Generalized linear models with a coarsened covariate

Stuart Lipsitz,
Medical University of South Carolina, Charleston, USA

Michael Parzen,
University of Chicago, USA

Sundar Natarajan,
Medical University of South Carolina, Charleston, USA

Joseph Ibrahim
University of North Carolina at Chapel Hill, USA

and Garrett Fitzmaurice
Harvard School of Public Health, Boston, USA

[Received April 2001. Revised July 2003]

Summary. We consider generalized linear models with a coarsened covariate. The term 'coarsened' is used here to refer to the case where the exact value of the covariate of interest is not fully observed. Instead, only some set or grouping that contains the exact value is observed. In particular, we propose a likelihood-based method for estimating regression parameters in a generalized linear model relating the mean of the outcome to covariates. We outline Newton–Raphson and EM algorithms for obtaining maximum likelihood estimates of the regression parameters. We also compare and contrast this likelihood-based approach with two somewhat ad hoc procedures: a complete-case analysis in which individuals with coarsened data are excluded and estimation is based on the remaining 'complete cases', and a coarsened data regression model in which the covariate values for all the complete cases are coarsened and then included in a regression model relating the mean to the coarsened covariate. The methodology that is presented is motivated by coarsened data on the racial–ethnicity categorization of patients in the US's National Ambulatory Medical Care Survey, a study to examine the medical care that is provided to a patient in a physician's office. In this study, the outcome of interest is the level of tests (none, non-invasive tests or invasive tests) ordered for the patient at the doctor's visit. One of the covariates of interest is the patient's four-level discrete covariate comprised of four racial–ethnicity categories: white–Hispanic, white–non-Hispanic, African-American–Hispanic and African-American–non-Hispanic. However, of the 19095 patients in the sample, 14955 (or 78%) have the exact category of race–ethnicity recorded and 4140 (or 22%) have race–ethnicity coarsened. For the latter group of 4140 individuals, the ethnicity is not recorded, but we know that 3683 are white and 457 are African-American.

Keywords: Coarsened data mechanism; EM algorithm; Logistic regression; Maximum likelihood estimation; Newton–Raphson algorithm

Address for correspondence: Stuart Lipsitz, Department of Biometry and Epidemiology, Medical University of South Carolina, Suite 1148, 135 Rutledge Avenue, PO Box 250551, Charleston, SC 29425, USA.
E-mail: lipsitzs@musc.edu

© 2004 Royal Statistical Society 0035–9254/04/53279
1. Introduction

Interval-censored data commonly arise in many medical and health-related studies. With an interval-censored variable, the value of the variable is known to fall between two values, but the exact value is unknown. Although interval-censored variables are often continuous, discrete or categorical variables with very similar properties also arise in many health-related studies. The main motivation for the methodology that is proposed in this paper arises from data from the National Ambulatory Medical Care Survey (Tenney et al., 1974), a study by the US Department of Health and Human Services to examine the medical care that is provided to a patient in a physician's office. In this study, we have data on an office visit of 19095 patients. Data are obtained on the patient's symptoms, the physician's diagnoses and medications and tests ordered or provided. The study also provides data on the demographic characteristics of patients, such as gender, age, race and ethnicity. In this study, the outcome of interest is the invasive level of tests (three levels: none, non-invasive tests and invasive tests) ordered for the patient at the visit. The main covariates of interest are age (three levels: 21 years or less, 22–59 years and 60 years and over), gender, insurance (three levels: private insurance, Medicaid or Medicare and none), reason for visit (two levels: acute problem versus other) and race–ethnicity (four levels: white–Hispanic, African-American–Hispanic and African-American–non-Hispanic). The last covariate was formed by combining race (two categories: white and African-American) and ethnicity (two categories: Hispanic and non-Hispanic); race and ethnicity are two separate questions on the survey, but the investigators were interested in the composite four-level covariate of race–ethnicity. Of the 19095 patients in the sample, 14955 (or 78%) have the exact category of race–ethnicity recorded and 4140 (or 22%) have race–ethnicity coarsened. For the latter group of 4140 individuals, we know that 3683 are white, and 457 are African-American, but the ethnicity is not recorded. In this particular example, the race–ethnicity covariate is considered to be a 'coarsened' variable (Heitjan and Basu, 1996). By the term coarsened, we mean that for at least some individuals in the sample the exact value of the categorical variable is not observed, but only that the true covariate value is in some subset of the possible values. In this example, the race–ethnicity covariate for an individual who is white with ethnicity unrecorded could be either white–Hispanic or white–non-Hispanic, but not African-American–Hispanic or African-American–non-Hispanic. This type of categorical data is sometimes also referred to as 'partially categorized categorical data' (Hocking and Oxspring, 1974).

Data from 30 participants in the National Ambulatory Medical Care Survey are displayed in Table 1. The main analytic goal is to relate the ordinal outcome invasiveness of the test (1, invasive tests; 2, non-invasive tests; 3, none) to these covariates. Furthermore, note that it is of interest to describe the dependence of the ordinal outcome on all four levels of race–ethnicity (and on the other covariates). However, recall that race–ethnicity is coarsened for approximately 22% of the study participants.

There are three general approaches for estimating the regression parameters in a model with a covariate that is coarsened. First, all individuals with the coarsened version of the covariate could be excluded from the analysis. The resulting analysis is based only on those individuals with complete covariates. For the example that was introduced earlier, the 4140 participants with the coarsened version of race–ethnicity could be excluded from the analysis, with estimation of the regression parameters based on the remaining 14955 participants. This general approach for handling coarsened data can be referred to as a 'complete-case' analysis. However, a complete-case analysis could yield biased estimates of the regression parameters unless the covariate is coarsened completely at random (Heitjan and Basu, 1996). This is a strong and often unrealistic assumption. Furthermore, even when the coarsening completely at random...
Table 1. Data from the National Ambulatory Medical Care Survey

| Subject | Test   | Reason          | Insurance               | Age (years) | Gender   | Race       | Ethnicity | """" |
|---------|--------|-----------------|-------------------------|-------------|----------|------------|-----------|
| 1       | None   | Other Private   | 21–59                   | Female      | White    | Not recorded|           |
| 2       | Invasive| Other Medicaid or Medicare | ≥60 | Female | White | Not recorded |           |
| 3       | Invasive| Other Medicaid or Medicare | ≥60 | Male   | White | Not recorded |           |
| 4       | Invasive| Acute Private   | 21–59                   | Female      | White    | Not recorded|           |
| 5       | None   | Other Medicaid or Medicare | ≥60 | Female | White | Not recorded |           |
| 6       | None   | Other Private   | <21                     | Male        | White    | Not recorded|           |
| 7       | None   | Other Medicaid or Medicare | <21 | Female | White | Not recorded |           |
| 8       | Invasive| Other None      | ≥60                     | Male        | White    | Not recorded|           |
| 9       | None   | Other Medicaid or Medicare | 21–59 | Female | African-American | Not recorded |           |
| 10      | Invasive| Acute None     | 21–59                   | Female      | African-American | Not recorded |           |
| 11      | Invasive| Other Private   | 21–59                   | Male        | African-American | Not recorded |           |
| 12      | Invasive| Other Private   | ≥60                     | Female      | African-American | Not recorded |           |
| 13      | Invasive| Other None      | <21                     | Female      | African-American | Not recorded |           |
| 14      | Non-invasive| Other Medicaid or Medicare | 21–59 | Female | White | Hispanic |           |
| 15      | Invasive| Other None      | <21                     | Female      | White | Hispanic |           |
| 16      | None   | Acute Private   | 21–59                   | Female      | White | Hispanic |           |
| 17      | Invasive| Acute Medicaid or Medicare | <21 | Male   | White | Hispanic |           |
| 18      | None   | Other Private   | 21–59                   | Female      | White | Non-Hispanic |           |
| 19      | Non-invasive| Other Private | 21–59 | Female | White | Non-Hispanic |           |
| 20      | Invasive| Other Private   | 21–59                   | Male        | White | Non-Hispanic |           |
| 21      | Non-invasive| Acute Private | <21                     | Female      | White | Non-Hispanic |           |
| 22      | None   | Other Medicaid or Medicare | ≥60 | Female | White | Non-Hispanic |           |
| 23      | Invasive| Other Medicaid or Medicare | ≥60 | Male   | White | Non-Hispanic |           |
| 24      | None   | Acute Private   | <21                     | Female      | White | Non-Hispanic |           |
| 25      | Invasive| Acute Private   | 21–59                   | Female      | African-American | Hispanic |           |
| 26      | Invasive| Other None      | <21                     | Female      | African-American | Hispanic |           |
| 27      | None   | Other Private   | 21–59                   | Female      | African-American | Hispanic |           |
| 28      | Non-invasive| Acute None | 21–59 | Male   | African-American | Non-Hispanic |           |
| 29      | None   | Other Private   | ≥60                     | Male        | African-American | Non-Hispanic |           |
| 30      | None   | Acute Private   | ≥60                     | Female      | African-American | Non-Hispanic |           |

assumption is tenable, and the complete-case analysis yields estimates that are asymptotically unbiased, there may be a substantial loss of efficiency. For example, a complete-case analysis of the National Ambulatory Medical Care Survey ignores data on approximately 22% of the participants in the study.

An alternative approach is to use the data from the entire sample to model directly the dependence of the mean on the coarsened version of the covariate (and on the other covariates), i.e., instead of modelling the dependence of the outcome on all levels of the categorical covariate, we model the dependence of the outcome on the coarsened levels. For example, in the National Ambulatory Medical Care Survey, we know that, of the patients with both race and ethnicity recorded, there are 13535 whites and 1420 African-Americans; of the patients with race recorded but ethnicity not recorded, there are 3683 whites and 457 African-Americans. In the entire sample of 19095 patients, there are 17218 whites and 1877 African-Americans. Thus, instead of using the four-level covariate race–ethnicity, the two-level coarsened covariate race (white and African-American) can then be used in all subsequent regression analyses based on the entire sample, without using ethnicity. The chief advantage of this approach is that data from all individuals contribute to the analysis. However, as discussed in greater detail in Section 2, this approach implicitly makes a strong assumption that there is no ethnicity effect, and one that cannot be verified easily from the data at hand. When the assumption that underlies this
approach is not tenable, this method can yield biased estimates of the regression parameters. The final approach, and the one that is the main focus of this paper, is to use maximum likelihood estimation under the assumption that the covariate is coarsened at random (Heitjan and Basu, 1996). Under the assumption of coarsening at random, the conditional probability of the observed coarsening indicator (an indicator of whether the data are coarsened), given the complete data and the parameters, is constant across all possible values that the coarsened variable can take on. Under coarsening at random, we can ignore the coarsening mechanism when using maximum likelihood. However, when using maximum likelihood, we need to specify correctly the conditional distribution of the possibly coarsened covariate given the other covariates.

In the next section we introduce some notation and propose a likelihood-based method for estimating regression parameters in a generalized linear model relating the mean of the outcome to covariates. We note that we can also treat the data in this example as a missing covariate problem with ‘ethnicity’ missing, instead of race–ethnicity as coarsened. However, missingness is a special form of coarsening, so using the coarsened data approach is more general. For this particular data set, treating ‘ethnicity’ as missing at random (Rubin, 1976), or race–ethnicity as coarsened at random will give the exact same maximum likelihood estimate of the regression parameters, since the likelihoods will be proportional to each other.

2. Notation and model

Consider a regression problem with independent observations on \( n \) subjects (or experimental units). Let \( Y_i (i = 1, \ldots, n) \) denote the response variable of interest for the \( i \)th subject. In addition, there is a \( P \times 1 \) vector of covariates \( x_i = (x_{i1}, \ldots, x_{ip})' \) (typically \( x_{i1} = 1 \) for all \( i \)) that is fully observed and a discrete covariate \( z_i \), with \( K \) levels. Later, we shall allow \( z_i \) to be possibly coarsened. Note that the \( K \) levels of \( z_i \) can be represented by \( K \) indicator random variables \( z_{ik} \), where \( z_{ik} = 1 \) if \( z_i = k \) and \( z_{ik} = 0 \) otherwise, \( k = 1, \ldots, K \). Thus, we can form a \( K \times 1 \) vector \( z_i = (z_{i1}, \ldots, z_{iK})' \) for each subject. The primary objective is to describe the dependence of \( \mu_i = E(Y_i) \) on the covariates \( x_i \) and \( z_i \) by using a generalized linear model.

Next, it is assumed that the mean response \( \mu_i \) is related to the covariates through

\[
g(\mu_i) = g\{E(y_i|x_i, z_{i1}, \ldots, z_{iK-1}, \beta)\} = \beta_1 x_i + \beta_2 z_i = \beta_1 x_i + \beta_{21} z_{i1} + \ldots + \beta_{2K} z_{iK},
\]

where \( \beta = (\beta_1', \beta_2')' \) is the regression parameter vector of primary interest and \( g(\cdot) \) is a known link function (a continuous, monotone, differentiable function). Often, there is an intercept in the model, so we must put constraints on \( \beta_2 \) for the model to be of full rank; for example, in a reference cell model, we can set \( \beta_{21} = 0 \). The variance of \( Y_i \) is assumed to be a function of the mean,

\[
\text{var}(Y_i) = \phi v(\mu_i),
\]

where \( v(\mu_i) \) is a known variance function (a function of the mean \( \mu_i \)) and \( \phi \) is a dispersion parameter (that may or may not be known). Finally, it is assumed that \( Y_i \) has a distribution in the exponential family; for example, \( Y_i \) may have a normal, gamma, Poisson, binomial or multinomial distribution. If the covariate \( z_i \) is fully observed, the maximum likelihood estimate of \( \beta \) can be obtained by using widely available statistical software for generalized linear models (e.g. procedure GENMOD in SAS).

Next, suppose that the covariate \( z_i \) is coarsened. By the term coarsened we mean that, instead of observing the exact value of \( z_i \), we know that it can equal some subset of \( \{1, \ldots, K\} \). In the
example that was introduced earlier, in which race–ethnicity is coded as 1, white–Hispanic, 2, white–non-Hispanic, 3, African-American–Hispanic, and 4, African-American–non-Hispanic, for participants with race–ethnicity coarsened as ‘white’, we know only that \( z_i \) can equal 1 or 2, but not 3 or 4. In general, in terms of the vector of indicator variables \( z_i \), we observe \( z_i^* = C_i z_i \) instead of \( z_i \), where \( C_i \) is a \((Q \times K)\)-matrix of 1s and 0s, and \( Q < K \). In the example that was introduced earlier, \( z_i \) has four levels with \( z_i = (z_{i1}, z_{i2}, z_{i3}, z_{i4})' \). For individuals with \( z_i \) coarsened to equal 1 or 2,

\[
C_i = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix},
\]

so that

\[
z_i^* = \begin{pmatrix} z_{i1}^* \\ z_{i2}^* \end{pmatrix} = C_i z_i = \begin{pmatrix} z_{i1} + z_{i2} \\ z_{i3} + z_{i4} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.
\]

When the covariate \( z_i \) is coarsened, there are several somewhat \textit{ad hoc} approaches for estimating \( \beta \). One simple approach is to estimate \( \beta \) by using a so-called complete-case analysis, i.e. all subjects with coarsened \( z_i \) are excluded from the analysis and estimation of \( \beta \) is based on the remaining complete cases. However, a complete-case analysis could potentially yield biased estimates of \( \beta \) unless \( z_i \) is coarsened completely at random (Heitjan and Basu, 1996), i.e. the conditional probability of the observed coarsening indicator (an indicator of whether \( z_i \) is coarsened), given \( x_i \) and the parameters, is constant across all possible values of \( z_i \). Furthermore, even when it is reasonable to assume that \( z_i \) is coarsened completely at random, the complete-case analysis will generally be very inefficient, especially when the fraction of coarsened data is large.

An alternative approach is to coarsen \( z_i \) into \( z_i^* \) for all the complete cases with fully observed \( z_i \) and then to relate the mean of \( Y_i \) to \( z_i^* \) instead of \( z_i \). However, this approach will in general result in biased estimates of \( \beta \). To see why, consider the example that was introduced earlier where a regression model of interest is

\[
g\{E(y_i|x_i, z_{i1}, z_{i2}, \beta)\} = \beta_1 x_i + \beta_{21} z_{i1} + \beta_{22} z_{i2} + \beta_{23} z_{i3} + \beta_{24} z_{i4}.
\]

If it is assumed that \( \beta_{21} = \beta_{22} \) and \( \beta_{23} = \beta_{24} \) then the regression model becomes

\[
g\{E(y_i|x_i, z_{i1}, z_{i2}, z_{i3}, \beta)\} = \beta_1 x_i + \beta_{21} (z_{i1} + z_{i2}) + \beta_{23} (z_{i3} + z_{i4})
\]

\[
= \beta_1 x_i + \beta_{21} z_{i1}^* + \beta_{23} z_{i2}^*,
\]

where

\[
z_i^* = \begin{pmatrix} z_{i1}^* \\ z_{i2}^* \end{pmatrix} = \begin{pmatrix} z_{i1} + z_{i2} \\ z_{i3} + z_{i4} \end{pmatrix}.
\]

In this case, \( z_i \) can be coarsened for all those individuals with \( z_i \) fully observed and \( z_i^* \) can be used in the regression model to obtain a consistent estimate of \((\beta_1, \beta_{21}, \beta_{23})\). Furthermore, the maximum likelihood estimate of \((\beta_{21}, \beta_{23})\) (and \( \beta_1 \)) can be obtained by using widely available statistical software for generalized linear models. However, if, in model (3), \( \beta_{21} \neq \beta_{22} \) and/or \( \beta_{23} \neq \beta_{24} \), then, for example, \( \hat{\beta}_{21} \) estimated from the coarsened data regression model given by equation (3) will be biased for both \( \beta_{21} \) and \( \beta_{22} \). Note that the assumption that \( \beta_{21} = \beta_{22} \) and \( \beta_{23} = \beta_{24} \) cannot be verified from the data at hand. For example, use of the complete cases to provide a test of hypothesis \( H_0: \beta_{21} = \beta_{22} \) is problematic, i.e. a complete-case analysis may fail to reject \( H_0 \) because of a lack of power (as a result of the inefficiency of the complete-case analysis) and/or because of bias in the estimation of \( \beta \) if coarsening is not completely at random.
As a result, the coarsened data regression model forces the data analyst to make unverifiable assumptions and also severely restricts the subject-matter questions that can be addressed by the analysis.

To avoid the aforementioned problems with these ad hoc approaches to handling coarsened data, we propose to use maximum likelihood estimation of $\beta$ under the assumption that $z_i$ is coarsened at random (Heitjan and Basu, 1996). With coarsened data, we can define an indicator random variable $R_i$, which equals 1 if $z_i$ is fully observed and 0 if $z_i$ is coarsened. Under the assumption that $z_i$ is coarsened at random, the conditional probability of the observed $R_i$, given $x_i$ and the parameters, is constant across all possible values of $z_i$ that produce the given value of $z_i^*$. Under the assumption that data are coarsened at random, it is not necessary to specify a model for the distribution of $R_i$ to make likelihood-based inferences about $\beta$, i.e. we only need to specify the distribution $p(y_i, z_i|x_i, \gamma)$.

When $z_i$ is coarsened ($R_i = 0$), we observe $(y_i, x_i, z_i^*)$; in contrast, when there is no coarsening ($R_i = 1$) of $z_i$, we observe $(y_i, x_i, z_i)$. Thus, the joint density of the data that we observe for the $i$th individual is $p(y_i, z_i|x_i, \gamma)$ if $z_i$ is fully observed and $p(y_i, z_i^*|x_i, \gamma)$ if $z_i$ is coarsened, where $\gamma$ is the parameter vector characterizing this distribution. The density $p(y_i, z_i|x_i, \gamma)$ can be expressed as

$$p(y_i, z_i|x_i, \gamma) = p(y_i|x_i, z_i, \beta) p(z_i|x_i, \alpha),$$  

(4)

where $p(z_i|x_i, \alpha)$ is the distribution of the possibly coarsened covariate $z_i$, given the fully observed covariates $x_i$. Since $x_i$ is a covariate vector that is always fully observed, we condition on $x_i$ throughout equation (4).

Recall that in this specification of the joint distribution $p(y_i, z_i|x_i, \gamma)$ the main interest is in estimating $\beta$, the vector of parameters from the conditional distribution $p(y_i|x_i, z_i, \beta)$. The parameters $\alpha$ from $p(z_i|x_i, \alpha)$ are largely regarded as nuisance parameters. Although $\alpha$ are nuisance parameters, $p(z_i|x_i)$ must be specified to make likelihood-based inferences about $\beta$. Thus, the $r$th individual’s contribution to the likelihood for $\gamma' = (\beta', \alpha')$ is given by

$$L_i(\gamma) = p(y_i, z_i|x_i, \gamma)^{r_i} p(y_i, z_i^*|x_i, \gamma)^{1-r_i}$$

$$= \{p(y_i|x_i, z_i, \beta) p(z_i|x_i, \alpha)\}^{r_i} p(y_i, z_i^*|x_i, \gamma)^{1-r_i}. $$  

(5)

In the following section, the likelihood for $\gamma' = (\beta', \alpha')$ is described in more detail, and the Newton–Raphson and EM algorithms for obtaining the maximum likelihood estimate are outlined.

3. Maximum likelihood estimation

The likelihood for $(\beta', \alpha')$ requires the specification of the two densities $p(y_i|x_i, z_i, \beta)$ and $p(z_i|x_i, \alpha)$. Recall that it is assumed that $p(y_i|x_i, z_i, \beta)$ has a distribution in the exponential family and follows a generalized linear model, with mean given by equation (1). Given that $z_i$ is discrete, it has a multinomial distribution (conditional on $x_i$) and a multinomial logistic or probit model can be assumed for $p(z_i|x_i, \alpha)$. However, if $z_i$ is ordinal, a cumulative logit or cumulative probit model may be more appropriate. Note that the density $p(y_i, z_i^*|x_i, \gamma)$ can be obtained by simply summing equation (4) over the possible values of $z_i$ that are coarsened. Letting $S_i$ be the set of possible values of $z_i$ that are coarsened,

$$p(y_i, z_i^*|x_i, \gamma) = \sum_{z_i \in S_i} p(y_i, z_i|x_i, \gamma) = \sum_{z_i \in S_i} p(y_i|x_i, z_i, \beta) p(z_i|x_i, \alpha).$$
For the example that was introduced earlier, in which race–ethnicity is coarsened as ‘white’, i.e. \( z_i = 1 \equiv \text{white–Hispanic} \) or \( z_i = 2 \equiv \text{white–non-Hispanic} \), then \( S_i = \{1, 2\} \) and

\[
p(y_i, z_i^* | x_i, \gamma) = p(y_i | x_i, z_i = 1, \beta) \text{pr}(z_i = 1 | x_i, \alpha) + p(y_i | x_i, z_i = 2, \beta) \text{pr}(z_i = 2 | x_i, \alpha).
\]

Then, using equation (5), the log-likelihood to be maximized is

\[
l(\gamma) = \sum_{i=1}^{n} r_i \log \left\{ p(y_i | x_i, z_i, \beta) \text{pr}(z_i | x_i, \alpha) \right\} + (1 - r_i) \log \left\{ \sum_{z_i \in S_i} p(y_i | x_i, z_i, \beta) \text{pr}(z_i | x_i, \alpha) \right\}.
\]

Thus, for the example that was introduced earlier, the log-likelihood to be maximized is

\[
l(\gamma) = \sum_{i=1}^{n} r_i \log \left\{ p(y_i | x_i, z_i, \beta) \text{pr}(z_i | x_i, \alpha) \right\} + \sum_{i=1}^{n} (1 - r_i) \left( z_{i1} + z_{i2} \right) \log \left\{ p(y_i | x_i, z_i = 1, \beta) \text{pr}(z_i = 1 | x_i, \alpha) \right\} + p(y_i | x_i, z_i = 2, \beta) \text{pr}(z_i = 2 | x_i, \alpha) \right\} + \sum_{i=1}^{n} (1 - r_i) \left( z_{i3} + z_{i4} \right) \log \left\{ p(y_i | x_i, z_i = 3, \beta) \text{pr}(z_i = 3 | x_i, \alpha) \right\} + p(y_i | x_i, z_i = 4, \beta) \text{pr}(z_i = 4 | x_i, \alpha) \right\}.
\]

Finally, the maximum likelihood estimate of \( \gamma' = (\beta', \alpha') \) can be obtained by using either a Newton–Raphson algorithm (Fletcher, 1987) or the EM algorithm (Dempster et al., 1977; Ibrahim, 1990); both of these algorithms are iterative. First, we discuss the Newton–Raphson algorithm, and then we discuss the EM algorithm. The Newton–Raphson algorithm yields the maximum likelihood estimate \( \hat{\gamma} \) as the solution to

\[
\left( \frac{\partial l(\gamma)}{\partial \gamma} \right)_{\gamma = \hat{\gamma}} = 0.
\]

In particular, at the \((m + 1)\)th step of the Newton–Raphson algorithm,

\[
\hat{\gamma}^{(m+1)} = \hat{\gamma}^m + \left\{ - \left( \frac{\partial l(\gamma)}{\partial \gamma \partial \gamma'} \right)_{\gamma = \hat{\gamma}^m} \right\}^{-1} \left( \frac{\partial l(\gamma)}{\partial \gamma} \right)_{\gamma = \hat{\gamma}^m},
\]

where \( \hat{\gamma}^m \) is the estimate of \( \gamma \) at the \( m \)th step. The maximum likelihood estimate \( \hat{\gamma} \), obtained at convergence of this algorithm, has all the usual asymptotic properties: consistency and asymptotic normality, with variance consistently estimated by the inverse of the observed information matrix (negative inverse of the Hessian matrix),

\[
\text{var}(\hat{\gamma}) = \left\{ - \left( \frac{\partial l(\gamma)}{\partial \gamma \partial \gamma'} \right)_{\gamma = \hat{\gamma}} \right\}^{-1}.
\]

Now, we describe the EM algorithm using the method of weights that was described by Ibrahim (1990). If there were no missing data, to obtain \( \hat{\gamma}' = (\hat{\beta}', \hat{\alpha}') \), we maximize the
complete-data log-likelihood function

\[ l(\gamma) = \sum_{i=1}^{n} l(\gamma; z_i, x_i, y_i) \]

\[ = \sum_{i=1}^{n} \log \{ p(y_i|x_i, z_i, \beta) \} + \log \{ p(z_i|x_i, \alpha) \}, \tag{9} \]

where \( l(\gamma; x_i, y_i) \) is the contribution to the complete-data log-likelihood for the \( i \)th observation.

Using the EM algorithm (Dempster et al., 1977) with \( z_i \) coarsened for some subjects, the maximum likelihood estimate of \( \gamma \) can be obtained by maximizing the conditional expectation of the complete-data log-likelihood, which is denoted \( l^* (\gamma) \). Here, the expectation is taken with respect to the conditional distribution of \( z_i \) given the observed data, i.e.

\[ l^* (\gamma) = \sum_{i=1}^{n} l^*_i (\gamma) = \sum_{i=1}^{n} E_{z_i} \left\{ I^*(\gamma; z_i, x_i, y_i) \right\} \text{ observed data for subject } i \}. \tag{10} \]

In particular, if \( z_i \) is fully observed, then the observed data are \((y_i, x_i, z_i)\) and subject \( i \)'s contribution to the expected log-likelihood is

\[ l^*_i (\gamma) = E_{z_i} \left\{ I(\gamma; z_i, x_i, y_i) | z_i, x_i, y_i \right\} = l(\gamma; z_i, x_i, y_i), \tag{11} \]

which is the same as its contribution to the complete-data likelihood in equation (9). If \( z_i \) is coarsened, then the observed data are \((y_i, x_i, z^*_i)\) and

\[ l^*_i (\gamma) = E_{z_i} \left\{ I(\gamma; z_i, x_i, y_i) | x_i, y_i, z^*_i \right\} = E_{z_i|y_i, x_i, z^*_i} \left[ \log \{ p(y_i|x_i, z_i, \beta) \} \right] + E_{z_i|y_i, x_i, z^*_i} \left[ \log \{ p(z_i|x_i, \alpha) \} \right]. \tag{12} \]

Combining equations (11) and (12), we obtain

\[ l^* (\gamma) = \sum_{i=1}^{n} l^*_i (\gamma) \]

\[ = \sum_{i=1}^{n} (r_i \log \{ p(y_i|x_i, z_i, \beta) \} + (1 - r_i) E_{z_i|y_i, x_i, z^*_i} \left[ \log \{ p(y_i|x_i, z_i, \beta) \} \right]) \]

\[ + \sum_{i=1}^{n} (r_i \log \{ p(z_i|x_i, \alpha) \} + (1 - r_i) E_{z_i|y_i, x_i, z^*_i} \left[ \log \{ p(z_i|x_i, \alpha) \} \right]). \tag{13} \]

Since \( z_i \) is categorical with \( K \) levels, we can write the two conditional expectations in equation (13) as

\[ E_{z_i|y_i, x_i, z^*_i} \left[ \log \{ p(y_i|x_i, z_i, \beta) \} \right] = \sum_{z \in \Delta_i} p(Z_i = z|y_i, x_i, z^*_i) \log \{ p(y_i|x_i, z, \beta) \}, \]

and

\[ E_{z_i|y_i, x_i, z^*_i} \left[ \log \{ p(z_i|x_i, \alpha) \} \right] = \sum_{z \in \Delta_i} p(Z_i = z|y_i, x_i, z^*_i) \log \{ p(z_i|x_i, \alpha) \}. \]

The expected log-likelihood can thus be written as

\[ l^* (\gamma) = \sum_{i=1}^{n} [r_i \log \{ p(y_i|x_i, z_i, \beta) \} + (1 - r_i) \sum_{z \in \Delta_i} w_{iz} \log \{ p(y_i|x_i, z, \beta) \}] \]

\[ + \sum_{i=1}^{n} [r_i \log \{ p(z_i|x_i, \alpha) \} + (1 - r_i) \sum_{z \in \Delta_i} w_{iz} \log \{ p(z_i|x_i, \alpha) \}], \tag{14} \]
where

\[ w_{iz} = \text{pr}(Z_i = z|y_i, x_i, z_i^*) = \frac{p(y_i|x_i, z, \beta) p(z|x_i, \alpha)}{\sum_{z \in S_i} p(y_i|x_i, z, \beta) p(z|x_i, \alpha)}. \]  

(15)

For example, if \( z_i \) is coarsened as ‘white’, then \( S_i = \{1, 2\} \), and

\[ w_{iz} = \frac{p(y_i|x_i, z, \beta) p(z|x_i, \alpha)}{p(y_i|x_i, 1, \beta) p(1|x_i, \alpha) + p(y_i|x_i, 2, \beta) p(2|x_i, \alpha)}, \]

for \( z = 1, 2 \). We can think of the expected log-likelihood (14) as a weighted likelihood, with weights given by \( w_{iz} \):

Intuitively, each ‘filled-in’ value \( z \) for a coarsened \( z_i \) is weighted by the probability that \( Z_i = z \), given the observed data for subject \( i \). We can maximize the expected log-likelihood by using the EM algorithm. Suppose that we define the function

\[ I^*(\gamma|\gamma^{(t)}) = \sum_{i=1}^{n} [r_i \log \{ p(y_i|x_i, z_i, \beta) \} + (1 - r_i) \sum_{z \in S_i} w_{iz}^{(t)} \log \{ p(y_i|x_i, z, \beta) \}] \]

\[ + \sum_{i=1}^{n} [r_i \log \{ p(z_i|x_i, \alpha) \} + (1 - r_i) \sum_{z \in S_i} w_{iz}^{(t)} \log \{ p(z|x_i, \alpha) \}], \]  

(16)

where \( w_{iz}^{(t)} \) is the conditional probability in equation (15) evaluated at the given value \( \gamma = \gamma^{(t)} \).

The EM algorithm to maximize \( I^*(\gamma) \) entails the following.

(a) Obtain an initial estimate \( \gamma = \gamma^{(1)} \), say, by complete cases. At the \( t \)th step, we have \( \gamma^{(t)} \).

(b) Using \( \gamma^{(t)} \), compute \( w_{iz}^{(t)} = w_{iz}^{(t)}(\gamma^{(t)}) \).

(c) Fixing \( w_{iz}^{(t)} = w_{iz}^{(t)}(\gamma^{(t)}) \), maximize \( I^*(\gamma|\gamma^{(t)}) \) to obtain \( \gamma^{(t+1)} \).

(d) Iterate until convergence.

To obtain the asymptotic covariance matrix of \( \hat{\gamma} \), we can use the method of Louis (1982).

### 4. Example

Next we present an analysis of the data from the National Ambulatory Medical Care Survey. Recall that the outcome variable of interest, \( Y_i \), is the invasive level of tests ordered, with three possible response categories: 1, invasive; 2, non-invasive; 3, none. Letting

\[ p_{ij} = \text{pr}(Y_i = j|x_i, z_i), \]

the distribution of \( Y_i \), given the covariates, is multinomial with

\[ p(y_i|x_i, z_i) = p_{i1} I(y_i=1) p_{i2} I(y_i=2) p_{i3} I(y_i=3), \]  

(17)

where \( I(\cdot) \) is an indicator function. Because the response variable is ordinal (1, invasive; 2, non-invasive; 3, none), we choose to model the ‘cumulative’ probabilities with a logistic link function (McCullagh, 1980) relating the cumulative \( p_{ij} \)s to the covariates. In particular, we simultaneously modelled the two logits,

\[ \log \left( \frac{p_{i1}}{1 - p_{i1}} \right) = \log \left( \frac{F_{i1}}{1 - F_{i1}} \right) \]

and

\[ \log \left( \frac{p_{i1} + p_{i2}}{1 - (p_{i1} + p_{i2})} \right) = \log \left( \frac{F_{i2}}{1 - F_{i2}} \right), \]
where $F_{i2} = p_{i1} + p_{i2}$ is the probability of having an ‘invasive or non-invasive test’ done and $F_{i1} = p_{i1}$ is the probability of having an invasive test done. Note that $p_{i2} = F_{i2} - p_{i1}$, and $p_{i3} = 1 - F_{i2}$. We considered the following proportional odds model (McCullagh, 1980) for the cumulative logits:

$$
\log \left( \frac{F_{ik}}{1 - F_{ik}} \right) = \beta_{0k} + \beta_1 I\{\text{age}_i \in (21, 60)\} + \beta_2 I(\text{age}_i \geq 60) + \beta_3 I(\text{insurance}_i = \text{private}) \\
+ \beta_4 I(\text{insurance}_i = \text{med}) + \beta_5 I(\text{gender}_i = \text{male}) + \beta_6 I(\text{acute}_i = \text{yes}) \\
+ \beta_7 I\{\text{(race, eth)}_i = (\text{white, his})\} + \beta_8 I\{\text{(race, eth)}_i = (\text{AA, his})\} \\
+ \beta_9 I\{\text{(race, eth)}_i = (\text{AA, no-his})\},
$$

(18)

for $k = 1, 2$. Recall that the possibly coarsened covariate race–ethnicity $z_i$ has four levels: 1, white–Hispanic; 2, white–non-Hispanic; 3, African-American–Hispanic; 4, African-American–non-Hispanic. In this study there are 4140 (22%) participants with race–ethnicity coarsened; 3683 are white, and 457 are African-American.

For likelihood-based inferences about $\beta$, we must also specify the distribution of $z_i$ given $x_i$, since $z_i$ is coarsened for some individuals. For the four possible values of $z_i$, let

$$
\pi_{ik} = \Pr(Z_i = k|x_i, \alpha),
$$

$k = 1, 2, 3, 4$. The distribution of $z_i$ given $x_i$ is also multinomial,

$$
p(z_i|x_i) = \pi_{i1}^{I(z_i=1)} \pi_{i2}^{I(z_i=2)} \pi_{i3}^{I(z_i=3)} \pi_{i4}^{I(z_i=4)},
$$

(19)

but not ordinal, so we choose to model $(\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})$ by using multinomial logistic regression. In particular, we modelled the following three logits:

$$
\log(\pi_{i1}/\pi_{i4}) = \alpha_{01} + \alpha_1^t x_i, \\
\log(\pi_{i2}/\pi_{i4}) = \alpha_{02} + \alpha_2^t x_i, \\
\log(\pi_{i3}/\pi_{i4}) = \alpha_{03} + \alpha_3^t x_i,
$$

(20)

Alternatively, this model can be expressed in terms of the $\pi_{ij}$s as

$$
\pi_{i1} = \frac{\exp(\alpha_{01} + \alpha_1^t x_i)}{1 + \exp(\alpha_{01} + \alpha_1^t x_i) + \exp(\alpha_{02} + \alpha_2^t x_i) + \exp(\alpha_{03} + \alpha_3^t x_i)},
$$

(21)

$$
\pi_{i2} = \frac{\exp(\alpha_{02} + \alpha_2^t x_i)}{1 + \exp(\alpha_{01} + \alpha_1^t x_i) + \exp(\alpha_{02} + \alpha_2^t x_i) + \exp(\alpha_{03} + \alpha_3^t x_i)},
$$

(22)

$$
\pi_{i3} = \frac{\exp(\alpha_{03} + \alpha_3^t x_i)}{1 + \exp(\alpha_{01} + \alpha_1^t x_i) + \exp(\alpha_{02} + \alpha_2^t x_i) + \exp(\alpha_{03} + \alpha_3^t x_i)},
$$

(23)

and

$$
\pi_{i4} = \frac{1}{1 + \exp(\alpha_{01} + \alpha_1^t x_i) + \exp(\alpha_{02} + \alpha_2^t x_i) + \exp(\alpha_{03} + \alpha_3^t x_i)}.
$$

(24)

Finally, we consider the likelihood under coarsening at random. Let $p_{ij} = \Pr(Y_i = j|x_i, z_i) = p_{ij}(z_{i1}, z_{i2}, z_{i3}, z_{i4})$ denote that $p_{ij}$ is a function of $(z_{i1}, z_{i2}, z_{i3}, z_{i4})$, with $z_{ik} = I(Z_i = k)$ for $k = 1 \equiv$ white–Hispanic, $k = 2 \equiv$ white–non-Hispanic, $k = 3 \equiv$ African-American–Hispanic and $k = 4 \equiv$ African-American–non-Hispanic respectively.

For the 3683 patients with race–ethnicity coarsened as ‘white’, we know that $(z_{i3}, z_{i4}) = (0, 0)$, and that $(z_{i1}, z_{i2})$ must equal either $(1, 0)$ or $(0, 1)$. For the 457 patients with race–ethnicity coarsened as ‘African-American’, we know that $(z_{i1}, z_{i2}) = (0, 0)$, and that $(z_{i3}, z_{i4})$ must equal either
(1, 0) or (0, 1). Using equation (17) with probabilities $p_{ij|z_1, z_2, z_3, z_4}$, and equation (19) with probabilities given by equations (21)–(24), the log-likelihood in equation (7) can be expressed as

$$I(\gamma) = \sum_{i=1}^{n} r_i \log(p_{i|y_i=1}) + \sum_{i=1}^{n} (1 - r_i)(z_{i1} + z_{i2}) \log(p_{i|y_i=2}) + \sum_{i=1}^{n} (1 - r_i)(z_{i3} + z_{i4}) \log(p_{i|y_i=3})$$

$$+ \frac{p_{i|y_i=1}}{p_{i|y_i=2}} \frac{p_{i|y_i=2}}{p_{i|y_i=3}} \frac{p_{i|y_i=3}}{\pi_{i1}} \frac{\pi_{i2}}{\pi_{i3}} \frac{\pi_{i4}}{\pi_{i4}}$$

$$+ \frac{p_{i|y_i=1}}{p_{i|y_i=2}} \frac{p_{i|y_i=2}}{p_{i|y_i=3}} \frac{p_{i|y_i=3}}{\pi_{i1}} \frac{\pi_{i2}}{\pi_{i3}} \frac{\pi_{i4}}{\pi_{i4}}$$

(25)

To obtain the maximum likelihood estimate, the log-likelihood in equation (25) was maximized by using the numerical Newton–Raphson algorithm implemented in SAS procedure NLMIXED. This SAS procedure requires only the specification of the log-likelihood that is given in equation (25); this prevented errors that could be caused by analytically miscalculating the first and second derivatives. The convergence criterion which was adopted was that a relative parameter change in subsequent iterations did not exceed $10^{-6}$. In this example, the number of iterations that are required for convergence was 4, with the complete-case estimates used as starting values. The SAS NLMIXED procedure required approximately 40 s of central processor unit time on a Pentium 4 2.5 GHz, 512 Mbytes random-access memory computer.

In Table 2 we present estimates of the regression parameters $\beta$ that were obtained from the following three approaches:

(a) a complete-case analysis (denoted CC in Table 2);
(b) a coarsened data regression model (denoted COARSE in Table 2);
(c) maximum likelihood estimation assuming that race is coarsened at random (denoted ML in Table 2).

We note that the coarsened data regression model in Table 2 is formed by setting $\beta_1 = 0$ (the white–non-Hispanic effect equals the white–Hispanic effect, which is constrained to be 0) in equation (18) and $\beta_3 = \beta_4$ (the African-American–Hispanic effect equals the African-American–non-Hispanic effect) in equation (18). Overall, the results in Table 2 indicate that there are differences between the estimates of $\beta$ from these three different approaches. The biggest difference is the MED effect (Medicare plus Medicare versus no insurance). In particular, comparing the maximum likelihood with the complete-case analysis, the maximum likelihood estimate (0.114) of the MED effect is almost nine times larger than the complete-case estimate (0.013). Further, the maximum likelihood estimate is significant, whereas the complete-case estimate is not. Although both are significant, we see that the private insurance effect is almost 65% larger by using maximum likelihood (0.284) versus complete-case estimation (0.170). These discernible differences in estimates for maximum likelihood and the complete-case analysis suggest that the complete-case estimates are biased for the insurance effects. We see also that the African-American–Hispanic effect is about 30% larger by using complete-case (0.472) than by maximum likelihood estimation (0.359); the complete-case estimate is significant, whereas the maximum likelihood estimate is not. Further, although both are significant, the African-American–non-Hispanic effect is about 38% larger by using complete-case (0.140) than by maximum likelihood.
### Table 2. Estimates for the National Ambulatory Medical Care Survey

<table>
<thead>
<tr>
<th>Effect</th>
<th>Approach†</th>
<th>( \hat{\beta} )</th>
<th>Standard error</th>
<th>Z-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept 1 (( \beta_{01} ))</td>
<td>CC</td>
<td>-0.953</td>
<td>0.062</td>
<td>-15.31</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>-1.698</td>
<td>0.120</td>
<td>-14.15</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-1.123</td>
<td>0.055</td>
<td>-20.45</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intercept 2 (( \beta_{02} ))</td>
<td>CC</td>
<td>-0.326</td>
<td>0.062</td>
<td>-5.29</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>-1.053</td>
<td>0.118</td>
<td>-8.92</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-0.494</td>
<td>0.054</td>
<td>-9.08</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Private insurance</td>
<td>CC</td>
<td>0.170</td>
<td>0.050</td>
<td>3.42</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>0.692</td>
<td>0.097</td>
<td>7.13</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.284</td>
<td>0.044</td>
<td>6.44</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MED</td>
<td>CC</td>
<td>0.013</td>
<td>0.056</td>
<td>0.24</td>
<td>0.814</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>0.435</td>
<td>0.110</td>
<td>3.95</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.114</td>
<td>0.050</td>
<td>2.28</td>
<td>0.023</td>
</tr>
<tr>
<td>Age (21–59 years)</td>
<td>CC</td>
<td>0.657</td>
<td>0.043</td>
<td>15.24</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>0.897</td>
<td>0.082</td>
<td>10.94</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.716</td>
<td>0.038</td>
<td>19.00</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age (( \geq 61 ) years)</td>
<td>CC</td>
<td>0.815</td>
<td>0.049</td>
<td>16.53</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>0.930</td>
<td>0.095</td>
<td>9.79</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.840</td>
<td>0.044</td>
<td>19.19</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Male</td>
<td>CC</td>
<td>-0.099</td>
<td>0.032</td>
<td>-3.12</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>-0.099</td>
<td>0.061</td>
<td>-1.62</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-0.101</td>
<td>0.028</td>
<td>-3.56</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Acute</td>
<td>CC</td>
<td>-0.209</td>
<td>0.034</td>
<td>-6.22</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>-0.139</td>
<td>0.066</td>
<td>-2.11</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-0.189</td>
<td>0.030</td>
<td>-6.36</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(White,Hispanic)</td>
<td>CC</td>
<td>0.160</td>
<td>0.053</td>
<td>3.02</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.163</td>
<td>0.053</td>
<td>3.04</td>
<td>0.002</td>
</tr>
<tr>
<td>(African,Hispanic)</td>
<td>CC</td>
<td>0.472</td>
<td>0.219</td>
<td>2.15</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>-0.012</td>
<td>0.097</td>
<td>-0.12</td>
<td>0.900</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.359</td>
<td>0.219</td>
<td>1.62</td>
<td>0.101</td>
</tr>
<tr>
<td>(African,Non-Hispanic)</td>
<td>CC</td>
<td>0.140</td>
<td>0.055</td>
<td>2.54</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>COARSE</td>
<td>-0.012</td>
<td>0.097</td>
<td>-0.12</td>
<td>0.900</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.102</td>
<td>0.049</td>
<td>2.07</td>
<td>0.038</td>
</tr>
</tbody>
</table>

†CC, complete-case analysis; COARSE, coarsened data regression analysis; ML, maximum likelihood; MED, Medicaid or Medicare.

estimation (0.102). Aside from the issue of potential bias, note that there is some gain in efficiency by using maximum likelihood, except for the race–ethnicity effects, in which the estimated standard errors are almost identical for complete-case and maximum likelihood estimation. In particular, for all effects other than the race–ethnicity effects, the estimated relative efficiencies (the ratio of estimated variances) are between 76% and 79%.

We can also compare the maximum likelihood with the coarsened data regression estimates. In Table 2, there are also discernible differences between maximum likelihood and the coarsened data regression estimates of the insurance effects, the age effects and the race–ethnicity effects. Thus, there is some evidence to suggest that the underlying assumption of the coarsened data regression analysis is untenable and that the resulting estimates of the effect of the coarsened covariate, as well as other covariates, are biased.

Finally, the maximum likelihood estimates in Table 2 can be given the following subject-matter interpretation; the effect of each of the covariates on the odds of having a test performed (responding at level \( k \) or lower) makes intuitive sense, as we now describe. A patient with pri-
vate insurance has $\exp(0.284) \approx 1.33$ times the odds of having a test performed when compared with a patient without insurance, and a patient on Medicare or Medicaid has $\exp(0.114) \approx 1.12$ times the odds of having a test performed when compared with a patient without insurance. When compared with a patient who is less than 21 years old, a patient who is 21–59 years old has $\exp(0.716) \approx 2.05$ times the odds of having a test performed, and a patient who is 60 years old or older has $\exp(0.834) \approx 2.32$ times the odds of having a test performed. A female patient has $\exp(0.101) \approx 1.11$ times the odds of having a test performed when compared with a male patient. A patient who comes for an acute care visit has $\exp(-0.189) \approx 0.828$ times the odds of having a test performed when compared with a patient who comes for a general physical examination, mainly because acute care patients coming for out-patient visits have illnesses (upper respiratory tract infections, lower back pain, etc.) that are usually easily diagnosed by a physical examination without the need for tests. Compared with a white–non-Hispanic, a white–Hispanic has $\exp(0.163) \approx 1.18$ times the odds of having a test performed, an African-American–Hispanic has $\exp(0.359) \approx 1.43$ times the odds of having a test performed and an African-American–non-Hispanic has $\exp(0.102) \approx 1.11$ times the odds of having a test performed.

It must be emphasized that the maximum likelihood estimates of $\beta$ that are reported here depend on the correct choice of model for $z_i$ given $x_i$. To assess the sensitivity of the maximum likelihood estimates of $\beta$ to assumptions concerning $p(z_i|x_i)$, we considered additional multinomial logistic regressions for $z_i$ given $x_i$ in equation (20) which included pairwise interactions between the vector of covariates, $x_i$. The overall results of these additional analyses were remarkably similar to those which are reported in Table 2. Thus, at least for the data that were considered here, it appears that the estimates of $\beta$ and their standard errors are not very sensitive to the choice of model for $z_i$ given $x_i$. The SAS \textsc{proc nlmixed} commands that were used to obtain the estimates are given at \url{http://gsbwww.uchicago.edu/fac/michael.parzen/research/programs.html}.

5. Conclusion

In this paper we have considered generalized linear models with a coarsened covariate and proposed a likelihood-based method for estimating the regression parameters of interest. The method that we proposed is relatively flexible and can be applied to a much wider class of models for the outcome variable than was considered here, e.g. proportional hazards models for survival data and also to a broad class of non-linear models for the outcome. We also outlined Newton–Raphson and EM algorithms for obtaining maximum likelihood estimates of the regression parameters. An attractive feature of this procedure is that it can be easily implemented by using widely available statistical software (such as SAS \textsc{proc nlmixed} (SAS Institute, 2000)). In the motivating example that was presented here, the Newton–Raphson algorithm was found to converge rapidly, suggesting that this approach is computationally quite feasible for many practical applications.

Although we have not encountered examples in which more than one covariate is coarsened, the extension of this methodology to cases where two or more covariates are coarsened is relatively straightforward. For example, the methods that are described in Lipsitz and Ibrahim (1996) for missing covariate data can be extended to handle coarsened covariate data. We note that Heitjan and Basu (1996) considered missing data as a special type of coarsened data, and, in that sense, we have encountered data sets with more than one coarsened covariate, i.e. we have encountered a data set in which one covariate has some values coarsened, and another covariate has some values missing. In this paper, however, to present our proposed method clearly, our focus is on one coarsened covariate.
It is also possible to relax the assumption that the data are coarsened at random. For example, if the probability that an individual’s covariate is coarsened depends on the exact (but unobserved) value of the coarsened data, the coarsening mechanism is said to be ‘non-ignorable’. When there is non-ignorable coarsening of the data extra caution must be exercised. It is well known that non-ignorable missing data models can suffer from a lack of identifiability and model sensitivity and are generally computationally intractable (see, for example, Lipsitz and Ibrahim (1996)). We conjecture that the same is true also for non-ignorable coarsened data. Finally, we note that the method that is proposed in this paper has potential applications in meta-analyses where a discrete covariate of interest has different possible categorizations in each of the constituent studies. However, a direct application of the method that is proposed here would require an assumption about the similarity of the regression parameters across the studies.

Acknowledgements

We are grateful for the support that was provided by grants HL69800, AHRQ 10871, HL52329, HL61769, GM 29745, MH 54693 and CA 70101 from the National Institutes of Health in the USA.

References