})/m(l, s, +, +),$$

is a consistent estimator for  $M(l, s, \mathbf{x}, \mathbf{t})$ . For age stratum 65-74, the same procedure is used, but  $m(l, 3, \mathbf{x}, \mathbf{t})$  is the weighted sum of counts in the GPS and supplemental sample.

Similarly, the basic case (patient) samples and the supplemental case samples in the age group 20-49 are simple random samples. Since the supplemental sample falls within a single age stratum, the counts of cases in each stratum are obtained without weighting. If  $d(l, s, +, \mathbf{t})$  is the number of cases with covariate value  $\mathbf{t}$  in either case sample at location  $l$  in stratum  $s$ , then  $E\{d(l, s, +, \mathbf{t})\} = d(l, s, +, +)D(l, s, +, \mathbf{t})/D(l, s, +, +)$ , where  $d(l, s, +, +)$  is the number of cases selected at location  $l$ , stratum  $s$ . It then follows that  $\hat{D}(l, s, \mathbf{x}, \mathbf{t}) = D(l, s, +, +)d(l, s, \mathbf{x}, \mathbf{t})/d(l, s, +, +)$  is a consistent estimator for  $D(l, s, \mathbf{x}, \mathbf{t})$ .

An estimate for  $\sum_{\mathbf{t}} \mathbf{t} D(+, +, +, \mathbf{t})$  based on the case samples is

$$\begin{aligned} & \sum_{l,s} \sum_{\mathbf{t}} \mathbf{t} d(l, s, +, \mathbf{t}) D(l, s, +, +) / d(l, s, +, +) \\ & \equiv \sum_{l,s} D(l, s, +, +) \bar{\mathbf{t}}(l, s), \end{aligned}$$

where  $\bar{\mathbf{t}}(l, s) \equiv \sum_{\mathbf{t}} \mathbf{t} d(l, s, +, \mathbf{t}) / d(l, s, +, +)$ .

Other estimates required for the pseudolikelihood equations are obtained analogously from the general population samples. First,  $\sum_{\mathbf{x}, \mathbf{t}} M(+, +, \mathbf{x}, \mathbf{t}) p(\alpha, \underline{\beta}, \underline{\gamma}, \mathbf{x}, \mathbf{t})$  is estimated by

$$\begin{aligned} & \sum_{\mathbf{x}, \mathbf{t}} \sum_{l,s} \frac{m(l, s, \mathbf{x}, \mathbf{t})}{m(l, s, +, +)} M(l, s, +, +) p(\alpha, \underline{\beta}, \underline{\gamma}, \mathbf{x}, \mathbf{t}) \\ & \equiv \sum_{l,s} M(l, s, +, +) \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s), \end{aligned}$$

where

$$\bar{p}(\alpha, \underline{\beta},$$

Similar

where

$$\bar{x}p($$

Finally,

$$\sum_{l,s}$$

where

$$\bar{t}p($$

The est reduce to

$$\sum_{\mathbf{x}} \mathbf{x} D$$

and

The third,...

of  $\mathbf{X}$ , and correspond collected valued fu can be s the Newt order Ta derivative

### 2.3 Cov

The covar mate  $\underline{A}\bar{v}$  var( $\underline{E}$ ). T tioned Ta  $\hat{\underline{\beta}}, \hat{\underline{\gamma}}$ , allo as  $-\underline{A}\underline{E}$ .

where

$$\bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) \equiv \sum_{\underline{x}, \underline{t}} p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) m(l, s, \underline{x}, \underline{t}) / m(l, s, +, +).$$

Similarly,  $\Sigma_{\underline{x}, \underline{t}} \underline{x} M(+, +, \underline{x}, \underline{t}) p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t})$  is estimated by

$$\begin{aligned} \sum_{\underline{x}, \underline{t}} \sum_{l, s} \frac{\underline{x} m(l, s, \underline{x}, \underline{t})}{m(l, s, +, +)} M(l, s, +, +) p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) \\ \equiv \sum_{l, s} M(l, s, +, +) \underline{x} \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s), \end{aligned}$$

where

$$\begin{aligned} \underline{x} \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) \\ \equiv \sum_{\underline{x}, \underline{t}} \underline{x} p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) m(l, s, \underline{x}, \underline{t}) / m(l, s, +, +). \end{aligned}$$

Finally,  $\Sigma_{\underline{x}, \underline{t}} \underline{t} M(+, +, \underline{x}, \underline{t}) p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t})$  is estimated by

$$\begin{aligned} \sum_{l, s} \sum_{\underline{x}, \underline{t}} \underline{t} \frac{m(l, s, \underline{x}, \underline{t})}{m(l, s, +, +)} M(l, s, +, +) p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) \\ \equiv \sum_{l, s} M(l, s, +, +) \underline{t} \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s), \end{aligned}$$

where

$$\begin{aligned} \underline{t} \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) \\ \equiv \sum_{\underline{x}, \underline{t}} \underline{t} p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) m(l, s, \underline{x}, \underline{t}) / m(l, s, +, +). \end{aligned}$$

The estimating equations based on the pseudolikelihood reduce to

$$D(+, +, +, +) - \sum_{l, s} M(l, s, +, +) \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) = 0,$$

$$\sum_{\underline{x}} \underline{x} D(+, +, \underline{x}, +) - \sum_{l, s} M(l, s, +, +) \underline{x} \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) = \mathbf{0},$$

and

$$\begin{aligned} \sum_{l, s} D(l, s, +, +) \underline{t}(l, s) \\ - \sum_{l, s} M(l, s, +, +) \underline{t} \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) = \mathbf{0}. \end{aligned}$$

The first equation corresponds to  $\alpha$ , the second, third, ...,  $(r_1 + 1)$ th equations correspond to the components of  $\underline{X}$ , and the  $r_1 + 2, r_1 + 3, \dots, r_1 + (r_2 + 1)$ th equations correspond to the components of  $\underline{T}$ . These equations can be collected into a  $(r_1 + r_2 + 1) \times 1$  vector  $\underline{E}$ , which is a vector-valued function of  $\alpha, \underline{\beta}$ , and  $\underline{\gamma}$ . The equation  $\underline{E}(\alpha, \underline{\beta}, \underline{\gamma}) = \mathbf{0}$  can be solved iteratively for estimates  $\hat{\alpha}, \hat{\underline{\beta}}$ , and  $\hat{\underline{\gamma}}$  using the Newton-Raphson procedure, which is based on the first-order Taylor series expansion of  $\underline{E}$ . The required vector of derivatives is provided in Appendix 1 and is denoted by  $\underline{E}^*$ .

### 2.3 Covariance of the Estimates

The covariance of  $\hat{\alpha}, \hat{\underline{\beta}}$ , and  $\hat{\underline{\gamma}}$  is based on the sandwich estimate  $\underline{A} \widehat{\text{var}}(\underline{E}) \underline{A}^T$ , where  $\underline{A}^{-1} = \underline{E}^*$  and  $\widehat{\text{var}}(\underline{E})$  estimates  $\text{var}(\underline{E})$ . This sandwich estimator arises from the aforementioned Taylor-series expansion, which, when evaluated at  $\hat{\alpha}, \hat{\underline{\beta}}, \hat{\underline{\gamma}}$ , allows one to approximate  $(\hat{\alpha} - \alpha, (\hat{\underline{\beta}} - \underline{\beta})', (\hat{\underline{\gamma}} - \underline{\gamma})')'$  as  $-\underline{A}\underline{E}$ . The quantities  $D(l, s, \underline{x}, \underline{t})$  are regarded as Poisson

variables with  $\text{var}[D(l, s, \underline{x}, \underline{t})]$  estimated by  $D(l, s, \underline{x}, \underline{t})$  and with  $\text{cov}[D(l, s, \underline{x}, \underline{t}), D(l', s', \underline{x}', \underline{t}')]$  estimated by zero unless  $l = l', s = s', \underline{x} = \underline{x}',$  and  $\underline{t} = \underline{t}'$ . Thus, the estimated variances of  $D(+, +, +, +)$  and  $\Sigma_{l, s} \underline{x} D(l, s, \underline{x}, +)$  are  $D(+, +, +, +)$  and  $\Sigma_{l, s} \underline{x} \underline{x}^T D(l, s, \underline{x}, +)$ , respectively. Their estimated covariance is  $\Sigma_{l, s} \underline{x} D(l, s, \underline{x}, +)$ . The estimates of the variance of  $\Sigma_{l, s} D(l, s, +, +) \underline{t}(l, s)$  and its covariances with  $D(+, +, +, +)$  and  $\Sigma_{l, s} \underline{x} D(l, s, \underline{x}, +)$  are given in Appendix 2. The general population control sample and the case sample are independent; thus, terms involving  $D$  and terms involving  $M$  in the log pseudolikelihood are independent. The  $M$  terms are sums of ratio estimators whose variances and covariances are obtained from classical sampling theory for cluster samples. While the method by which self-weighted samples were obtained does not alter the expectation of sample means, the variances are affected. Details are provided in Appendix 3.

### 3. A Numerical Example

As described in Section 1, the nonmelanoma skin cancer study provided complete population information on incidence rates by 5-year age groups and UVB levels. Thus, the estimating equations for a logistic model based only on these two factors, the first two equations in expression (2.2), can be specified without recourse to the case-control data. We used this approach to develop a risk model involving age and UVB but not other factors such as skin color. Five-year incidence rates for males 20-74 were analyzed using the SAS procedure, PROC LOGISTIC (SAS Institute, 1989). Models were compared to a main effects model with 8 d.f. for the nine locations and with 10 d.f. for the 11 age groups (see Table 2). Taking age to be the midage of an age group, either age or  $\ln(\text{age})$  could be regarded as a continuous covariate to replace the 10 age group variables, but together they were nearly as effective as the full model (a chi-square on 8 d.f. of  $197,593 - 197,581 = 12, p = 0.15$ ). On the other hand, the fit is not good when the eight location variables are replaced with simple models using UVB or  $\ln(\text{UVB})$  (e.g., for the model using  $\ln(\text{UVB})$ , age, and  $\ln(\text{age})$ ,  $\chi^2 = 198,545 - 197,593 = 952, p < 0.001$ ) or with complex models using these variables (e.g., for the models using UVB, UVB<sup>2</sup>, age, and  $\ln(\text{age})$ ,  $\chi^2 = 198,432 - 197,593 = 839, p < 0.001$ ), indicating that host factors or location-specific factors other than UVB partly account for variation in rates. Despite the poor fit, we estimated the effect of  $\ln(\text{UVB})$  to measure UVB exposure because the coefficient of  $\ln(\text{UVB})$  has been used in previous studies and is readily interpretable. From Table 2, the coefficient of  $\ln(\text{UVB})$  is 1.46 with standard error 0.03. According to this model, the relative odds of disease in a person exposed to a 1% higher level of UVB, compared to a baseline exposure level, is odds ratio =  $1.01^{1.46} = 1.0146$ . The excess relative risk (in percent) corresponding to a 1% increase in UVB exposure is  $(1.0146 - 1) \times 100 = 1.46\%$ , which is called the biological amplification factor (BAF) in the literature on risk from UVB (Scotto et al., 1982).

To obtain a better fitting model and to adjust the estimated effect of  $\ln(\text{UVB})$  for host factors that were measured only in the case-control surveys, we used the pseudolikelihood method in Section 2. First, adjusting only for age and  $\ln(\text{age})$  in the pseudolikelihood but not for other covariates, we obtain

**Table 2**  
Comparison of models

Data set	Method	Model	d.f.	ln(UVB) coefficient ± SE	-2 × log likelihood	-2 × log pseudo likelihood
Population with 5-year age groups	Logistic	Age groups + locations	19	na	197,581	
Population with 5-year age groups	Logistic	Age + ln(age) + locations	11	na	197,593	
Population with 5-year age groups	Logistic	Age + ln(age) + ln(UVB)	4	1.46 ± .03	198,545	
Weighted samples	Pseudolikelihood	Age + ln(age) + ln(UVB)		1.39 ± .04		198,874
Weighted samples	Pseudolikelihood	Age + ln(age) + ln(UVB) + ancestry		1.43 ± .04		198,502
Weighted samples	Pseudolikelihood	Age + ln(age) + ln(UVB) + burn/tan + outdoor occupation		1.31 ± .065		195,448

an estimate for the coefficient of ln(UVB) of  $1.39 \pm 0.04$ . The BAF is then estimated as  $\{(1.01)^{1.39} - 1\} \times 100 = 1.39$  with a 95% confidence interval (CI) of 1.31–1.48. Note that the estimate 1.39 differs slightly from the maximum likelihood estimate 1.46 (Table 2) because the former method relies on averages of sampled exact ages within broad age strata, whereas the latter method uses the population age distribution in 5-year groups.

Another important host factor is skin color. Dark-skinned individuals tend to be resistant to UVB-induced skin cancer. This characteristic is especially common among people of Mexican ancestry. An indicator of Mexican ancestry in the presence of the other terms in the previous model is significant ( $p \leq 0.001$  by the Wald test) and suggests an especially low odds of disease (relative odds: 0.08, CI: 0.03–0.23). However, the distribution of individuals with Mexican ancestry across the study locations is such that the coefficient of ln(UVB), 1.43, is hardly changed (BAF: 1.43, CI: 1.34–1.52). The fit of the model, however, is apparently improved by inclusion of Mexican ancestry, as indicated by a decrease in  $-2 \times$  pseudo log likelihood from 198,874 to 198,502 (Table 2). Although exact significance tests for pseudolikelihood ratios are difficult (Liang and Self, 1996), this large decrease is indicative of improved model fit.

Other strong host factors include sunburn or suntan characteristics and patterns of outdoor exposure. Indicators for sunburn–suntan characteristics and outdoor exposure on principal occupation were significant when added to the basic model as judged by the Wald test (Table 2). The referent group was males who are rarely or never outdoors for their principal occupation and who develop a dark tan and do not sunburn. Those who burn but do not develop any tan have a relative odds of 3.35 (CI: 2.51–4.48); those who burn and develop a light tan have a relative odds of 2.73 (CI: 2.19–3.40); and those who burn and develop an average tan have a relative odds of 1.83 (CI: 1.53–2.17). Those who are outdoors frequently or occasionally on their principal occupation have a relative odds of 1.19 (CI: 1.01–1.39). Inclusion of all these factors in the model resulted in a lower, but somewhat less precise, estimate of the effect of ln(UVB), 1.31 (BAF: 1.31%, CI: 1.17–1.43). Notably,  $-2 \times$  pseudo log likelihood is reduced

substantially to 195,448, which is even smaller than  $-2 \times$  log likelihood in the model with saturated location effects (row 1, Table 2).

Suppose conventional logistic methods are applied to analyze the case-control data. Because the ratio of cases to controls in each location does not reflect the ratio of total cases to population size in each location, a standard logistic analysis that includes location or a location-level factor, such as UVB, will yield meaningless results. This distortion is analogous to naive analysis of the second-stage in a two-stage case-control design (Breslow and Cain, 1988).

Another, more appropriate logistic analysis of the case-control data would reweight cases and controls in each location and age stratum to represent all the cases and the total population in each location and age stratum. Because this reweighting inflates the apparent study population, a naive logistic analysis of the reweighted data will have an unrealistically small estimate of variance and covariance of exposure effects. Within each of eleven 5-year age groups at each location, the base weights for cases and population size are separately adjusted (or poststratified) so that the sum of poststratified weights within a location and age group equals the known total number of cases and controls. The variability of these poststratified weights is ignored. Adjusting for age, ln(age), sunburn–suntan characteristics, and exposure habits, this naive logistic analysis yields an estimate for the coefficient of ln(UVB) of 1.33 with standard error (SE) of 0.033. The correct standard error estimate, based on the methods of this paper, is 0.065. The standard errors of other coefficients are also much too small. For example, the SE for the estimated coefficient for outdoor exposure is 0.016, compared to the more appropriate estimate 0.081. Thus, the use of standard logistic methods with poststratified weights results in a serious underestimation of standard errors.

The use of cluster sampling with subsampling (RDD) allowed identification of general population controls with enormous cost savings compared with simple random sampling of individuals, but how seriously were standard errors affected? To assess this issue, we considered a setting in which all households have only one phone; thus, RDD results in a



Graubard, B. and Gail, M. (1992). Cluster sampling of controls in population-based case-control studies. In *Accomplishments in Cancer Research 1991*, J. G. Fortner and J. E. Rhoads (eds), 366-367. Philadelphia: Lippincott.

Graubard, B. I., Fears, T. R., and Gail, M. H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case-control studies. *Biometrics* **45**, 1053-1071.

Liang, K. Y. and Self, S. G. (1996). On the asymptotic behavior of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society, Series B* **58**, 785-796.

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. New York: Wiley.

SAS Institute. (1989). *SAS/STAT User's Guide*, Version 6, 4th edition, Volume 2. Cary, North Carolina: SAS Institute.

Scotto, J., Fears, T. R., and Fraumeni, J. F., Jr. (1981). *Incidence of nonmelanoma skin cancer in the United States*, Publication (NIH) 82-2433, Department of Health and Human Services, U.S. Government Printing Office, Washington, D.C.

Scotto, J., Fears, T. R., and Fraumeni, J. F., Jr. (1982). Solar radiation. In *Cancer Epidemiology and Prevention*, D. Schottenfeld and J. F. Fraumeni, Jr. (eds), 254-276. Philadelphia: W. B. Saunders.

Scotto, J., Cotton, G., Urbach, F., Berger, D., and Fears, T. R. (1988). Biologically effective ultraviolet radiation: Surface measurements in the United States, 1974 to 1985. *Science* **239**, 762-763.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association* **73**, 40-46.

White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* **115**, 119-128.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Received October 1997. Revised June 1999.  
Accepted July 1999.

APPENDIX 1

The Derivatives of the Pseudo Log Likelihood Equations

Assume  $r_1 = r_2 = 1$  so that  $E' \equiv (E_1, E_2, E_3)$  has only three components. Then

$$\begin{aligned} \frac{dE_1}{d\alpha} &= - \sum_{l,s} M(l, s, +, +) \\ &\quad \times \sum_{x,t} \frac{m(l, s, x, t)}{m(l, s, +, +)} p(\alpha, \beta, \gamma, x, t) \{1 - p(x, t, \alpha, \beta, \gamma)\} \\ \frac{dE_1}{d\beta} &= - \sum_{l,s} M(l, s, +, +) \\ &\quad \times \sum_{x,t} x \frac{m(l, s, x, t)}{m(l, s, +, +)} p(\alpha, \beta, \gamma, x, t) \{1 - p(x, t, \alpha, \beta, \gamma)\} \\ \frac{dE_1}{d\gamma} &= - \sum_{l,s} M(l, s, +, +) \end{aligned}$$

$$\begin{aligned} &\quad \times \sum_{x,t} t \frac{m(l, s, x, t)}{m(l, s, +, +)} p(\alpha, \beta, \gamma, x, t) \{1 - p(x, t, \alpha, \beta, \gamma)\} \\ \frac{dE_2}{d\beta} &= - \sum_{l,s} M(l, s, +, +) \\ &\quad \times \sum_{x,t} tx^2 \frac{m(l, s, x, t)}{m(l, s, +, +)} p(\alpha, \beta, \gamma, x, t) \\ &\quad \times \{1 - p(x, t, \alpha, \beta, \gamma)\} \\ \frac{dE_2}{d\gamma} &= - \sum_{l,s} M(l, s, +, +) \\ &\quad \times \sum_{x,t} xt \frac{m(l, s, x, t)}{m(l, s, +, +)} p(\alpha, \beta, \gamma, x, t) \\ &\quad \times \{1 - p(x, t, \alpha, \beta, \gamma)\} \\ \frac{dE_3}{d\gamma} &= - \sum_{l,s} M(l, s, +, +) \\ &\quad \times \sum_{x,t} t^2 \frac{m(l, s, x, t)}{m(l, s, +, +)} p(\alpha, \beta, \gamma, x, t) \\ &\quad \times \{1 - p(x, t, \alpha, \beta, \gamma)\}. \end{aligned}$$

Note that

$$\frac{dE_2}{d\alpha} = \frac{dE_1}{d\beta}, \quad \frac{dE_3}{d\alpha} = \frac{dE_1}{d\gamma}, \quad \text{and} \quad \frac{dE_3}{d\beta} = \frac{dE_2}{d\gamma}.$$

Extensions to  $r_1$  or  $r_2 > 1$  are straightforward.

APPENDIX 2

Variations and Covariances of Estimates Based on the Case Sample

In this section, we provide an estimate for the variance of the case sample-based estimate,  $\Sigma_{l,s} D(l, s, +, +) \bar{l}(l, s)$ , as well as estimates of its covariances with the incidence survey counts,  $D(+, +, +, +)$  and  $\Sigma_{\underline{x}} \underline{x}D(+, +, \underline{x}, +)$ .

The variance estimate was obtained using a conditioning argument. First, conditional on the  $D(l, s, +, \underline{t})$ 's and regarding the sample sizes  $d(l, s, +, +)$  as fixed constants, we compute the conditional expectation of  $\Sigma_{l,s} D(l, s, +, +) \bar{l}(l, s)$  as

$$\sum_{l,s} \sum_{\underline{t}} \underline{t} D(l, s, +, \underline{t}). \tag{A2.1}$$

Again, conditional on the  $D(l, s, +, \underline{t})$ 's, the conditional variance of  $\Sigma_{l,s} D(l, s, +, +) \bar{l}(l, s)$  is

$$\begin{aligned} &\sum_{l,s} D^2(l, s, +, +) \left\{ 1 - \frac{d(l, s, +, +)}{D(l, s, +, +)} \right\} \\ &\quad \times \left[ \frac{\sum_{\underline{t}} \underline{t} \underline{t}^T D(l, s, +, \underline{t})}{[D(l, s, +, +) - 1]d(l, s, +, +)} \right. \\ &\quad \left. - \frac{\left\{ \sum_{\underline{t}} \underline{t} D(l, s, +, \underline{t}) \right\} \left\{ \sum_{\underline{t}} \underline{t} D(l, s, +, \underline{t}) \right\}^T}{D(l, s, +, +)[D(l, s, +, +) - 1]d(l, s, +, +)} \right], \end{aligned} \tag{A2.2}$$

Material may be protected by copyright law (Title 17, U.S. Code)

which includes a finite sampling correction factor for SRS. The unconditional variance of  $\Sigma_{l,s} D(l, s, +, +) \bar{\mathbf{t}}(l, s)$  is then the variance of (A2.1) plus the expectation of (A2.2). Now the variance of (A2.1) is  $\Sigma_{l,s} \Sigma_{\mathbf{t}} \mathbf{t} \mathbf{t}^T \text{var}\{D(l, s, +, \mathbf{t})\}$ , which is estimated by

$$\sum_{l,s} D(l, s, +, +) \sum_{\mathbf{t}} \mathbf{t} \mathbf{t}^T d(l, s, +, \mathbf{t}) / d(l, s, +, +). \quad (\text{A2.3})$$

The estimate

$$\begin{aligned} & \sum_{l,s} D^2(l, s, +, +) \left\{ 1 - \frac{d(l, s, +, +)}{D(l, s, +, +)} \right\} \\ & \times \left[ \frac{\sum_{\mathbf{t}} \mathbf{t} \mathbf{t}^T d(l, s, +, \mathbf{t})}{\{d(l, s, +, +) - 1\} d(l, s, +, +)} \right. \\ & \left. - \frac{\left\{ \sum_{\mathbf{t}} \mathbf{t} d(l, s, +, \mathbf{t}) \right\} \left\{ \sum_{\mathbf{t}} \mathbf{t} d(l, s, +, \mathbf{t}) \right\}^T}{d(l, s, +, +) \{d(l, s, +, +) - 1\} d(l, s, +, +)} \right] \end{aligned} \quad (\text{A2.4})$$

has the same expectation as (A2.2) and can therefore be used to estimate the expectation of (A2.2). The estimate of the variance of  $\Sigma_{l,s} D(l, s, +, +) \bar{\mathbf{t}}(l, s)$  is then obtained by summing (A2.3) and (A2.4).

Consider the estimation of the covariance of  $\Sigma_{l,s} D(l, s, +, +) \bar{\mathbf{t}}(l, s)$  and  $D(+, +, +, +) = \Sigma_{l,s} D(l, s, +, +)$ . The covariance of  $D(l, s, +, +)$  and  $D(l', s', +, +) \bar{\mathbf{t}}(l', s')$  is zero for  $l \neq l'$  or  $s \neq s'$  because  $D(l, s, +, +)$  and  $D(l', s', +, +)$  are independent. Thus, we can restrict attention to the covariance of  $D(l, s, +, +) \bar{\mathbf{t}}(l, s)$  with  $D(l, s, +, +)$ . The expectation of their product conditional on the  $D(l, s, +, \mathbf{t})$ 's is  $D(l, s, +, +) \times \Sigma_{\mathbf{t}} \mathbf{t} D(l, s, +, \mathbf{t})$ , which can be written

$$\begin{aligned} & \sum_{\mathbf{t}'} D(l, s, +, \mathbf{t}') \sum_{\mathbf{t}} \mathbf{t} D(l, s, +, \mathbf{t}) \\ & = \sum_{\mathbf{t}} \mathbf{t} D^2(l, s, +, \mathbf{t}) + \sum_{\mathbf{t}} \mathbf{t} D(l, s, +, \mathbf{t}) \sum_{\mathbf{t}' \neq \mathbf{t}} D(l, s, +, \mathbf{t}'). \end{aligned}$$

The unconditional expectation of the product of  $D(l, s, +, +) \times \bar{\mathbf{t}}(l, s)$  and  $D(l, s, +, +)$  is then

$$\begin{aligned} & \sum_{\mathbf{t}} \mathbf{t} [E\{D(l, s, +, \mathbf{t})\} + E^2\{D(l, s, +, \mathbf{t})\}] \\ & + \sum_{\mathbf{t}} \mathbf{t} E\{D(l, s, +, \mathbf{t})\} \\ & \quad \times [E\{D(l, s, +, +)\} - E\{D(l, s, +, \mathbf{t})\}] \\ & = \sum_{\mathbf{t}} \mathbf{t} E\{D(l, s, +, \mathbf{t})\} \\ & \quad + E \left\{ \sum_{\mathbf{t}} \mathbf{t} D(l, s, +, \mathbf{t}) \right\} E\{D(l, s, +, +)\}. \end{aligned}$$

It follows that the covariance of  $D(l, s, +, +) \bar{\mathbf{t}}(l, s)$  and  $D(l, s, +, +)$  is the first term,  $\Sigma_{\mathbf{t}} \mathbf{t} D(l, s, +, \mathbf{t})$ . Thus, the covariance of  $\Sigma_{l,s} D(l, s, +, +) \bar{\mathbf{t}}(l, s)$  and  $D(+, +, +, +)$  is  $\Sigma_{l,s} \Sigma_{\mathbf{t}} \mathbf{t} D(l, s,$

$+, \mathbf{t})$ , which is estimated by

$$\sum_{l,s} D(l, s, +, +) \bar{\mathbf{t}}(l, s).$$

Similarly, the covariance of  $\Sigma_{\mathbf{x}} \mathbf{x} D(+, +, \mathbf{x}, +)$  and  $\Sigma_{l,s} \mathbf{t} \times D(l, s, +, \mathbf{t})$  is estimated by  $\Sigma_{l,s} D(l, s, +, +) \underline{\mathbf{x}} \bar{\mathbf{t}}(l, s)$ , where  $\underline{\mathbf{x}} \bar{\mathbf{t}}(l, s) = \Sigma_{\mathbf{x}, \mathbf{t}} \mathbf{x} \mathbf{t} d(l, s, \mathbf{x}, \mathbf{t}) / d(l, s, +, +)$ .

### APPENDIX 3

#### Variance and Covariance Estimates That Utilize the General Population Samples

It is sufficient to obtain the variances and covariances for  $\bar{p}$ ,  $\underline{\mathbf{x}} \bar{\mathbf{p}}$ , and  $\bar{\mathbf{t}} \bar{\mathbf{p}}$ . We first consider only covariances that involve  $\bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s)$ , the estimator for

$$\begin{aligned} & \bar{P}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) \\ & = \Sigma_{\mathbf{x}, \mathbf{t}} M(l, s, \mathbf{x}, \mathbf{t}) p(\alpha, \underline{\beta}, \underline{\gamma}, \mathbf{x}, \mathbf{t}) / M(l, s, +, +), \end{aligned}$$

the mean value of  $p$  at location  $l$  in stratum  $s$ . For  $s = 1, 2$ , these estimates are based only on a random digit dialing (RDD) sample in which the first-stage clusters are selected with replacement with probability proportional to the number of households in each cluster. The probability of selecting the  $c$ th cluster is  $\xi_c = H(l, c) / H(l, +)$ , where  $H(l, c)$  is the number of households in the  $c$ th cluster at location  $l$  and  $H(l, +)$  is the sum of  $H(l, c)$  over all  $K_l$  clusters in location  $l$ . In the second stage,  $r$  households are randomly selected from each cluster without replacement.

The variance of  $\bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s)$  can be obtained following classical sampling results (e.g., Cochran, Section 11.14). In our notation,

$$\begin{aligned} & \text{var} [\{\bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s)\}] \\ & = \frac{1}{kM^2(l, s, +, +)} \\ & \quad \times \sum_{c=1}^{K_l} \frac{1}{\xi_c} \left[ \sum_{\mathbf{x}, \mathbf{t}} M(l, s, c, \mathbf{x}, \mathbf{t}) \right. \\ & \quad \left. \times \{p(\alpha, \underline{\beta}, \underline{\gamma}, \mathbf{x}, \mathbf{t}) - \bar{P}(\alpha, \underline{\beta}, \underline{\gamma}, l, s)\} \right]^2 \\ & \quad + \frac{1}{kM^2(l, s, +, +)} \sum_{c=1}^{K_l} \frac{H^2(l, c)}{\xi_c} \left\{ 1 - \frac{r}{H(l, c)} \right\} \frac{S_{p2c}^2}{r}, \end{aligned} \quad (\text{A3.1})$$

where  $M(l, s, c, \mathbf{x}, \mathbf{t})$  denotes the number in the population with exposures  $\mathbf{x}$  and  $\mathbf{t}$  on the  $c$ th cluster and, with  $\delta(l, s, c, h, \mathbf{x}, \mathbf{t})$  as the number in the  $h$ th household with exposures  $\mathbf{x}$  and  $\mathbf{t}$ ,  $S_{p2c}^2$  is the variance of

$$\sum_{\mathbf{x}, \mathbf{t}} \delta(l, s, c, h, \mathbf{x}, \mathbf{t}) \{p(\alpha, \underline{\beta}, \underline{\gamma}, \mathbf{x}, \mathbf{t}) - \bar{P}(\alpha, \underline{\beta}, \underline{\gamma}, l, s)\}.$$

The first component of  $\text{var}[\{\bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s)\}]$  in (A3.1) arises from variation between clusters and the second component arises from variation within clusters. The quantities  $\text{var}[\{\underline{\mathbf{x}} \bar{\mathbf{p}}(\alpha, \underline{\beta}, \underline{\gamma}, l, s)\}]$  and  $\text{var}[\{\bar{\mathbf{t}} \bar{\mathbf{p}}(\alpha, \underline{\beta}, \underline{\gamma}, l, s)\}]$  are obtained in exactly the same way.

Using the classical methods, we also obtain covariances. For  $s = 1, 2$  and  $s' = 1, 2$ ,

$$\begin{aligned} \text{cov} \left[ \left\{ \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s), \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s') \right\} \right] \\ = \frac{1}{kM(l, s, +, +)M(l, s', +, +)} \\ \times \sum_{c=1}^{K_l} \frac{1}{\xi_c} \left[ \sum_{\underline{x}, \underline{t}} M(l, s, c, \underline{x}, \underline{t}) \right. \\ \left. \times \left\{ p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) - \bar{P}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) \right\} \right] \\ \times \left[ \sum_{\underline{x}, \underline{t}} M(l, s', c, \underline{x}, \underline{t}) \right. \\ \left. \times \left\{ p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) - \bar{P}(\alpha, \underline{\beta}, \underline{\gamma}, l, s') \right\} \right] \\ + \frac{1}{kM(l, s, +, +)M(l, s', +, +)} \\ \times \sum_{c=1}^{K_l} \frac{H^2(l, c) [1 - r/H(l, c)]}{r\xi_c} S_{p, p2c}(s, s'), \end{aligned}$$

where  $S_{p, p2c}$  is the covariance of

$$\sum_{\underline{x}, \underline{t}} \delta(l, s, c, h, \underline{x}, \underline{t}) \left\{ p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) - \bar{P}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) \right\}$$

and

$$\sum_{\underline{x}, \underline{t}} \delta(l, s', c, h, \underline{x}, \underline{t}) \left\{ p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) - \bar{P}(\alpha, \underline{\beta}, \underline{\gamma}, l, s') \right\}.$$

Also for  $s = 1, 2$  and  $s' = 1, 2$ ,

$$\begin{aligned} \text{cov} \left[ \left\{ \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s), \bar{x}\bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s') \right\} \right] \\ = \frac{1}{kM(l, s, +, +)M(l, s', +, +)} \\ \times \sum_{c=1}^{K_l} \frac{1}{\xi_c} \left[ \sum_{\underline{x}, \underline{t}} M(l, s, c, \underline{x}, \underline{t}) \right. \\ \left. \times \left\{ p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) - \bar{P}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) \right\} \right] \\ \times \left[ \sum_{\underline{x}, \underline{t}} M(l, s', c, \underline{x}, \underline{t}) \right. \\ \left. \times \left\{ \bar{x}p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) - \bar{x}\bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s') \right\} \right] \end{aligned}$$

$$\begin{aligned} + \frac{1}{kM(l, s, +, +)M(l, s', +, +)} \\ \times \sum_{c=1}^{K_l} \frac{H^2(l, c) [1 - r/H(l, c)]}{r\xi_c} S_{p, xp2c}(s, s'), \end{aligned}$$

where  $S_{p, xp2c}(s, s')$  is the covariance of

$$\sum_{\underline{x}, \underline{t}} \delta(l, s, c, h, \underline{x}, \underline{t}) \left\{ p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) - \bar{P}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) \right\}$$

and

$$\sum_{\underline{x}, \underline{t}} \delta(l, s', c, h, \underline{x}, \underline{t}) \left\{ \bar{x}p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) - \bar{x}\bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s') \right\}.$$

Other needed covariances are obtained using obvious substitutions.

Consistent estimates of these quantities are obtained from Cochran (Section 11.4, equation [11.39]) after  $M(l, s, +, +)/H(l, +)$  is replaced by  $m(l, s, +, +)/rk$ .

For example, an asymptotically unbiased estimate of the covariance (A3.2) would be

$$\begin{aligned} \frac{k}{(k-1)m(l, s, +, +)m(l, s', +, +)} \\ \times \sum_{c=1}^k \left[ \sum_{\underline{x}, \underline{t}} m(l, s, c, \underline{x}, \underline{t}) \right. \\ \left. \times \left\{ p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) - \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s) \right\} \right] \\ \times \left[ \sum_{\underline{x}, \underline{t}} m(l, s', c, \underline{x}, \underline{t}) \right. \\ \left. \times \left\{ p(\alpha, \underline{\beta}, \underline{\gamma}, \underline{x}, \underline{t}) - \bar{p}(\alpha, \underline{\beta}, \underline{\gamma}, l, s') \right\} \right]. \end{aligned}$$

The sampling for  $s = 3$  is more complex. At each location, there is the base RDD sample (clusters  $c = 1, \dots, k$ ) with an enlarged sample of  $r^*$  households and an SRS supplement. We regard the pooled estimate for  $s = 3$  as a weighted average of the ratio estimate from the RDD samples and the classical ratio estimate from the SRS. Weights are the expected sample sizes divided by the total expected sample size. Covariance and estimates can then be obtained easily. The expected sample size of the RDD sample is  $M(l, 3, +, +)r^*k/H(l, +)$ , which is unknown, and we replace  $M(l, 3, +, +)H(l, +)$  with  $m(l, 3, +, +)/r^*k$ . Weights become the observed weighted sample sizes divided by the total sample size. Note that in all covariance estimates we also replace  $\bar{P}$  with the pooled estimate rather than estimates based only on the RDD sample or only the SRS.

1. Intro  
Interval  
panel st  
virus (AIDS),  
the even  
will refe

To es  
data, Pe  
a Newto  
ster, Laf  
maximu  
Wellner  
tablished  
the NPN  
tional ha  
a Newto  
the regr  
lished the  
sion coef  
baseline  
Datta, a  
posed th  
sored su  
review o  
for a tut

In this  
ple impu  
censored  
times fr