REGRESSION METHODS FOR DATA WITH INCOMPLETE COVARIATES

ALICE S. WHITTEMORE AND STELLA GROSSER

TECHNICAL REPORT NO. 81
SEPTEMBER 1985
REGRESSION METHODS FOR DATA WITH INCOMPLETE COVARIATES

by

ALICE S. WHITTEMORE

and

STELLA GROSSER

STANFORD UNIVERSITY

TECHNICAL REPORT NO. 81
SEPTEMBER 1985

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH (SIMS)

PREPARED UNDER A GRANT TO SIMS FROM
ENVIRONMENTAL PROTECTION AGENCY (EPA)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Regression Methods for Data with Incomplete Covariates
Alice S. Whittemore* and Stella Grosser*

Abstract. Modern statistical methods in chronic disease epidemiology allow simultaneous regression of disease status on several covariates. These methods permit examination of the effects of one covariate while controlling for those of others that may be causally related to the disease. However they do not accommodate data in which one or more covariates are incomplete, e.g. missing or measured with error. This paper uses assumptions about the probability laws governing covariate incompleteness to obtain estimates of regression coefficients relating disease to the unobserved complete covariates. The estimates are obtained by maximizing the likelihood of the observed, incomplete data via the EM algorithm [1].

1. Introduction. The multivariate procedures available for the analysis of epidemiological studies of chronic disease provide useful techniques for the control of confounding and evaluation of combined effects and interactions. However, application of these techniques is problematical when study subjects lack accurate or complete data on one or more potential risk factors to be tested in multiple regressions. The following examples illustrate several types of covariate incompleteness characteristic of epidemiological data.

1.1 Cervical dysplasia and DES. To examine any relationship between in-utero exposure to diethylstilbesterol (DES) and subsequent cervical dysplasia, a cross-sectional study ascertains indicators for the presence or absence of these two variables for a sample of women seen at a cervical cancer screening clinic. The women (or their mothers) may report prior DES exposure information incorrectly. Thus each woman has a probability of being misclassified into the wrong exposure category.

*Department of Family, Community and Preventive Medicine, Stanford University School of Medicine, Stanford, CA 94305. Supported by NIH Grants CA-23214 and CA-34817 and by grants to the Society for Industrial and Applied Mathematics Institute for Mathematics and Society, from the Sloan Foundation, the Environmental Protection Agency, and the National Science Foundation.
1.2 Colorectal cancer and diet. To examine any relationship between dietary intake of certain nutrients and cancers of the large intestine, a case control study ascertains estimates of frequency and amount of all foods eaten in the previous year, for a sample of cancer patients (cases) and nonpatients (controls). These data are reported with error.

1.3 Death due to breast cancer and extent of disease. To examine any relationship between weight-for-height and survival from breast cancer after adjustment for extent of disease at diagnosis, a prospective study ascertains age, height, weight and stage of disease from a sample of women with newly diagnosed breast cancer who are monitored for subsequent mortality. The medical records of a woman usually provide enough information to place her in exactly one of five stages. Occasionally a woman’s stage is ambiguous due to her incomplete records, so that she is known only to be in one of two stages. Alternatively, she may be lacking the information to stage her at all, or she may be misclassified. Recent findings have shown poor survival among heavy breast cancer patients, relative to survival among those who are leaner [2]. These findings may not reflect an adverse effect of obesity per se, but rather a greater tendency for oncologists to understage obese women, for whom nodal involvement and tumor size may be more difficult to determine.

1.4 Adult mortality and childhood asthma. To examine any relationship between history of childhood asthma and subsequent mortality among a cohort of former college students, Whittemore et al. [3] reviewed characteristics of youth as reported on physical examination at college entrance, as well as alumni characteristics obtained from questionnaires mailed in 1962 or 1966. Virtually complete mortality data were available from alumni records. Some individuals, however, lacked data on childhood asthma because this variable was not included in the physical exam for their college year. Others lacked data on alumni cigarette smoking history, because they died prior to questionnaire mailing.

In all of these examples, an individual’s covariate vector may be incomplete. The incompleteness is due to partially observed, missing, or erroneously measured components. A common approach to such missing covariate data is to delete the individual from the analysis. This approach involves loss of information and possible bias. For example, deleting those alumni who died prior to 1960 restricts the study to those who survived, who may be less susceptible to the effects of smoking than those who died earlier. The alternate approach of omitting from analysis any incomplete covariate components can also lead to bias. Ignoring cigarette smoking can induce a spurious protective effect of childhood asthma, because asthmatics tend to smoke less.

The problem of measurement error applies especially to data from studies of adverse effects of exposures to air or water contaminants, and to data from case-control studies of chronic disease vs. diet, tobacco consumption, or energy expenditures in physical activity. This errors-in-variables problem has been studied extensively when covariates and disease response can be assumed to have a joint multivariate normal distribution (e.g., [4], Chapter 29; [5]). Prentice [6-8] has studied covariate error effects on the Cox regression model. There is need for further work to incorporate such errors into
multivariate regressions when the response and/or covariates are nonnormal. The results of Gail et al., in this volume and elsewhere [9], indicate biased estimates of treatment effect when omitting a balanced covariate. These results suggest more careful attention to incomplete covariates in both the design and analysis phase of epidemiological studies.

This paper develops and illustrates a general method for obtaining multivariate risk estimates using all of the available data. Computational aspects of the method are handled via the EM algorithm [1]. Section 2 defines assumptions about incomplete covariates and describes the general theory for dealing with them. It is shown that in many situations of practical utility, the EM algorithm provides a heuristically appealing and computationally simple method for obtaining maximum likelihood estimates of parameters relating disease to true covariates, based on the observed, incomplete data. At each iteration of the algorithm, an individual’s incomplete covariates are replaced by their expected values, given his disease status, his observed covariates, and the current parameter estimates. Section 3 uses the first three examples outlined above to illustrate the general theory. Section 4 contains a discussion of issues in need of further work.

2. General Theory. Let the random variable $Y$ represent response (e.g. time-to-disease or disease status), and let $X$ be an $m$-dimensional vector of covariates that is observed incompletely. That is, components of $X$ may be missing, truncated, censored, or measured with error, and one observes only a vector $Z$ of incomplete covariates. Although some components of $X$ may be fixed by design, we shall regard them all as random variables with (possibly degenerate) distributions.

We make the following assumptions: i) the observed covariate $Z$ is incomplete because of some stochastic deletion mechanism $D$. That is, $D$ is a random variable and $Z = \Phi(X, D)$ is a known function of $X$ and $D$. Examples include a mechanism $D$ that randomly deletes one or more components of $X$ (e.g. [10]), or an error vector $D$ with $Z = X + D$ (e.g. [5]); ii) conditional on $X$, $Y$ and $Z$ are stochastically independent random variables whose distributions depend on functionally independent parameters $\beta$ and $\tau$, respectively. In symbols,

$$f_{YZ|X}(y, z|x) = f_{Y|X}(y|x, \beta)f_{Z|X}(z|x, \tau);$$

iii) the parameter $\eta$ governing the marginal distribution of $X$ is functionally independent of both $\beta$ and $\tau$.

Assumption (ii) would be violated in the first three examples of Section 1 if women with dysplasia report prenatal DES exposure more accurately than do those without it, if colorectal cancer cases recall their diet more accurately than controls, or if breast cancer patients with poorer prognosis are staged less accurately than those with more favorable prognosis. Assumption (ii) is violated by the asthma mortality data. Conditional on an alumnus’ asthma and smoking status $X$, the probability that his observed covariate $Z$ is complete depends on his time of death $Y$. The theory described below is therefore inappropriate for this data set, and we do not discuss it further.
We wish to make inferences about the parameter $\beta$ governing the relationship between response and the incompletely observed covariates. Assumptions (i)-(iii) imply that the joint ("complete data") distribution for $X, Y$ and $Z$ is

$$f_{XYZ}(x,y,z|\eta,\beta,\tau) = f_X(x,\eta)f_Z(z|x,\tau)f_Y(y|x,\beta).$$

According to (2.1), $X$ and $Z$ are jointly ancillary for $\beta$ in the complete data distribution. That is, $(X,Z)$ are jointly sufficient for $(\eta,\tau)$, and their joint distribution does not depend on $\beta$. Such ancillarity suggests that if the investigator could observe the value assumed by $X$, he should base inferences for $\beta$ on the distribution $f_Y|X$ of response conditional on that value [11].

The marginal distribution of the observed data $(Y,Z)$ is

$$f_{YZ}(y,z|\beta,\eta,\tau) = \int f_X(x|\eta)f_Z(z|x,\tau)f_Y(y|x,\beta)dx,$$

obtained by integrating (2.1) with respect to the $m$ components of $X$ over the sample space of $X$. (Throughout the paper, summation replaces integration for discrete components of $X$.) The conditional distribution of response given the observed covariate $Z$ is obtained by dividing (2.2) by the marginal distribution for $Z$:

$$f_{Y|Z}(y|z,\beta,\eta,\tau) = \int f_{X|Z}(x|z,\eta,\tau)f_Y(y|x,\beta)dx.$$

Note from (2.3) that $Z$ is ancillary for $\beta$ in the observed data distribution (2.2) only if $f_{X|Z}$ is completely specified, i.e. with no unknown parameters. In the absence of such complete specification, inferences based on (2.3) may involve some loss of information for $\beta$. (See [12-15] for further discussion of information and ancillarity in the presence of nuisance parameters.)

We assume that $f_{X|Z}$ is specified by the investigator. Then given independent, identically distributed observations $(y_i,z_i), i = 1,\ldots,n$, we can infer $\beta$ from the conditional loglikelihood

$$l(\beta) = \sum \log f_{Y|Z}(y_i|z_i,\beta)$$
$$= \sum \log \int f_{X|Z}(x|z_i)f_Y(y|x,\beta)dx.$$

The second equality in (2.4) follows from (2.3) and the assumption that $f_{X|Z}$ involves no unknown parameters. Unless otherwise specified, all summations are taken over $i = 1,\ldots,n$.

We will need to evaluate the loglikelihood of the complete data $(x_i,y_i)$, conditional on the $z_i$, $i = 1,\ldots,n$. From (2.1) the probability density for a complete data point $(x,y)$, given the observed covariate $z$, is
\[ f_{X|Y|Z}(x, y|z) = f_{X|Z}(x|z)f_{Y|X}(y|x, \beta). \]

Since \( f_{X|Z} \) is assumed known, the loglikelihood kernel for the complete data is

\[ l_0(\beta) = \sum \log f_{Y|X}(y_i|x_i, \beta). \]  

(2.5)

Of course (2.5) arises from the observed data loglikelihood (2.4) when each \( f_{X|Z}(x|z_i) \) is degenerate, with all of its mass at \( x_i = z_i \).

In this conditional setting, using the terminology of Dempster, Laird and Rubin [1], the complete data are the pairs \((y, x)\). The observed data are the responses \( y \), i.e. the \( x \)'s are "missing."

Specifying \( f_{X|Z} \) implies some knowledge of the marginal covariate distribution \( f_X \), and of the deletion mechanism giving rise to \( f_{Z|X} \). In some applications it may be more reasonable to specify only the distribution \( f_{Z|X} \) of incomplete covariates, conditional on their actual values. Then efficient inferences for \( \beta \) would be based on the joint distribution \( f_{YZ} \) in (2.2), which involves the nuisance parameter \( \eta \) governing \( f_X \). In this unconditional setting, the observed data are the pairs \((y, z)\), and the complete data are the triples \((x, y, z)\). From (2.2) the observed data loglikelihood kernel would be

\[ l^*(\beta, \eta) = \sum \log \int f_{XZ}(x, z|\eta)f_{Y|X}(y|x, \beta)dx, \]  

(2.6)

and from (2.1) the complete data loglikelihood kernel would be

\[ l_0^*(\beta, \eta) = \sum \log f_X(x|\eta) + \sum \log f_{Y|X}(y_i|x_i, \beta)dx. \]  

(2.7)

Maximizing (2.7) with respect to \( \beta \) and \( \eta \) involves maximizing separately each of its two summands. The latter are often assumed to be functions with maxima available in closed form or via standard packaged programs. By contrast (2.4) and (2.6) are loglikelihood kernels for a collection of independent mixture densities, with "mixing parameters" \( f_{X|Z} \) and \( f_{XZ} \) dependent on the individual observed covariates \( z_i \). The decreased tractability of the observed data kernels (2.4) and (2.6) relative to the complete data kernels (2.5) and (2.7) suggests using the EM algorithm to maximize the former.

The EM algorithm is an iterative procedure for maximizing the observed data loglikelihood kernel with respect to \( \beta \). Each iteration consists of an expectation (E) step and a maximization (M) step. The E step uses a current estimate \( \beta^c \) to compute the expected value of the complete data kernel with respect to the \( x \)'s, given the data points \((y, z)\), with \( \beta = \beta^c \). Thus the E step computes

\[ Q(\beta|\beta^c) = E[l_0(\beta)|y, z, \beta^c], \]  

(2.8)
where \( y, z \) represents the observations \((y_i, z_i), \ i = 1, ..., n\). The function \( Q \) is assumed to exist for all possible values of \((\beta, \beta^C, y, z)\). The \( M \) step finds \( \beta^{C+1} \) to maximize \( Q(\beta|\beta^C) \) over \( \beta \). The two steps are iterated, with \( \beta^{C+1} \) replacing \( \beta^C \) until convergence to an estimate \( \hat{\beta} \). Under mild regularity conditions \( \hat{\beta} \) maximizes the loglikelihood \( l(\beta) \) based on (2.4) \([1, 16, 17]\).

The covariance matrix for \( \hat{\beta} \) can be estimated as the inverse of the observed information matrix for \( l(\beta) \), which can be computed using a method due to Louis \([18]\). Let \( S(\beta) \) denote the vector of partials of \( l(\beta) \) with respect to \( \beta \), and let \( -l(\beta) \) denote the matrix of second partials. Let \( S_0(\beta) \) and \( -l_0(\beta) \) denote the corresponding derivatives of \( l_0(\beta) \). \( l(\hat{\beta}) \) can be computed from the latter according to the formula

\[
l(\hat{\beta}) = E[l_0(\hat{\beta}) - S_0(\hat{\beta})S_0^T(\hat{\beta}) | (y_i, z_i), \ i = 1, ..., n, \ \hat{\beta}],
\]

where the expectation is taken over the unobserved \( x_i, \ i = 1, ..., n \) \([18]\).

The algorithm simplifies when \( f_{y|x} \) belongs to the regular exponential family. This means that, up to an additive data-dependent constant independent of \( \beta \),

\[
(2.9) \quad \log f_{y|x}(y|x, \beta) = w^T(y) G(\beta, x) - A(\beta, x).
\]

Here \( w \) is a sufficient statistic for \( \beta \). The term regular means that the range of the vector valued function \( G \) is restricted to a convex set \( \Omega \) such that \( f_{y|x} \) defines a distribution for all \( G(\beta, x) \) in \( \Omega \). Substituting (2.9) into (2.5) and (2.5) into (2.8) gives

\[
(2.10) \quad Q(\beta|\beta^C) = \sum_i \{w^T(y_i)E_i^C[G(\beta, x)] - E_i^C[A(\beta, x)]\}.
\]

In (2.10) \( E_i^C[G(\beta, x)] \) denotes expectation with respect to the conditional distribution of \( X \) given the data \((y_i, z_i)\) and the current parameter \( \beta^C \):

\[
E_i^C[G(\beta, x)] = \int G(\beta, x)f_{x|yz}(x|y_i, z_i, \beta^C)dx.
\]

Thus the \( E \) step computes for each distinct observation \((y, z)\) the expected values of \( G \) and \( A \) with respect to the posterior distribution of \( X \) obtained from the prior \( f_x \) and the data \((y, z)\), with \( \beta \) at its current value.

When response \( Y \), conditional on \( X = x \), is normally distributed about a linear function of \( x \), then \( G \) and \( A \) are linear in \( x \), and the \( E \) step computes only the expected values of \( X \). In general, however, \( G \) and \( A \) are nonlinear in \( x \), and computing \( Q \) requires numerical integrations for each distinct pair \((y, z)\), and each trial value of \( \beta \) examined in the following \( M \) step. Substantial computing time may be involved unless the data are grouped so that many observations have the same \((y, z)\) values, or unless the needed expectations can be approximated.
An important exceptional case occurs when the sample space of $X$ contains only finitely many values $c_1,\ldots,c_K$ in $\mathbb{R}^m$. Then one can linearize the E step as follows. Introduce the K-dimensional parameter vector $\xi$ with $k^{th}$ component $\xi_k = \beta^Tc_k$, and let the K-dimensional random vector $V = V(X)$ take value the $k^{th}$ unit vector in $\mathbb{R}^K$ when $X$ takes the value $c_k$, $k = 1,\ldots,K$. Then $\beta^T x = \xi^T v$, and for any $r$-dimensional vector valued function $G$, $G(\beta^T x) = \Gamma(\beta)v$, where $\Gamma(\beta)$ is the $rxK$ matrix whose $k^{th}$ column is $G(\beta,c_k)$, $k = 1,\ldots,K$. Thus

$$E_\mathcal{C}[G(\beta,x)] = \Gamma(\beta)E_\mathcal{C}[\nu(x)] \equiv \Gamma(\beta)x,$$

and (2.10) becomes

$$(2.11) \quad Q(\beta, \beta^C) = \sum\{w^T(y_i)\Gamma(\beta)x_i - [a(\beta)]^T x_i\},$$

where $a(\beta) = [A(\beta,c_1),\ldots,A(\beta,c_K)]^T$. Hence the E step requires for each $(y,z)$ only the multinomial probability vector $\chi = E_\mathcal{C}[V(X)]$, whose $k^{th}$ component is the probability that $X = c_k$, conditional on $(y,z)$ and $\beta^C$.

To determine the $\chi_i$'s, let $z$ be an L dimensional column vector of indicators for one of L observed values, and specify $f_{X\mid Z}(x\mid z) = v^T\Theta z$, where $\Theta = (\theta_{kl})$ is a known $K\times L$ probability matrix. Then from Bayes' Theorem the $k^{th}$ component of $\chi$ is

$$(2.12a) \quad \chi^{(k)} = \varepsilon_k\theta_{kl}(\varepsilon^T\Theta z)^{-1},$$

where $\varepsilon$ is a K-dimensional column vector with $k^{th}$ component

$$(2.12b) \quad \varepsilon_k = \exp[w^T(y)G(\beta,c_k) - A(\beta,c_k)].$$

The likelihood equation needed for maximizing (2.11) in the M step is

$$(2.13) \quad \sum_i w^T(y_i) \nabla_\beta \Gamma(\beta)x_i = [\nabla_\beta a(\beta)]^T \sum x_i.$$

Here $\nabla_\beta$ denotes the gradient vector of partial derivatives with respect to $\beta$. Solving (2.13) for $\beta$ may itself involve an iterative procedure, such as Newton-Raphson.

3. Examples. We illustrate the preceding theory with examples from Section 1.

3.1 Dysplasia and DES. Here $Y,X$ and $Z$ are indicators for presence of dysplasia, DES exposure and a positive DES report, respectively. We parameterize $f_{Y\mid X}$ as
\begin{align}
\text{logit } f_{Y|X}(1|x, \beta) &= \beta_0 + \beta_1 x,
\end{align}

where \text{logit } p = p/(1-p). \text{ We also assume as known the probabilities } \theta_z \text{ of DES exposure, given recalled exposure } Z = z, z = 1, 0. \text{ The conditional probability of dysplasia status } y \text{ given recall } z \text{ is then the mixture}

\begin{align}
f_{Y|Z}(y|z, \beta) &= \theta_z f_{Y|X}(y|1, \beta) + (1 - \theta_z) f_{Y|X}(y|0, \beta).
\end{align}

The observed data loglikelihood is

\begin{align}
l(\beta) &= \sum_i \log f_{Y|Z}(y_i|z_i, \beta),
\end{align}

where \( f_{Y|Z} \) is given by (3.2). Several papers (e.g. [19-21]) have shown that erroneously specifying no error (i.e. \( \theta_z = z \)) leads to maximum likelihood estimates for \( \beta_1 \) that are biased toward zero.

The corresponding complete data loglikelihood kernel \( l_0(\beta) \) is the usual logistic loglikelihood, given by (2.9) with \( w = y, G(\beta, x) = \beta_0 + \beta_1 x, \) and \( A(\beta, x) = \log[1 + \exp(\beta_0 + \beta_1 x)]. \) Since \( x \) assumes \( K = 2 \) values \( c_1 = 0 \) and \( c_2 = 1, \) \( x \) is the binomial probability vector \( (1 - \pi, \pi)^T \) where \( \pi \) is the probability of DES exposure, conditional on disease and recall status, and on the current parameter estimate \( \beta^C. \) From (2.12),

\begin{align}
\pi &= \frac{1 + [(\theta_2)^{-1} - 1] \exp(-\beta_1^C y)[1 + \exp(\beta_0^C)]^{-1}[1 + \exp(\beta_0 + \beta_1^C)]^{-1}}{1 + [(\theta_2)^{-1} - 1] \exp(-\beta_1^C y)[1 + \exp(\beta_0^C)]^{-1}[1 + \exp(\beta_0^C + \beta_1^C)]^{-1}}.
\end{align}

The E step involves computing four \( \pi \) values, corresponding to the four values assumed by the pair \((y, z)\).

In this example,

\begin{align}
\Gamma(\beta) = (\beta_0, \beta_0 + \beta_1)^T,
\end{align}

and

\begin{align}
\alpha(\beta) = \{\log[1 + \exp(\beta_0)], \log[1 + \exp(\beta_0 + \beta_1)^T]\}.
\end{align}

Therefore the M step likelihood equation (2.13) determines \((\beta_0^{C+1}, \beta_1^{C+1})\) as the usual logistic estimator

\begin{align}
\beta_0^{C+1} &= \log \left[ \frac{\sum y_i(1 - \pi_i)}{\sum (1-y_i)(1 - \pi_i)} \right],
\end{align}

\begin{align}
\beta_1^{C+1} &= \log \left[ \frac{\sum y_i \pi_i}{\sum (1-y_i)(1 - \pi_i)} \right]/\left[ \frac{\sum (1-y_i) \pi_i}{\sum y_i(1 - \pi_i)} \right],
\end{align}

with each woman’s DES exposure indicator \( x_i \) replaced by her probability \( \pi_i \) of exposure, given her dysplasia and DES recall status.

3.2 Colorectal cancer and diet. Here \( Y \) is an indicator for case (vs. control) status, and \( X \) and \( Z \) represent actual and recalled average daily nutrient in the year before interview. For simplicity we consider an unmatched case-control
study consisting of a random sample \( z_{y1}, \ldots, z_{yn} \) from each of the distributions \( f_{X|Z}(z|y) \), \( y = 0,1 \). We make the standard discriminant analysis assumption that the joint distribution of \( X \) and \( Z \), conditional on disease status \( y \), is bivariate normal with mean

\[
(3.2a) \quad (\mu + \Delta y, \mu + \Delta y) ,
\]

and covariance matrix \( \Sigma \). Further we assume that

\[
(3.2b) \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 + \tau \end{pmatrix}, \quad \sigma_x^2 > 0, \quad \tau > 0.
\]

Interest centers on testing the null hypothesis that \( \Delta = 0 \), and on estimating the standardized difference \( \Delta/\sigma_x \).

Standard normal theory [22] verifies the basic assumption (2.0) of independence between \( Y \) and \( Z \), conditional on the value \( x \) assumed by \( X \). Indeed \( f_{Z|X,Y} = f_{Z|X} \) is given by the normal distribution with mean \( x \) and variance \( \tau \). The distribution \( f_{Z|Y} \) governing the observed data is normal with mean \( \mu + \Delta y \) and variance \( \sigma_x^2 + \tau \). This variance is the sum of the between-person variance component \( \sigma_x^2 \), representing the population variance of actual nutrient intakes, and the within-person component \( \tau \), representing the variance of recall errors. Thus in this special case of joint normality for \( X \) and \( Z \), the observed data distribution is available in closed form, and the EM algorithm is not needed.

The usual estimate of \( \Delta/\sigma_x \) is given by the difference \( z_1 - z_0 \) between case and control means, divided by the square root of the pooled variance estimate. This estimate converges in probability to \( \Delta/(\sigma_x^2 + \tau)^{1/2} \). Therefore it underestimates \( \Delta/\sigma_x \) by the factor \( \rho^{1/2} \), where \( \rho = \sigma_x^2/(\sigma_x^2 + \tau) \) is the intraclass correlation coefficient, i.e. the proportion of total variation in reported intakes due to population heterogeneity of actual intakes.

To see how this downward bias can apply to estimates for \( \beta_1 \) obtained from unconditional logistic analysis of case control studies, we invoke the normality assumptions (3.2) and Bayes Theorem to obtain the logistic distribution (3.1) for \( f_{Y|X} \) with \( \beta_1 = \Delta/\sigma_x^2 \) [23]. The same arguments give \( f_{Y|Z} \) as logistic with \( \beta_1 \) replaced by \( \beta_1^* = \Delta/(\sigma_x^2 + \tau) \). Anderson [24] and Prentice and Pyke [25] have shown that one obtains consistent estimates for \( \beta_1 \) from complete data \( x_{yi}, y = 0,1 \), by maximizing the likelihood based on \( f_{Y|X} \) as if the data were obtained from a prospective or cross-sectional study. When recalled intakes \( y_i \)
replace the actual $x_{yi}$, these estimates converge in probability to the deflated coefficient $\beta_1^* = \beta_1 \rho$.

Wu *et al.* [26] estimated the variance components $\sigma_X^2$ and $\tau$ and the intraclass correlation coefficient $\rho$ for several nutrients by administering repeat dietary history questionnaires to the same healthy white middle-aged U.S. subjects. The estimates for $\rho$ ranged from 83% to 21%, depending on the nutrient. They showed that these values lead to substantial power loss for the moderate to small odds-ratios expected of dietary-chronic-disease associations. The results suggest that case-control studies of diet be restricted to populations with larger dietary heterogeneity (i.e. larger $\sigma_X^2$) than typically found in U.S. populations, since reduction of the recall error variance $\tau$ by averaging repeated dietary assessments is infeasible.

The normality assumptions in this example make it special in two ways: i) the observed data loglikelihood is easily maximized without special numerical algorithms; ii) covariate errors lead to estimates for the regression coefficients that are deflated toward zero. These properties do not hold in general.

3.3 Breast cancer survival and extent of disease. We shall use the EM algorithm to investigate the effects of covariate misclassification and grouping on the relationship between survival time and disease stage for 112 women aged 45-49 diagnosed with breast cancer in the San Francisco Bay Area during the period 1972 to 1977. To simplify the presentation, we suppose first that each woman's time to death is uncensored by loss-to-followup or study termination. Then her data consist of a pair $(y,z)$, with $y$ representing her time to death, and $z$ representing an $L$-dimensional stage indicator vector that may be incomplete.

We assume that a woman's correct complete stage indicator $x^T = (x^{(1)},\ldots,x^{(K)})$ and her observed indicator $z^T = (z^{(1)},\ldots,z^{(L)})$ are related by

$$f_{X|Z}(x|z) = x^T \Theta z.$$  

Here $\Theta$ is a known $K \times L$ matrix of classification probabilities, with $\theta_{kl}$ denoting the probability that a woman observed in stage $i$ is truly in stage $k$. (In general, the matrix could vary from woman to woman.)

We also assume that the death times $y$ arise independently from an exponential density whose hazard rate depends on the corresponding stage indicators $x$ via the form $\exp(\beta^T x)$. Thus

$$\log f_{Y|X}(y|x,\beta) = \beta^T x - y \exp(\beta^T x),$$

and the exponential family form (2.9) holds with $w(y) = y$, $G(\beta,x) = -\exp(\beta^T x)$, and $A(\beta,x) = -\beta^T x$. Moreover $v = x$, $\xi = \beta$, $\Gamma(\beta) = [-\exp(\beta_1),\ldots,-\exp(\beta_K)]^T$, and $\alpha(\beta) = -\beta$. Substituting these values into (2.11) gives the expected complete data loglikelihood.
\[
Q(\beta|\beta^C) = \beta^T \sum x_i - \sum y_i \sum_k x_i(k) \exp(\beta_k)
\]

where \(x_i^{(k)}\) is the probability that the \(i^{th}\) woman was in stage \(k\) at diagnosis, given her death time \(y\), her observed stage \(l\), and the current parameter \(\beta^C\). From (2.12) and (3.3), this probability is

\[
(x_i^{(k)}) = \exp[\beta_k^C - y] \exp(\beta_k^C) \theta_{kl} \left\{ \sum_j \exp[\beta_j^C - y] \exp(\beta_j^C) \theta_{jl} \right\}^{-1}.
\]

The M-step likelihood equation (2.13) is

\[
\sum x_i - \sum y_i \sum_k x_i(k) \exp(\beta_k)
\]

with solution

\[
\exp(\beta_k^C + 1) = \frac{\sum_i x_i^{(k)}}{\sum_i x_i^{(k)} \theta_{ij}}.
\]

Equation (3.6) shows that the updated death rate \(\exp(\beta_k^C + 1)\) is the reciprocal of a weighted average time in stage \(k\), with the \(i^{th}\) woman’s contribution weighted by the proportion \(x_i^{(k)}, 0 \leq x_i^{(k)} \leq 1, k = 1, \ldots, K\). In the absence of misclassification or grouping, the \(x_i = x_i\) are the multinomial indicators for stage at diagnosis, and the estimated death rate \(\hat{\exp(\beta_k)}\) is the usual reciprocal of average woman-months of survival time for stage \(k\).

When a woman’s death time \(U\) can be censored by a random variable \(C\), we observe for her a triple \((t, \delta, z)\), where \(t = \min(u, c)\) and \(\delta\) is an indicator for the event that \(t = u\). Both response \(y = (t, \delta)^T\) and covariate \(z\) are incomplete. We assume that the complete observations \((u_i, c_i, x_i, z_i)\) are a random sample of size \(n\) from a distribution function that satisfies

\[
f_{UCXZ}(u, c, x, z) = f_{U|X}(u|x, \beta)f_{X|Z}(x|z)f_{C|Z}(c|z)f_Z(z).
\]

The factorization (3.7) implies that censoring times are stochastically independent of death times, given \(x\) and \(z\). This assumption is basic to regression analysis of censored data. The form (3.7) also implies that, conditional on a woman’s observed stage, her censoring time is independent of her true stage vector \(x\). This second assumption, which may hold only approximately in practice, assures that the expected values of \(X\) obtained at each E step do not depend on the censoring mechanism.

We infer \(\beta\) from the loglikelihood \(l(\beta)\) of (2.4). The \(i^{th}\) summand of (2.4) is

\[
\log f_{Y|Z}(y_i|z_i, \beta), \text{ where } f_{Y|Z}(y|z, \beta) \text{ is obtained by integrating the first three terms of (3.7) over all values of } x \text{ and over the space } S(y) \text{ of pairs } (u, c) \text{ corresponding to } y:
\]

\[
f_{Y|Z}(y|z, \beta) = \int f_{X|Z}(x|z) \left[ \int_{S(y)} f_{U|X}(u|x, \beta)f_{C|Z}(c|z)dudc \right] dx.
\]
The integral in square brackets equals

\[(3.9) \quad \left[ \int_{t \mid X(t \mid x, \beta)} \delta \left[ F_{U \mid X(t \mid x, \beta)} \right] \right]^{\delta - t} \int_{t \mid X(t \mid x, \beta)} f_{U \mid X(t \mid x, \beta)} f_{C \mid Z(t \mid z)} \right]^{1 - \delta}, \]

where \( F(t) = \int_{t}^{\infty} f(s) ds \) is the survivor function. Substitution of (3.9) into (3.8) and (3.8) into (2.4) gives, up to an additive term,

\[(3.10) \quad l(\beta) = \sum \log \int f_{X \mid Z}(x \mid z \mid i) \left[ f_{U \mid X}(t_i \mid x, \beta) \delta_i F_{U \mid X}(t_i \mid x, \beta) \right]^{1 - \delta_i} dx. \]

The corresponding complete data loglikelihood is

\[(3.11) \quad l_o(\beta) = \sum \log f_{U \mid X}(t_i \mid x_i, \beta) \left[ \delta_i \log f_{U \mid X}(t_i \mid x_i, \beta) \right] F_{U \mid X}(t_i \mid x_i, \beta). \]

When \( f_{U \mid X} \) is the exponential density (3.4), (3.11) reduces to (2.9) with \( w = y = (t, \delta)^T \), \( G(\beta, x) = (\exp(\beta^T x), \beta^T x)^T \), and \( A(\beta, x) = 0 \). As in the uncensored case, \( v = x \) and \( \xi = \beta \). The 2 \times K matrix \( \Gamma(\beta) \) has \( k \)th column \( [\exp(\beta_k), \beta_k]^T \). From (2.11)

\[Q(\beta \mid \beta^C) = \beta^T \sum_i \delta_i x_i - \sum_i t_i \sum_k x_i \exp(\beta_k),\]

with likelihood equation

\[\sum_i \delta_i x_i \exp(\beta_k) = \sum_i t_i x_i \exp(\beta_k), \quad k = 1, ..., K.\]

Thus \( \beta^{C+1} \) satisfies

\[(3.12) \quad \exp(\beta_k^{C+1}) = \sum \delta_i x_i \exp(\beta_k) / \sum t_i x_i \exp(\beta_k).\]

The updated death rates (3.12), which generalize the uncensored ones (3.6), represent the number of deaths in stage \( k \) divided by the total time in the stage, with each woman contributing a fractional part \( x_i \) of her time (and death) to the stage. From (2.12)

\[x_i = \theta_{kl} \exp(\delta \beta_{kl} - t \exp \beta_{kl}) / \sum_j \theta_{jl} \exp(\delta \beta_{lj} - t \exp \beta_{lj}).\]

a generalization of (2.5).

Table 1 shows numbers of women, person-months of survival time, and deaths in each of five stages. Stages 1-4 are defined by tumor size and number of involved lymph nodes (see [27] for details). Stage 5 consists of women diagnosed with metastatic disease. An additional "Stage" 2.5 consists of women known only to be in Stages 2 or 3. "Stage" 2.5 accounts for a large proportion of women, person-months and deaths. The death rates are virtually
indistinguishable for Stages 1 and 2, and for Stages 2.5, 3 and 4. Those with metastatic disease exhibit substantially higher rates, with a rate-ratio of 32.5 relative to women diagnosed in Stage 1.

Table 1

Survival vs. Disease Stage for Women Aged 45-49 at Breast Cancer Diagnosis

<table>
<thead>
<tr>
<th>Stage</th>
<th>Number of Women</th>
<th>Number of Person-Months</th>
<th>Number of Deaths</th>
<th>Death Rate (Deaths/10^3 Woman-Months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>967</td>
<td>2</td>
<td>2.068</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>2314</td>
<td>5</td>
<td>2.161</td>
</tr>
<tr>
<td>2.5a</td>
<td>46</td>
<td>3141</td>
<td>17</td>
<td>5.412</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>754</td>
<td>4</td>
<td>5.305</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>362</td>
<td>2</td>
<td>5.510</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>88</td>
<td>6</td>
<td>68.182</td>
</tr>
<tr>
<td>Total</td>
<td>112</td>
<td>7626</td>
<td>36</td>
<td>4.721</td>
</tr>
</tbody>
</table>

a) Women in "Stage" 2.5 were known only to be in either Stage 2 or Stage 3.

Table 2 shows the classification matrix \( \Theta \) introduced in (3.3) and represented as a function of a parameter \( \delta, 0 \leq \delta < 1 \). The value \( \delta = 0 \) corresponds to no misclassification for women observed in Stages 1-5, with those in Stage 2.5 having 67% probability of being in Stage 2. Positive values for \( \delta \) indicate positive probabilities of understaging for women in Stages 1-4; it is assumed that no one was overstaged.

Table 2

CLASSIFICATION MATRIX \( \Theta \)

<table>
<thead>
<tr>
<th>Actual Stage x</th>
<th>Observed Stage z</th>
<th>1</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-(\delta)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>(\delta/2)</td>
<td>1-(\delta)</td>
<td>.67-(\delta/4)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>(\delta/3)</td>
<td>(\delta/2)</td>
<td>.33-(\delta/4)</td>
<td>1-(\delta)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>(\delta/6)</td>
<td>(\delta/3)</td>
<td>(\delta/4)</td>
<td>2(\delta/3)</td>
<td>1-(\delta)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>(\delta/6)</td>
<td>(\delta/4)</td>
<td>(\delta/3)</td>
<td>(\delta)</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Tables 3 and 4 show the values of $\beta^c$ and the maximized observed data loglikelihood for the first 12 iterations of the EM algorithm corresponding to $\delta = 0$ and $\delta = 0.10$, respectively. The initial values $\beta^0$ were taken to be the logs of the death rates in Table 1, ignoring the death rate for Stage 2.5. The maximized loglikelihood at the 12th iteration is higher for $\delta = 0$ than $\delta = 0.10$. Neither of them exceeds the maximized loglikelihood of -215.2398 for the model with one additional parameter obtained by fitting six separate death rates to the data of Table 1.

### Table 3

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_4$</th>
<th>$\beta_5$</th>
<th>Observed Data loglikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-6.181</td>
<td>-6.137</td>
<td>-5.239</td>
<td>-5.198</td>
<td>-2.686</td>
<td>-217.3847</td>
</tr>
<tr>
<td>1</td>
<td>-6.181</td>
<td>-5.820</td>
<td>-4.931</td>
<td>-5.198</td>
<td>-2.686</td>
<td>-216.2042</td>
</tr>
<tr>
<td>4</td>
<td>-6.181</td>
<td>-5.825</td>
<td>-4.911</td>
<td>-5.198</td>
<td>-2.686</td>
<td>-216.2018</td>
</tr>
<tr>
<td>5</td>
<td>-6.181</td>
<td>-5.822</td>
<td>-4.907</td>
<td>-5.198</td>
<td>-2.686</td>
<td>-216.2017</td>
</tr>
<tr>
<td>7</td>
<td>-6.181</td>
<td>-5.823</td>
<td>-4.906</td>
<td>-5.198</td>
<td>-2.686</td>
<td>-216.2017</td>
</tr>
<tr>
<td>8</td>
<td>-6.181</td>
<td>-5.824</td>
<td>-4.906</td>
<td>-5.198</td>
<td>-2.686</td>
<td>-216.2017</td>
</tr>
</tbody>
</table>

Death Rate: $2.068 \times 10^{-3}$

While the death rates in Table 3 are identical to the ones in Table 1 for Stages 1, 4 and 5, those in Stages 2 and 3 are elevated. The increase in Stage 2 reflects the fractional contribution of deaths and person-months from women in Stage 2.5, whose mortality experience was less favorable than that of women in Stage 2. The increase in Stage 3 reflects the contribution from those women in Stage 2.5 with shorter survival times, and puts women in Stage 3 at greater risk than those in Stage 4.
Table 4

Iterations of EM Algorithm for Data in Table 1 using Classification Matrix $\Theta$ of Table 2 with $\delta = 0.10$

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_4$</th>
<th>$\beta_5$</th>
<th>Observed loglikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-6.181</td>
<td>-6.137</td>
<td>-5.239</td>
<td>-5.198</td>
<td>-2.686</td>
<td>218.2181</td>
</tr>
<tr>
<td>1</td>
<td>-6.238</td>
<td>-5.911</td>
<td>-5.003</td>
<td>-5.201</td>
<td>-2.862</td>
<td>217.4241</td>
</tr>
<tr>
<td>2</td>
<td>-6.268</td>
<td>-5.899</td>
<td>-4.972</td>
<td>-5.281</td>
<td>-2.894</td>
<td>217.3955</td>
</tr>
<tr>
<td>3</td>
<td>-6.272</td>
<td>-5.900</td>
<td>-4.960</td>
<td>-5.319</td>
<td>-2.903</td>
<td>217.3907</td>
</tr>
<tr>
<td>4</td>
<td>-6.272</td>
<td>-5.903</td>
<td>-4.953</td>
<td>-5.336</td>
<td>-2.906</td>
<td>217.3895</td>
</tr>
<tr>
<td>5</td>
<td>-6.272</td>
<td>-5.905</td>
<td>-4.949</td>
<td>-5.343</td>
<td>-2.907</td>
<td>217.3892</td>
</tr>
<tr>
<td>6</td>
<td>-6.272</td>
<td>-5.907</td>
<td>-4.946</td>
<td>-5.346</td>
<td>-2.908</td>
<td>217.3890</td>
</tr>
<tr>
<td>7</td>
<td>-6.272</td>
<td>-5.909</td>
<td>-4.945</td>
<td>-5.348</td>
<td>-2.908</td>
<td>217.3889</td>
</tr>
<tr>
<td>8</td>
<td>-6.272</td>
<td>-5.910</td>
<td>-4.944</td>
<td>-5.349</td>
<td>-2.908</td>
<td>217.3889</td>
</tr>
<tr>
<td>9</td>
<td>-6.272</td>
<td>-5.911</td>
<td>-4.943</td>
<td>-5.349</td>
<td>-2.908</td>
<td>217.3889</td>
</tr>
<tr>
<td>10</td>
<td>-6.272</td>
<td>-5.911</td>
<td>-4.942</td>
<td>-5.349</td>
<td>-2.908</td>
<td>217.3889</td>
</tr>
<tr>
<td>11</td>
<td>-6.272</td>
<td>-5.911</td>
<td>-4.942</td>
<td>-5.349</td>
<td>-2.908</td>
<td>217.3889</td>
</tr>
<tr>
<td>12</td>
<td>-6.272</td>
<td>-5.911</td>
<td>-4.942</td>
<td>-5.349</td>
<td>-2.908</td>
<td>217.3889</td>
</tr>
<tr>
<td>Death Rate $e^{\beta x 10^3}$</td>
<td>1.888</td>
<td>2.707</td>
<td>7.140</td>
<td>4.753</td>
<td>54.585</td>
<td></td>
</tr>
</tbody>
</table>

Comparison of the death rates in Tables 3 and 4 shows that taking account of a 10% probability of understaging has the surprising effect of lowering the rates in all stages. (This phenomenon has been dubbed "the Will Rogers effect", because of his remark that when the Okies moved from Oklahoma to California, they raised the average I.Q. in both states.) Rates in the lower stages are decreased due to loss to higher stages of part of the mortality experience of women who died early. Rates in the higher stages are decreased due to gain of experience from women observed in the lower stages, who apparently survived longer than did women observed in the higher recipient stage. If $\delta = 10\%$ were the correct model, misspecifying $\delta = 0$ can be shown to produce death rate estimates for Stage 5 that are biased toward zero. Therefore the elevated Stage 5 rates and elevated loglikelihood for the model $\delta = 0$, relative to the model $\delta = 10\%$, suggest that the latter fits poorly.
4. **Discussion.** The preceding examples show that the EM algorithm provides a feasible and heuristically appealing procedure for estimating regression coefficients relating disease to true exposures by maximizing the likelihood of the observed, incomplete data. In these examples the updated estimates obtained at each M step were available in closed form. More generally, standard software packages such as GLIM-3 [28] can be embedded in the algorithm as subroutines. Usually the M step is less computer intensive than the E step. The latter is particularly time consuming when response y or observed covariates z are continuous, necessitating separate numerical integrations for each individual under study. Further work is needed to shorten computing time needed by the algorithm, and to compare its performance to that of alternative "one-pass-through-the-data" procedures [29].

It should be emphasized that there is no satisfactory substitute for complete or nearly complete data. The conditional independence assumption used for response and observed covariates may fail in practice, and the models assumed for true covariates given the observed ones may be difficult to verify. Nevertheless, the procedures described above allow one to test robustness of inferred associations against plausible departures from the simple assumptions typically used to deal with incomplete covariates, and to make alternative, more realistic assumptions when warranted. Therefore the procedures provide a useful tool for analyzing observational data.

**REFERENCES**


