Binomial Regression with Misclassification

Carlos Daniel Paulino,1∗ Paulo Soares,1 and John Neuhaus2

1Instituto Superior Técnico, e Centro de Matemática e Aplicações, Universidade Técnica de Lisboa, Portugal
2Department of Epidemiology and Biostatistics, University of California, San Francisco, California, U.S.A.
∗email: dpaulino@math.ist.utl.pt

Summary. Motivated by a study of human papillomavirus infection in women, we present a Bayesian binomial regression analysis in which the response is subject to an unconstrained misclassification process. Our iterative approach provides inferences for the parameters that describe the relationships of the covariates with the response and for the misclassification probabilities. Furthermore, our approach applies to any meaningful generalized linear model, making model selection possible. Finally, it is straightforward to extend it to multinomial settings.

Key words: Bayesian analysis; Binomial regression; Generalized linear model; Misclassification.

1. Introduction

Currently the analysis of binary response data is still dominated by the classical approach under which inferences are based on asymptotic theory. However, the Bayesian literature on this topic and related ones is steadily growing. See, for example, Dey, Ghosh, and Mallick (2000). Most importantly, the greatest contribution to this increasing ability to deal with more-complex problems and models is the continuous development of computational tools, in particular those based on Monte Carlo methods. Under the Bayesian paradigm it is then natural that the treatment of more problematic cases, such as misclassification, incomplete data, or measurement errors, would receive increasing attention. See, for instance, Geng and Asano (1989), Evans et al. (1996), Mendoza-Blanco, Tu, and Iyengar (1996), and Rekaya, Weigel, and Gianola (2001) for different approaches to misclassified categorical data under several sampling schemes. Soares and Paulino (2001) present an analysis of incomplete categorical data under informative censoring and Thürligen et al. (2000) offer a review of methods for measurement errors in the covariates.

Motivation for this work comes from data gathered in an ongoing study of human papillomavirus (HPV) infection at the University of California-San Francisco (UCSF) (see Moscicki et al. (2001)). The purpose of the investigation is to examine the association of several potential risk factors with HPV cervical infection among females who tested negative at entry into the study. The study screened 104 women aged 13 to 21 years who attended family planning clinics in the San Francisco Bay Area; it recorded for each woman her infection status at the end of the study (HPVS) by testing for HPV DNA in cervical samples, whether she had a history of vulvar warts (VW), whether she had any new sexual partner in the last two months at baseline (NSP), and whether she had an history of herpes simplex (HS). The median follow-up time was 26 months for those women who remained HPV-negative.

HPV is really a family of viruses responsible for various epithelial lesions of which over 90 subtypes have been described. From those, around 30 subtypes have a clear preference for the genital tissues and certain ones are commonly associated with cervical cancer. Since any test for HPV infection is limited to one subtype or a group of subtypes, it will miss a certain number of infections and, therefore, the response variable HPVS is bound to be affected by misclassification, producing some false negative results. Although less probable, false positive results are also possible due to sample contamination and other reasons associated with laboratory work. The definition of initial negative results was made using a conservative criterion: only those from the cohort who at baseline and first follow-up had negative results were included.

In this work, we present a fully Bayesian analysis of binomial regression data with a misclassified response and error-free covariates. In Section 2, the problem is described, along with an informative misclassification model. Section 3 focuses on the use of generalized linear models to analyze the association of the response with the covariates, embodying the ideas found in Bedrick, Christensen, and Johnson (1996) on prior specification. Section 4 discusses the use of data augmentation and other computational issues. In Section 5, the HPV data above mentioned serve as an illustration of the analysis developed and, finally, Section 6 contains some concluding remarks.
2. Model for the Misclassified

Binomial Regression Data

Consider regression data \((n_k, N_k, x_k), k = 1, \ldots, N\), where the \(n_k\)'s represent the number of successes from independent binomial distributions, Binomial\((n_k, \phi_k)\), the \(x_k\)'s are known \(p \times 1\) vectors of covariates, and the index \(k\) denotes covariate patterns. Due to the action of some corrupting mechanism, the response variable is often classified incorrectly. To accommodate this misclassified response, we conceptually split the data-collecting process into two stages: an unobserved sampling stage related to the true response \(R^T\) followed by a reporting stage where an observed and possibly wrong response \(R^O\) is reported.

If we associate a success with \(R^T = 1\) and write \(\theta_{ki} = P(R^T = i \mid x_k), k = 1, \ldots, N, i = 0, 1\), with \(\sum_i \theta_{ki} = 1\) and \(\lambda_{kij} = P(R^O = j \mid R^T = i, x_k), k = 1, \ldots, N, i, j = 0, 1\), with \(\sum_j \lambda_{kij} = 1\), then the probability of success for the observed response of an individual with covariates \(x_k\) is \(\phi_k = \sum_j \lambda_{kij} \theta_{ki}\). The probability model for the observed regression data, \(n = (n_k, N_k)\), is described by the nonidentifiable product-binomial likelihood

\[
L(\theta, \lambda \mid n) = \prod_{k=1}^{N_k} \left( \frac{N_k}{n_k} \right) \left( \sum \lambda_{kij} \theta_{ki} \right)^{n_k} \left( \sum \lambda_{kij} \theta_{ki} \right)^{N_k - n_k}, \tag{1}
\]

where \(\theta\) and \(\lambda\) stand for the sets of parameters \(\theta_{ki}\) and \(\lambda_{kij}\), respectively. Note that both kinds of parameters may depend on the covariates. When the covariate dependence of \(\lambda\) is allowed, this means that a differential misclassification mechanism is being adopted. A standard method of analyzing the association between a response variable and several covariates is using generalized linear models. In this case, we can express the expected value of the success proportions as a function of a linear predictor

\[
E\left(\frac{n_k}{N_k} \mid \theta_{ki} \right) = \theta_{ki}, \quad \theta_{ki} = \frac{\exp(x_k' \beta)}{1 + \exp(x_k' \beta)} = g(x_k' \beta),
\]

where \(\beta\) is an unknown \(p \times 1\) vector of regression coefficients and \(g(\cdot)\) can be an arbitrary c.d.f. Common choices for \(g\) include the logistic, normal, and Gumbel (for minima) distribution functions, defined by

\[
g(x' \beta) = \begin{cases} \exp(x' \beta)/(1 + \exp(x' \beta)), & x' \beta \leq 0, \\ \Phi(x' \beta), & 0 < x' \beta < \infty, \\ 1 - \exp(-\exp(x' \beta)), & x' \beta \geq \infty, \end{cases}
\tag{2}
\]

Henceforth, we will always assume that, a priori, the distinct quantities \(\theta\) and \(\lambda\) are independent, motivated by the occurrence of misclassification being inherent to the clinical procedure regardless of \(\theta\). Moreover, the inclusion of any form of dependence between these parameters would make the analysis more difficult, without any obvious practical benefit.

3. Subjective Prior Distribution

In a subjective Bayesian analysis, the introduction of further regression parameters leads to a potentially serious problem. Because these parameters do not relate directly to the data, they can look rather esoteric to the practitioner’s eyes. This is particularly true when competing choices of the link function, \(g^{-1}(\cdot)\), must be considered, each one with a different interpretation of \(\beta\). In practice, this can be a subtle, but nevertheless powerful inducement for using convenience priors and, sometimes, to skip altogether the expert prior elicitation step and stick to noninformative or diffuse prior distributions. A recent example of this approach can be seen in Rekaya et al. (2001); they use hierarchical priors of convenience in problems of this type, but with a simplified misclassification structure.

Bedrick et al. (1996) considered this problem of generalized linear models and proposed a method of overcoming it using a so-called conditional means prior (CMP). In brief, a CMP is obtained as follows:

1. Choose \(p\) covariate vectors \(\tilde{x}_l, l = 1, \ldots, p\);
2. Assign a prior on \(\{\theta_0(\tilde{x}_l' \beta), \ldots, \theta_0(\tilde{x}_p' \beta)\}\);
3. Obtain the induced prior on \(\beta\) by using the change-of-variables method.

The choice of the \(p\) covariate vectors \(\tilde{x}_l\) should also be subjected to expert opinion and be made so that those vectors are widely spaced in the predictor space in order to make it reasonable to assume prior independence of the quantities \(\theta_0(\tilde{x}_l' \beta)\). On this subject, we refer the reader to Bedrick et al. (1996) for general guidelines and diagnostic measures for the choice of the \(\tilde{x}_l\)'s, and to Bedrick, Christensen, and Johnson (1997) for an illustration.

In the following, we will abandon some of the previous generality and focus on a particular case that will be important in the applications. This is when one specifies that, independently, \(\theta_{0}(\tilde{x}_l' \beta) \sim \text{Beta}(c_l, d_l), l = 1, \ldots, p\), and that \(g(\cdot)\) is a continuous c.d.f. \(F(\cdot)\) with density \(f(\cdot)\). Denoting \(\theta_{0}(\tilde{x}_l' \beta) = 1 - \theta_0(\tilde{x}_l' \beta)\), the induced prior on \(\beta\) is

\[
\pi(\beta) \propto \prod_{l=1}^{p} \{\theta_0(\tilde{x}_l' \beta)]^{c_l} - 1 \theta_0(\tilde{x}_l' \beta)]^{d_l} f(\tilde{x}_l' \beta). \tag{3}
\]

To illustrate a simple logistic regression model, the Jacobian of the transformation \(\theta_0(\tilde{x}_l' \beta), \theta_0(\tilde{x}_0' \beta)) \rightarrow \beta \equiv (\tilde{\beta}_1, \tilde{\beta}_2)\), where \(\tilde{x}_l = (1, x_l)\), is \((x_2 - x_1)\prod_{l=1}^{p} \theta_0(\tilde{x}_l' \beta)/\theta_0(\tilde{x}_0' \beta)\) and hence (3) is \(\pi(\beta) = \chi_2 - x_1 \prod c_l \theta_0(\tilde{x}_l' \beta)]^{c_l} \theta_0(\tilde{x}_0' \beta)]^{d_l} f(\tilde{x}_0' \beta)\). The hyperparameters \(c_l\) and \(d_l\) are determined (indirectly in general) from expert prior judgments on features of the prior distribution for \(\{\theta_0(\tilde{x}_l' \beta)\}\). An analogous process is used to obtain the (possibly beta) prior hyperparameters for \(\lambda\).

Comparing this prior with the likelihood \(L(\beta, \lambda \mid n) = L(\theta(\beta), \lambda \mid n)\), we see that the presence of a misclassified response, along with the probabilistic model described in Section 2 to explain that misclassification, leads to a complicated posterior kernel that makes impossible any straightforward inferences using analytical methods. We will see in the next section how the use of data augmentation can alleviate this problem by separating the parameters \(\beta\) and \(\lambda\) in the likelihood.

4. Data Augmentation

Let \(m_{ki}\) be the number of observations with \(R^T = i\) and \(R^O = j\) among those observations with a covariate pattern \(x_k\). For these unobserved quantities, we have \(m_{ki} = \sum m_{kij} = n_k\) and \(m_{ki} = N_k - n_k\), while any other partial sum is also unobserved. The augmented data \(m = (m_{ki})\) is an hypothetical sample from a product of multinomial distributions.
$M\{N_k, (\lambda_{ki}\theta_{ki})\}$ with a corresponding likelihood under the
generalized linear model parameterization,

$$L(\beta, \lambda \mid m) \propto \prod_{k,i} \{\theta_i(x_i^k, \beta)\}^{n_{ki}} \prod_{k,i,j} \lambda_{kij}^{m_{kij}}.$$  

This likelihood shows that a data augmentation approach can
serve our inferential purposes, in the sense that it leads to a
factorization $L(\beta, \lambda \mid m) = L(\beta \mid m) \times L(\lambda \mid m)$ that, in a
sense, conjugates nicely with the induced prior for $\beta$ in expres-
sion (3). In fact, the augmented data posterior distribution is

$$\pi(\beta, \lambda \mid m) \propto \pi(\beta \mid m)\pi(\lambda) \prod_{k,i,j} \lambda_{kij}^{m_{kij}}, \quad (4)$$

where $\pi(\lambda)$ is a prior distribution for $\lambda$ and

$$\pi(\beta \mid m) \propto \prod_{i=1}^{p} \{\theta_i(\tilde{x}_i^k, \beta)\}^{c_i-1} \{\theta_i(\tilde{x}_i^k, \beta)\}^{d_i-1} f(\tilde{x}_i^k \mid \beta) \times \prod_{k,i} \{\theta_i(x_i^k, \beta)\}^{m_{kij}}. \quad (5)$$

For the reporting stage parameters $\lambda$, it may be generally
reasonable to assume prior independence among the sets
\{\lambda_{kij} \mid j = 0, 1\}, $\forall k, i$, and to use a beta distribution for each.
In this case, their posterior distribution given $m$ is a prod-
uct of beta distributions with their hyperparameters updated
by $m$.

Conditionally on the observed data, the augmented data is
distributed according to independent binomial distributions for
each $k$

$$m_{k01} \mid \beta, \lambda, n \sim \text{Binomial}\left\{n\beta, \lambda_{k01}\theta(x, \beta) + \sum_{i} \lambda_{ki0}\theta_i(x_i^k, \beta)\right\}$$

$$m_{k10} \mid \beta, \lambda, n \sim \text{Binomial}\left\{n - n\beta, \lambda_{k10}\theta(x, \beta) + \sum_{i} \lambda_{k01}\theta_i(x_i^k, \beta)\right\}. \quad (6)$$

From this setup, it seems now possible to draw inferences
based on a data augmentation algorithm (see Tanner, 1996),
such as the chained data augmentation algorithm (CDA).
This algorithm can be viewed as a Gibbs sampler and is
formed by the following steps:

1. Choose adequate initial values $\beta^0$ and $\lambda^0$;
2. For $i = 1, \ldots, t$:
   a. Imputation step
      i. sample $m_i$ from the independent binomial distrib-
         butions in (6) given $\beta^{i-1}, \lambda^{i-1},$ and $n$;
   b. Posterior step
      i. sample $\lambda^i$ from the independent beta distrib-
        utions given $m^i$;
      ii. sample $\beta^i$ from $\pi(\beta \mid m)$ in (5) given $m^i$.

Then, under general conditions, $\pi(\beta, \lambda \mid m^i)$ will converge to
$\pi(\beta, \lambda \mid m)$ as $i$ goes to infinity (Tanner and Wong, 1987).

The remaining problematic issue is to sample from $\pi(\beta \mid m)$
in step (2)(b)(ii). A possible solution is to resort to the
sampling-importance resampling (SIR) method as described
by Gelman et al. (1995). For the importance distribution,

we found it adequate to use a multivariate student $t$-density
with $\nu$ degrees of freedom, with a mode equal to the mode
of $\pi(\beta \mid m^i)$ and dispersion proportional to the asymptotic co-
variance matrix evaluated at the mode. The value of $\nu$ should
ensure that the tails of the importance distribution decay
slower than those of the target distribution $\pi(\beta \mid m^i)$.

5. The HPV Data
We will now analyze the problem described in Section 1. Writing $x' = (VW, NSP, HS)$, the collected data is pre-
sented in Table 1 and a brief descriptive analysis is given
in Figure 1, where we can see that, for example, 54 women
were diagnosed as infected by the HPV during the study,
while 50 others maintained a negative result in the clinical
tests.

Following Bedrick et al. (1997), we carried out the prior
elicitation with the kind collaboration of Dr Anna-Barbara
Moscicki from the Department of Pediatrics at UCSF. Since
we have three binary covariates, we needed to choose four
covariate configurations to induce the prior on $\beta$. So, Dr
Moscicki was asked to choose those four configurations and
to provide values for the 1%, 50%, and 99% quantiles of her
prior density for the probability of HPV infection of some-
one with those covariate characteristics. This is a purposeful
overspecification of the beta distributions, so we used the two
quantiles on which the expert was more confident, 50% and
99%, to calculate the prior hyperparameters, and the third
one to assess the consistency of the choice. This process was
iterated until a consistent set of values was found and then
the hyperparameters were obtained numerically.

<table>
<thead>
<tr>
<th>$x_k^*$</th>
<th>$n_k$</th>
<th>$N_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0)</td>
<td>12</td>
<td>44</td>
</tr>
<tr>
<td>(0,0,1)</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>(0,1,0)</td>
<td>29</td>
<td>40</td>
</tr>
<tr>
<td>(0,1,1)</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>(1,0,0)</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>(1,0,1)</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>(1,1,0)</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 1. Brief descriptive analysis of the HPV data.
Table 2

| Expert elicitation of prior quantiles for $\theta$ and $\lambda$ |
|-----------------|-------------|-------------|-------------|-----------------|
| $x_1 = (1, 1, 1)$ | 0.30 0.70 0.95 | 5.944 2.731 |
| $x_2 = (1, 0, 0)$ | 0.09 0.30 0.60 | 4.692 10.512 |
| $x_3 = (0, 1, 0)$ | 0.15 0.40 0.70 | 5.897 8.682 |
| $x_4 = (0, 0, 1)$ | 0.14 0.30 0.50 | 10.142 23.224 |
| $P(\text{false positive})$ | 0.02 0.05 0.10 | 8.386 153.386 |
| $P(\text{false negative})$ | 0.03 0.07 0.20 | 3.135 37.650 |

Regarding the misclassification probabilities, it is quite reasonable to argue that the presence of misclassification is not related to the covariates and, consequently, a nondifferential misclassification mechanism can be used. This means that we only have to deal with two parameters: $P(\text{false positive}) = 1$-specificity $= \lambda_{01}$ and $P(\text{false negative}) = 1$-sensitivity $= \lambda_{10}$. The elicitation of the prior hyperparameters for these two misclassification probabilities followed the lines described above. Assessments of the sensitivity and specificity of the HPV diagnostic using known samples (Moscicki et al., 2001) formed the basis of Dr Moscicki’s prior opinion about misclassification rates. The complete set of elicited quantiles is given in Table 2.

In this article, we will restrict the analysis to three common generalized linear models, the logistic ($M_1$), probit ($M_2$), and complementary log-log models ($M_3$), for which the c.d.f. is defined by the function $g$ in (2) for which $x'\beta = \beta_0 + \beta_1VW + \beta_2\text{NSP} + \beta_3\text{HS}$.

A preliminary simulation study of the importance-sampling weights showed that the degrees of freedom of the importance distribution, $\nu$, need not be very small and that a value of 30 would be fine for all three models in terms of the usual requirement of a small variability for the importance weights. We believe that this is due to the light tails of the posterior of $\beta$, since we found that the posterior results were only mildly affected by the variation of the importance-sampling degrees of freedom.

To select the most appropriate model for the data, we computed Bayes factors to compare $M_i$ and $M_j$, $BF_{ij} = P(n | M_i) / P(n | M_j)$, where $P(n | M_i)$ is the marginal probability of observing the data $n$ from model $M_i$, that is,

$$P(n | M_i) = \int \mathcal{L}_i(\beta, \lambda | n) \pi(\beta, \lambda) d\beta d\lambda,$$

where $\mathcal{L}_i(\beta, \lambda | n)$ is the likelihood for the model $M_i$. That marginal probability can be estimated by sampling from the prior distribution $\pi(\beta, \lambda)$ of Sections 3 and 4, and then evaluating

$$P(n | M_i) \approx \frac{1}{t} \sum_{l=1}^{t} \mathcal{L}_i(\beta^l, \lambda^l | n).$$

To sample from the marginal prior of $\beta$, it is sufficient to sample from the prior distributions of $\theta_l(\tilde{x} | \beta)$, $l = 1, \ldots, 4$, and then solve the set of equations $\{\theta_1 = g(\tilde{x} | \beta)\}$ to obtain samples from $\beta$.

To obtain sufficiently stable estimates, it was necessary to use $t = 10^5$, from which the values obtained were $BF_{12} = 1.0058$, $BF_{23} = 1.0015$, and $BF_{13} = 1.0073$. These values are all close to 1 so if the prior odds for any pair of models are equal to 1, then they are practically unaltered by the data. Therefore, this criterion for model selection does not provide enough evidence for choosing any particular one. This being so, the following results were computed from a sample of 100,000 points obtained with the logistic model, after a proper convergence analysis using the methods available in the CODA (Best, Cowles, and Vines, 1997) or BOA (Smith, 2001) software.

Figures 2 and 3 display the plots of the prior (dashed lines) and posterior densities (solid lines) for $\lambda_{01}$, $\lambda_{10}$, and $\theta_l(\tilde{x} | \beta)$, $l = 1, \ldots, 4$, where the posteriors are drawn as smoothed histograms. These figures show that there is no conflict between the prior information and the data, and that the expert’s opinion about the misclassification probabilities was very precise; this is particularly true in the false positive case, in the sense that there are no significant differences between the respective prior and posterior densities. In this regard, we should add that a further simulation study (based on uniform priors for $\lambda_{01}$ and $\lambda_{10}$) showed that the data can strongly affect the estimation of the probabilities of a false positive and a false negative (and of $\theta$ as well). Furthermore, point and interval summaries of the posterior distribution of the misclassification probabilities were found to be rather insensitive to changes in $\theta$’s prior distribution strength.

Other posterior estimates are shown in Table 3. The highest priority density (HPD) credible intervals were obtained with Chen and Shao’s method (described in Chen, Shao, and Ibrahim (2000)). Those estimates show that NSP is the covariate with the strongest association with HPV infection.

Figure 2. Prior and posterior distributions for $\lambda$ (the dashed lines are for the priors and the solid ones for the posteriors).
This result supports the hypothesis that HPV is a sexually transmitted virus and that most new infections are due to exposure rather than to the development of latent infections.

All the computations were made with specially written OX or Mathematica code that is available from the authors upon request.

6. Concluding Remarks

The presented approach extends the work of Bedrick et al. (1996) to accommodate the observation of binomial regression data with misclassification on the response. In Section 5, the described illustration misses an important point—there are no continuous covariates. However, by construction, this approach does not depend on the covariate nature and so it is immediately applicable to the general case. Naturally, the larger the number of covariates, the more extensive the eliciting work will be and the more complex it is to assure the prior independence between the $\theta_i(\tilde{x}_i/\beta)$’s.

Despite the undeniable advantages of the Bayesian operation on drawing inferences in settings like this, model nonidentifiability requires special care in the prior elicitation process because, as is well known, the posterior inferences of nonidentifiable parameters are strongly influenced by the prior, even for increasingly large sample sizes. Another simple extension would be the consideration of polychotomous response data.

On the negative side, the use of the SIR method within the chained data augmentation algorithm is currently a weakness of this approach. Since there is no way to completely ensure the “quality” of the SIR results, alternatives, such as advanced importance sampling, slice sampling, or even perfect sampling (see Liu (2001)), are badly needed. Ongoing preliminary work has already allowed us to validate the SIR results using a much faster slice sampling algorithm.

Table 3

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>MC error</th>
<th>S.d.</th>
<th>Median</th>
<th>HPD CI (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$(false positive)</td>
<td>0.0568</td>
<td>0.0001</td>
<td>0.0188</td>
<td>0.0549</td>
<td>0.0228</td>
</tr>
<tr>
<td>$P$(false negative)</td>
<td>0.0586</td>
<td>0.0001</td>
<td>0.0311</td>
<td>0.0535</td>
<td>0.0082</td>
</tr>
<tr>
<td>Intercept</td>
<td>−1.0215</td>
<td>0.0013</td>
<td>0.3321</td>
<td>−1.0108</td>
<td>−1.6715</td>
</tr>
<tr>
<td>VW</td>
<td>0.3861</td>
<td>0.0016</td>
<td>0.4579</td>
<td>0.3824</td>
<td>−0.5071</td>
</tr>
<tr>
<td>NSP</td>
<td>1.5506</td>
<td>0.0014</td>
<td>0.3933</td>
<td>1.5452</td>
<td>0.7732</td>
</tr>
<tr>
<td>HS</td>
<td>0.3028</td>
<td>0.0014</td>
<td>0.4230</td>
<td>0.3017</td>
<td>−0.5249</td>
</tr>
</tbody>
</table>
Binomial Regression with Misclassification

ACKNOWLEDGEMENTS

The authors thank Dr. Anna-Barbara Moscicki from the University of California-San Francisco for the use of the data and for her kind collaboration in the prior elicitation. They also gratefully acknowledge the helpful comments and questions raised by the associate editor and the referees. The first author’s research was partially supported by the Luso-American Foundation for Development (FLAD) and the Centre for Mathematics and Its Applications (CEMAT). Grants from the U.S. National Institutes of Health and the Portuguese Studies Program at the University of California-Berkeley also supported this research.

RÉSUMÉ

Motivés par l’étude de l’infection à virus papilloma humain chez la femme, nous présentons une analyse de régression binomiale bayésienne dans laquelle la réponse est sujette au processus d’erreur de classement sans contrainte. Notre approche itérative fournit des inférences pour les paramètres qui décrivent la relation des covariables avec la réponse, et pour les probabilités d’erreur de classement. De plus, notre approche s’applique à n’importe quel modèle linéaire généralisé pertinent, rendant possible la sélection de modèle, et il est immédiat de l’étendre à un contexte multinomial.

REFERENCES


Received March 2002. Revised February 2003. Accepted February 2003.