University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2006 *Paper* 68

A note on bias due to fitting prospective multivariate generalized linear models to categorical outcomes ignoring retrospective sampling schemes

Bhramar Mukherjee* Ivy Liu†

*University of Michigan, bhramar@umich.edu

†Victoria University of Wellington, i-ming.liu@vuw.ac.nz

A note on bias due to fitting prospective multivariate generalized linear models to categorical outcomes ignoring retrospective sampling schemes

Bhramar Mukherjee and Ivy Liu

Abstract

Outcome dependent sampling designs are commonly used in economics, market research and epidemiological studies. Case-control sampling design is a classic example of outcome dependent sampling, where exposure information is collected on subjects conditional on their disease status. In many situations, the outcome under consideration may have multiple categories instead of a simple dichotomization. For example, in a case-control study, there may be disease sub-classification among the "cases" based on progression of the disease, or in terms of other histological and morphological characteristics of the disease. In this note, we investigate the issue of fitting prospective multivariate generalized linear models to such multiple-category outcome data, ignoring the retrospective nature of the sampling design. We first provide a set of necessary and sufficient conditions for the link functions that will allow for equivalence of prospective and retrospective inference for the parameters of interest. We show that for categorical outcomes, prospective-retrospective equivalence does not hold beyond the generalized multinomial logit link. We then derive an approximate expression for the bias incurred when link functions outside this class are used. We illustrate the extent of bias through a real data example, based on the ongoing Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial by the National Cancer Institute.

A note on bias due to fitting prospective multivariate generalized linear models to categorical outcomes ignoring retrospective sampling schemes

BHRAMAR MUKHERJEE[1] AND IVY LIU[2]
[1]Department of Biostatistics, University of Michigan
Ann Arbor, Michigan, USA.
email: bhramar@umich.edu

[2]Department of Mathematics, Statistics and Computer Science
Victoria University of Wellington
Wellington, New Zealand.
email: i-ming.liu@vuw.ac.nz

SUMMARY

Outcome dependent sampling designs are commonly used in economics, market research and epidemiological studies. Case-control sampling design is a classic example of outcome dependent sampling, where exposure information is collected on subjects conditional on their disease status. In many situations, the outcome under consideration may have multiple categories instead of a simple dichotomization. For example, in a case-control study, there may be disease sub-classification among the "cases" based on progression of the disease, or in terms of other histological and morphological characteristics of the disease. In this note, we investigate the issue of fitting prospective *multivariate* generalized linear models to such multiple-category outcome data, ignoring the retrospective nature of the sampling design. We first provide a set of necessary and sufficient conditions for the link functions that will allow for equivalence of prospective and retrospective inference for the parameters of interest. We show that for categorical outcomes, prospective-retrospective equivalence does not hold beyond the generalized multinomial logit link. We then derive an approximate expression for the bias incurred when link functions outside this class are used. We illustrate the extent of bias through a real data example, based on the ongoing Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial by the National Cancer Institute.

KEYWORDS: Choice-based sampling, Colorectal adenoma, Cumulative logit, Link function, Model mis-specification, Ordered response.

1 Introduction

Case-control study is a prime example of outcome dependent sampling where individuals are sampled conditional on their disease status, and exposure information is then collected on the sampled individuals. Several other forms of outcome dependent sampling are commonly observed in econometric and social research, where explanatory variables are related to the discrete choices already made by individuals (Manski and McFadden, 1981). For binary outcomes, it is well-known that the disease-exposure (response-explanatory variable) association can be consistently estimated using a prospective logistic model (Andersen, 1970; Prentice and Pyke, 1979) under outcome dependent sampling. The prospective-retrospective equivalence does not hold for any other generalized linear model (GLM) for binary data, beyond the logistic link function (Kagan, 2001). Ignoring the outcome-dependent nature of sampling and fitting any arbitrary link function (such as probit, complimentary log-log) could produce biased estimates of the regression parameters of interest, and the bias could be substantial depending on the sampling rates from the two response categories (Neuhaus, 2002).

In modern medicine, with precise characterization of diseases in histological and morphological terms, it is natural to consider disease states with more than one category, i.e., there may be subdivisions within the "cases". For example, patients diagnosed with cancer may have cancer of stage-I, stage-II or stage-III at the time of the diagnosis or may simply be classified in terms of the number/size of adenomas/tumors present. There are several popular models for analyzing categorical response (Agresti, 2002), for instance, the cumulative logit model for ordered outcomes, that one may want to fit in such scenarios. It may also be desirable to select a fixed number of subjects from each disease category through an outcome dependent sampling scheme. The purpose of this note is to establish an approximation to the bias when multivariate generalized linear models (which includes many common models for outcomes with multiple categories) are fitted to data collected by retrospective sampling. An additional objective is to illustrate the degree and extent of bias through a real example based on the PLCO cancer screening trial (based on data available

in Ji et al, 2006;). In our example, we consider disease outcomes that are classified according to number of colorectal adenomas detected in a subject by sigmoidoscopy screening of the distal colon (descending colon and sigmoid or rectum). We investigate the association between smoking (never vs. ever) and number of adenomas and illustrate the extent of bias that may result with a naive prospective analysis of data sampled retrospectively from the PLCO cohort. This dataset is also used to assess the accuracy of our analytical approximation to the bias.

We would like to emphasize that there exists a rich literature on appropriate estimation techniques for fitting prospective models under outcome-dependent or choice-based sampling schemes. We refer the reader to the pioneering work by Scott and Wild (1986) and Breslow and Cain (1988). Their work spurred further research in this area (Breslow and Holubkov, 1997a, 1997b; Breslow and Chatterjee 1999; Chatterjee 2004; Scott and Wild 1991, 1997; Wild 1991; Wang et al 1997). Pfeffermann *et al* (1998) and Pfeffermann and Sverchov (1999) also considered outcome dependent-sampling in the context of sample surveys. The purpose of this note is not to develop new inferential procedures, but to provide an analytical description of the bias for the situation with multiple outcome categories, and to leave the reader with an intuitive sense of the bias mechanism via our real data example.

The rest of the article is organized as follows. In Section 2, we introduce the model, notations, and provide a characterization of the link functions in a multivariate generalized linear model for categorical outcomes (MVGLM) which allow prospective-retrospective equivalence of likelihood inference regarding the regression parameters of interest. In Section 3, we provide an approximation to the bias when a prospective MVGLM is fitted to retrospective data, completely ignoring the sampling design. In Section 4, we illustrate the magnitude of the bias and the quality of our approximation through a real data example. Section 5 presents concluding remarks.

2 Model and Notations

2.1 Multivariate Generalized Linear Models

Let Y_i be a K-category outcome variable scaled from $1, \ldots, K$, and let $\boldsymbol{x}_i$ denote the $s \times 1$ vector of covariates, both measured for subject i, $i = 1, \ldots, n$. Let us define a set of $q = K - 1$ indicator variables $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq})'$, where $y_{im} = 1$ if subject i belongs to response class m and 0 otherwise, $m = 1, \cdots, q$.

Following the notational convention of Fahrmeir and Tutz (2001), we express the multinomial distribution for a general categorical variable Y_i, in terms of the vector $\boldsymbol{y}_i$ as

$$f(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_i, \phi, w_i) = \exp\left[\frac{\boldsymbol{y}_i'\boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{\phi} w_i + c(\boldsymbol{y}_i, \phi, w_i)\right],$$

where

$$\begin{aligned}
\boldsymbol{\theta}_i' &= \left[\log\left(\frac{\pi_{i1}}{1 - \sum_{j=1}^{q} \pi_{ij}}\right), \ldots, \log\left(\frac{\pi_{iq}}{1 - \sum_{j=1}^{q} \pi_{ij}}\right)\right] \\
b(\boldsymbol{\theta}_i) &= -\log\left(1 - \sum_{j=1}^{q} \pi_{ij}\right) \\
c(\boldsymbol{y}_i, \phi, w_i) &= -\log\left(y_{i1}! \cdots y_{iq}!(1 - \sum_{j=1}^{q} y_{ij})!\right).
\end{aligned}$$

Here $\pi_{im} = P(y_{im} = 1) = P(Y_i = m)$. Typically, π_{im} is modeled as a function of the covariates $\boldsymbol{x}_i$ for all $m = 1, \ldots, q$. In that case, we can express the model as

$$\boldsymbol{\pi}(\boldsymbol{x}_i) = \boldsymbol{h}(\boldsymbol{Z}_i\boldsymbol{\beta}) \tag{1}$$

where $\boldsymbol{\pi}(\boldsymbol{x}_i) = (\pi_{i1}(\boldsymbol{x}_i), \ldots, \pi_{iq}(\boldsymbol{x}_i))'$; $\boldsymbol{Z}_i$ is a $q \times p$ design matrix involving $\boldsymbol{x}_i$; $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters; and $\boldsymbol{h} = (h_1, \ldots, h_q)'$ is a vector valued function operator

$$\boldsymbol{h} : S \subset \boldsymbol{R}^q \quad \rightarrow \quad M \subset \boldsymbol{R}^q$$

where M is the q dimensional simplex representing the admissible set of probabilities $M = \left\{(\eta_1, \ldots, \eta_q) \,|\, 0 < \eta_j < 1, \sum_{j=1}^{q} \eta_j < 1\right\}$.

Let us now consider the class of MVGLMs for categorical data with the design matrix $\boldsymbol{Z}_i$ of the following particular structure,

$$\boldsymbol{Z}_i = \begin{bmatrix} 1 & \boldsymbol{x}_i' & 0 & \mathbf{0}' & \cdots & 0 & \mathbf{0}' \\ 0 & \mathbf{0}' & 1 & \boldsymbol{x}_i' & \cdots & 0 & \mathbf{0}' \\ \vdots & & & & & & \\ 0 & \mathbf{0}' & 0 & \mathbf{0}' & \cdots & 1 & \boldsymbol{x}_i' \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_{01} \\ \boldsymbol{\beta}_1 \\ \beta_{02} \\ \boldsymbol{\beta}_2 \\ \vdots \\ \beta_{0q} \\ \boldsymbol{\beta}_q \end{bmatrix}$$

In this model, the total number of parameters is given by $p = (s+1)q$. The model in (1) can also be expressed as

$$\pi_{im}(\boldsymbol{x}_i) = P(Y_i = m \,|\, \boldsymbol{x}_i) = h_m(\beta_{01} + \boldsymbol{x}_i'\boldsymbol{\beta}_1, \ldots, \beta_{0q} + \boldsymbol{x}_i'\boldsymbol{\beta}_q),\ m = 1, \ldots, q,$$

where $\boldsymbol{h} = \{h_1, \ldots, h_q\}'$ is the multidimensional response function and $h_m : \boldsymbol{R}^q \rightarrow \boldsymbol{R}$ is the response function corresponding to the mth component (or category) of Y for all $m = 1, \ldots, q$. We assume that for all $m = 1, \cdots, q$, h_m is differentiable with respect to each co-ordinate.

2.2 Likelihood under outcome-dependent sampling scheme

Let us assume that the sampling probabilities for each individual in the population depend only on the outcomes and let λ_m denote the sampling rate at which subjects from response category $Y = m$ is sampled, $m = 1, \ldots, K$. Let n_m be the number of subjects selected from outcome category m and let N_m be the total number of subjects available in category m for the population under study. Then $\lambda_m = n_m/N_m$. Typically, the sampling rates are unknown, as N_ms are unknown except for some special cases. Let S_i be an indicator variable denoting whether subject i is selected or not from the population. Instead of the assumption of sampling without replacement, we will assume that the sampling model is iid Bernoulli sampling where each member from category $Y = m$ is selected by the result of a coin toss with equal selection probability λ_m. Therefore,

$$P(S_i = 1 \,|\, Y_i = m, \boldsymbol{x}_i) = \lambda_m.$$

By Bayes theorem, we have

$$
\begin{aligned}
P(Y_i = m \,|\, \boldsymbol{x}_i, S_i = 1) &= \frac{P(S_i = 1 \,|\, Y_i = m, \boldsymbol{x}_i) P(Y_i = m \,|\, \boldsymbol{x}_i)}{P(S_i = 1 \,|\, \boldsymbol{x}_i)} \\
&= \frac{\lambda_m h_m(\beta_{01} + \boldsymbol{x}_i'\boldsymbol{\beta}_1, \ldots, \beta_{0q} + \boldsymbol{x}_i'\boldsymbol{\beta}_q)}{\sum_{j=1}^{q} \lambda_j h_j(\beta_{01} + \boldsymbol{x}_i'\boldsymbol{\beta}_1, \ldots, \beta_{0q} + \boldsymbol{x}_i'\boldsymbol{\beta}_q) + \lambda_{q+1}(1 - \sum_{j=1}^{q} h_j(\beta_{01} + \boldsymbol{x}_i'\boldsymbol{\beta}_1, \ldots, \beta_{0q} + \boldsymbol{x}_i'\boldsymbol{\beta}_q))} \\
&= \frac{\lambda_m h_m(u_{i1}, \ldots, u_{iq})}{\sum_{j=1}^{q} \lambda_j h_j(u_{i1}, \ldots, u_{iq}) + \lambda_{q+1}(1 - \sum_{j=1}^{q} h_j(u_{i1}, \ldots, u_{iq}))}, \qquad (2)
\end{aligned}
$$

where $u_{im} = \beta_{0m} + \boldsymbol{x}_i'\boldsymbol{\beta}_m$, $m = 1, \cdots, q$. Without loss of generality, let the response category $K = q + 1$ denote the baseline category. The retrospective likelihood based on the above sampling scheme is

$$
\begin{aligned}
& L_R(\beta_{01}, \ldots, \beta_{0q}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q | \boldsymbol{x}_i, y_i, i = 1, \ldots, n) \\
\propto & \prod_{i=1}^{n} \left[\prod_{m=1}^{q} \left\{ \frac{\lambda_m h_m(u_{i1}, \ldots, u_{iq})}{\sum_{j=1}^{q} \lambda_j h_j(u_{i1}, \ldots, u_{iq}) + \lambda_{q+1} \left(1 - \sum_{j=1}^{q} h_j(u_{i1}, \ldots, u_{iq})\right)} \right\}^{y_{im}} \right. \\
& \left. \times \left\{ \frac{\lambda_{q+1} \left(1 - \sum_{j=1}^{q} h_j(u_{i1}, \ldots, u_{iq})\right)}{\sum_{j=1}^{q} \lambda_j h_j(u_{i1}, \ldots, u_{iq}) + \lambda_{q+1} \left(1 - \sum_{j=1}^{q} h_j(u_{i1}, \ldots, u_{iq})\right)} \right\}^{1 - \sum_{j=1}^{q} y_{ij}} \right]
\end{aligned}
$$

However, the prospective likelihood assuming that the data was obtained through a cohort study is given by

$$
\begin{aligned}
& L_P(\beta_{01}, \ldots, \beta_{0q}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q | \boldsymbol{x}_i, y_i, i = 1, \ldots, n) \\
\propto & \prod_{i=1}^{n} \left[\prod_{m=1}^{q} \{h_m(u_{i1}, \ldots, u_{iq})\}^{y_{im}} \left(1 - \sum_{j=1}^{q} h_j(u_{i1}, \ldots, u_{iq})\right)^{1 - \sum_{j=1}^{q} y_{ij}} \right]
\end{aligned}
$$

We now establish the following theorem which provides necessary and sufficient conditions for the response functions which will allow the effect of sampling rates in L_R to be absorbed in the intercept parameters β_{0m}, $m = 1, \cdots, q$, and thus allow L_R to differ from L_P by intercept terms only. Consequently, only for such link functions, the regression parameters $\boldsymbol{\beta}_m$, $m = 1, \cdots, q$ remain identifiable via the prospective likelihood.

Theorem 1: Suppose that $h_1, \cdots, h_q$ are real valued functions and for $m = 1, \cdots, q$, $\theta_m(\boldsymbol{\lambda})$ is a

real valued function of the sampling ratios, with $\boldsymbol{\lambda} = (\log(\lambda_1/\lambda_{q+1}), \cdots, \log(\lambda_q/\lambda_{q+1}))'$. Then,

$$\begin{aligned}
&\prod_{i=1}^{n}\left[\prod_{m=1}^{q}\left\{\frac{\lambda_m h_m(u_{i1},\ldots,u_{iq})}{\sum_{j=1}^{q}\lambda_j h_j(u_{i1},\ldots,u_{iq})+\lambda_{q+1}\left(1-\sum_{j=1}^{q}h_j(u_{i1},\ldots,u_{iq})\right)}\right\}^{y_{im}}\right.\\
&\left.\left\{\frac{\lambda_{q+1}\left(1-\sum_{j=1}^{q}h_j(u_{i1},\ldots,u_{iq})\right)}{\sum_{j=1}^{q}\lambda_j h_j(u_{i1},\ldots,u_{iq})+\lambda_{q+1}\left(1-\sum_{j=1}^{q}h_j(u_{i1},\ldots,u_{iq})\right)}\right\}^{1-\sum_{j=1}^{q}y_{ij}}\right]\\
=\ &\prod_{i=1}^{n}\left[\prod_{m=1}^{q}\{h_m(u_{i1}+\theta_1(\boldsymbol{\lambda}),u_{i2}+\theta_2(\boldsymbol{\lambda}),\ldots,u_{iq}+\theta_q(\boldsymbol{\lambda}))\}^{y_{im}}\right.\\
&\left.\left(1-\sum_{j=1}^{q}h_j(u_{i1}+\theta_1(\boldsymbol{\lambda}),u_{i2}+\theta_2(\boldsymbol{\lambda}),\ldots,u_{iq}+\theta_q(\boldsymbol{\lambda}))\right)^{1-\sum_{j=1}^{q}y_{ij}}\right] \qquad (3)
\end{aligned}$$

iff

$$h_m(u_1,\ldots,u_q)=\frac{\exp(d_m+\sum_{j=1}^{q}c_{mj}u_j)}{1+\sum_{l=1}^{q}\exp(d_l+\sum_{j=1}^{q}c_{lj}u_j)} \qquad (4)$$

and

$$\log\left(\frac{\lambda_m}{\lambda_{q+1}}\right)=\log\left(\frac{\lambda_m}{\lambda_K}\right)=\sum_{j=1}^{q}c_{mj}\theta_j(\boldsymbol{\lambda}).$$

for some set of scalars $\{d_m, c_{mj}, m = 1,\ldots,q, j = 1,\ldots,q\}$. The theorem holds under the assumption that the map $f : \boldsymbol{\lambda} = (\log(\lambda_1/\lambda_{q+1}), \cdots, \log(\lambda_q/\lambda_{q+1}))' \rightarrow \boldsymbol{\theta}(\boldsymbol{\lambda}) = (\theta_1(\boldsymbol{\lambda}), \cdots, \theta_q(\boldsymbol{\lambda}))'$ is one to one and onto, that is, if we know one vector we can retrieve the other.

Proof: The proof of this theorem resemble the argument in Kagan (2001) where an analogous characterization for the logistic link function is presented for all GLMs for binary data. The mathematical argument has to be modified for MVGLMs for outcomes with multiple categories and a rigorous complete proof is contained in the Appendix (A.1). Examples of commonly used link functions which satisfy the above characterization are the multinomial and adjacent category logit links, or any other generalized logit link functions (McCullagh and Nelder, 1999).

3 Magnitude of bias by ignoring the sampling scheme

From Theorem 1, we know that by using L_P in MVGLM model with link functions beyond the multiplicative intercept and odds structure, one is not able to estimate the true model parameters

by a naive prospective analysis. We now present an approximation to the bias incurred by fitting a prospective MVGLM to these categorical observations. We treat the problem of ignoring the sampling design as a model mis-specification problem (Neuhaus, 1999, 2002) and use classical results from (Huber, 1967; White, 1982) to derive properties of MLEs under the mis-specified model ignoring the sampling design.

From (2), we know that the true model which acknowledges the retrospective sampling scheme is given by

$$\begin{aligned}\pi_m^{\text{T}}(\boldsymbol{x}) &= P_{\text{T}}(Y = m \,|\, \boldsymbol{x}, S = 1) \\ &= \frac{\lambda_m h_m(\beta_{01} + \boldsymbol{x}'\boldsymbol{\beta}_1, \ldots, \beta_{0q} + \boldsymbol{x}'\boldsymbol{\beta}_q)}{\sum_{j=1}^{q} \lambda_j h_j(\beta_{01} + \boldsymbol{x}'\boldsymbol{\beta}_1, \ldots) + \lambda_{q+1}(1 - \sum_{j=1}^{q} h_j(\beta_{01} + \boldsymbol{x}'\boldsymbol{\beta}_1, \ldots))}, \end{aligned} \tag{5}$$

for $m = 1, \ldots, q$. The false model that ignores the retrospective sampling scheme is described by

$$\pi_m^{\text{F}}(\boldsymbol{x}) = P_{\text{F}}(Y = m \,|\, \boldsymbol{x}, S = 1) = h_m(\beta_{01}^* + \boldsymbol{x}'\boldsymbol{\beta}_1^*, \ldots, \beta_{0q}^* + \boldsymbol{x}'\boldsymbol{\beta}_q^*).$$

Note that when $\lambda_1 = \lambda_2 = \cdots = \lambda_{q+1}$, then $\pi_m^{\text{T}}(\boldsymbol{x}) = \pi_m^{\text{F}}(\boldsymbol{x})$ for all m and the two likelihoods agree perfectly. However, in a typical outcome dependent design, sampling rates for the rare outcome categories are much higher than sampling rates for the controls or the commonly prevalent outcome category, and this equality is extremely unlikely to hold in any practical situation.

It is well known that the MLEs from the false model converge to $(\beta_{01}^*, \ldots, \beta_{0q}^*, \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_q^*)$ that minimizes the Kullback-Leibler (KL) divergence between the true model and the false model (Akaike, 1973 and Huber, 1967). The KL-divergence between the two models is defined as

$$\begin{aligned} KLD(\text{T, F}) &= E_X\left[E_{Y/X}\left\{\log \frac{\boldsymbol{\pi}_Y^{\text{T}}(\boldsymbol{x})}{\boldsymbol{\pi}_Y^{\text{F}}(\boldsymbol{x})}\right\}\right] \\ &= E_X\left[\sum_{j=1}^{q} \pi_j^{\text{T}}(\boldsymbol{x}) \log \frac{\pi_j^{\text{T}}(\boldsymbol{x})}{\pi_j^{\text{F}}(\boldsymbol{x})} + \left\{1 - \sum_{j=1}^{q} \pi_j^{\text{T}}(\boldsymbol{x})\right\} \log \frac{\{1 - \sum_{j=1}^{q} \pi_j^{\text{T}}(\boldsymbol{x})\}}{\{1 - \sum_{j=1}^{q} \pi_j^{\text{F}}(\boldsymbol{x})\}}\right] \end{aligned}$$

So $(\beta_{01}^*, \ldots, \beta_{0q}^*, \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_q^*)$, which minimize $KLD(T, F)$, solve the system of equations:

$$\begin{aligned} \frac{\partial}{\partial \beta_{0m}^*} KLD(\text{T, F}) &= 0 \quad \text{for } m = 1, \ldots, q, \\ \frac{\partial}{\partial \boldsymbol{\beta}_m^*} KLD(\text{T, F}) &= 0 \quad \text{for } m = 1, \ldots, q. \end{aligned} \tag{6}$$

Let us consider a single covariate x, to simplify the notations. The results and proof directly translate to multiple covariates. With a single x, the equations in (6) can be expressed as,

$$E_X\left[\sum_{j=1}^{q}\frac{\pi_j^{\mathrm{T}}(x)}{\pi_j^{\mathrm{F}}(x)}\frac{\partial}{\partial\beta_{0m}^*}\pi_j^{\mathrm{F}}(x)+\frac{\{1-\sum_{j=1}^{q}\pi_j^{\mathrm{T}}(x)\}}{\{1-\sum_{j=1}^{q}\pi_j^{\mathrm{F}}(x)\}}\frac{\partial}{\partial\beta_{0m}^*}\left\{1-\sum_{j=1}^{q}\pi_j^{\mathrm{F}}(x)\right\}\right]=0 \qquad (7)$$

and

$$E_X\left[x\left\{\sum_{j=1}^{q}\frac{\pi_j^{\mathrm{T}}(x)}{\pi_j^{\mathrm{F}}(x)}\frac{\partial}{\partial\beta_{m}^*}\pi_j^{\mathrm{F}}(x)+\frac{\{1-\sum_{j=1}^{q}\pi_j^{\mathrm{T}}(x)\}}{\{1-\sum_{j=1}^{q}\pi_j^{\mathrm{F}}(x)\}}\frac{\partial}{\partial\beta_{m}^*}\left\{1-\sum_{j=1}^{q}\pi_j^{\mathrm{F}}(x)\right\}\right\}\right]=0 \qquad (8)$$

for $m=1,\ldots,q$.

Remark 1: Suppose there is no association between Y and X, i.e., $\beta_1=\beta_2=\ldots=\beta_q=0$, then $\pi_j^{\mathrm{T}}(x)$ is independent of X. Without loss of generality, let $E(X)=0$. Then, if $\beta_1^*=\beta_2^*=\ldots=\beta_q^*=0$, each equation in (8) is a multiple of X and has expected value 0. Therefore, $\beta_1^*=\beta_2^*=\ldots=\beta_q^*=0$ is a solution to the equations in (8). Thus, under the null model, using a prospective likelihood, ignoring the sampling scheme, does provide consistent ML estimation for β_m, $m=1,\cdots,q$.

Remark 2: Values of $(\beta_{01}^*,\ldots,\beta_{0q}^*,\beta_1^*,\ldots,\beta_q^*)$ which result in

$$\pi_j^{\mathrm{T}}(x)=\pi_j^{\mathrm{F}}(x)$$

for all x, trivially satisfy (7) and (8); the right hand sides of these equations then reduce to the expectation of true score function, which is zero by classical ML theory.

In a general setting, solving (7) and (8) is considerably difficult. We adopt the route followed in Neuhaus (1999, 2002) by solving an alternate system of equations.

For the multivariate generalized linear model as described in (1), namely, $\boldsymbol{\pi}(\boldsymbol{x}_i)=\boldsymbol{h}(\boldsymbol{Z}_i\boldsymbol{\beta})$, consider the link function denoted by $\boldsymbol{g}=\boldsymbol{h}^{-1}$. The equivalent model is written as

$$\boldsymbol{g}(\boldsymbol{\pi}(\boldsymbol{x}_i))=\boldsymbol{Z}_i\boldsymbol{\beta},$$

where $\boldsymbol{g}=(g_1,\ldots,g_q)'$ is a vector function from $\boldsymbol{R}^q\rightarrow\boldsymbol{R}^q$. For a simple case with only one

covariate x, the model in terms of the link functions can be written as,

$$\begin{bmatrix} g_1(\pi_1(x),\ldots,\pi_q(x)) \\ g_2(\pi_1(x),\ldots,\pi_q(x)) \\ \vdots \\ g_q(\pi_1(x),\ldots,\pi_q(x)) \end{bmatrix} = \begin{bmatrix} \beta_{01}+\beta_1 x \\ \beta_{02}+\beta_2 x \\ \vdots \\ \beta_{0q}+\beta_q x \end{bmatrix}$$

Therefore, the covariate effects under the FALSE prospective model are measured by

$$\begin{bmatrix} g_1(\pi_1^{\mathrm{F}}(x+1),\ldots,\pi_q^{\mathrm{F}}(x+1)) - g_1(\pi_1^{\mathrm{F}}(x),\ldots,\pi_q^{\mathrm{F}}(x)) \\ g_2(\pi_1^{\mathrm{F}}(x+1),\ldots,\pi_q^{\mathrm{F}}(x+1)) - g_2(\pi_1^{\mathrm{F}}(x),\ldots,\pi_q^{\mathrm{F}}(x)) \\ \vdots \\ g_q(\pi_1^{\mathrm{F}}(x+1),\ldots,\pi_q^{\mathrm{F}}(x+1)) - g_q(\pi_1^{\mathrm{F}}(x),\ldots,\pi_q^{\mathrm{F}}(x)) \end{bmatrix} = \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_q^* \end{bmatrix}. \tag{9}$$

Similarly, the covariate effects under the TRUE retrospective model are measured by

$$\begin{bmatrix} g_1(\pi_1^{\mathrm{T}}(x+1),\ldots,\pi_q^{\mathrm{T}}(x+1)) - g_1(\pi_1^{\mathrm{T}}(x),\ldots,\pi_q^{\mathrm{T}}(x)) \\ g_2(\pi_1^{\mathrm{T}}(x+1),\ldots,\pi_q^{\mathrm{T}}(x+1)) - g_2(\pi_1^{\mathrm{T}}(x),\ldots,\pi_q^{\mathrm{T}}(x)) \\ \vdots \\ g_q(\pi_1^{\mathrm{T}}(x+1),\ldots,\pi_q^{\mathrm{T}}(x+1)) - g_q(\pi_1^{\mathrm{T}}(x),\ldots,\pi_q^{\mathrm{T}}(x)) \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{bmatrix}. \tag{10}$$

To relate the β^*s to the βs we try to find an approximate solution for which, $\boldsymbol{g}(\boldsymbol{\pi}^T(\boldsymbol{x})) \approx \boldsymbol{g}(\boldsymbol{\pi}^F(\boldsymbol{x}))$. This is achieved by first equating the LHS of (10) to the RHS of (9).

$$\begin{bmatrix} H_1(\beta_1,\ldots,\beta_q) \\ H_2(\beta_1,\ldots,\beta_q) \\ \vdots \\ H_q(\beta_1,\ldots,\beta_q) \end{bmatrix} = \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_q^* \end{bmatrix} \tag{11}$$

where $H_l(\beta_1,\ldots,\beta_q) = g_l(\pi_1^{\mathrm{T}}(x+1),\ldots,\pi_q^{\mathrm{T}}(x+1)) - g_l(\pi_1^{\mathrm{T}}(x),\ldots,\pi_q^{\mathrm{T}}(x))$, for $l = 1,\cdots,q$.

Next, we carry out a first order multivariate Taylor's expansion of the elements $H_l(\beta_1,\ldots,\beta_q)$ around $\boldsymbol{\beta} = (0,\ldots,0)$. Note that $H_l(0,\ldots,0) \equiv 0$ for all $l = 1,\cdots,q$. The details of the Taylor's expansion are relegated to the Appendix (A.2). Combining the first order Taylor's expansion with the matrix equation in (11) we have,

$$\begin{bmatrix} \frac{\partial}{\partial\beta_1}H_1(\beta_1,\ldots,\beta_q)\,|_{(0,\ldots,0)} & \cdots & \frac{\partial}{\partial\beta_q}H_1(\beta_1,\ldots,\beta_q)\,|_{(0,\ldots,0)} \\ \frac{\partial}{\partial\beta_1}H_2(\beta_1,\ldots,\beta_q)\,|_{(0,\ldots,0)} & \cdots & \frac{\partial}{\partial\beta_q}H_2(\beta_1,\ldots,\beta_q)\,|_{(0,\ldots,0)} \\ \vdots & & \\ \frac{\partial}{\partial\beta_1}H_q(\beta_1,\ldots,\beta_q)\,|_{(0,\ldots,0)} & \cdots & \frac{\partial}{\partial\beta_q}H_q(\beta_1,\ldots,\beta_q)\,|_{(0,\ldots,0)} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{bmatrix} = \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_q^* \end{bmatrix}$$

Where the derivative at the null model for each H_l (generically denoted as H in the following) can be evaluated as,

$$\begin{aligned} & \frac{\partial}{\partial \beta_m} H(\beta_1, \ldots, \beta_q) \mid_{(0,\ldots,0)} \\ = & \sum_{j=1}^{q} g^{(j)}(\pi_{10}^{\mathrm{T}}, \ldots, \pi_{q0}^{\mathrm{T}}) \times \left[\frac{G_{jm}(\beta_{01}, \ldots, \beta_{0q})}{\{\sum_{t=1}^{q}(r_t - 1)h_t(\beta_{01}, \ldots, \beta_{0q}) + 1\}^2} \right], \end{aligned} \tag{12}$$

where $r_t = \lambda_t/\lambda_{q+1}$, and we follow the convention that for any function $f(u_1, \cdots, u_q)$, $f^{(i)}(u_1, \cdots, u_q)$ is the partial derivative of f with respect to the i-th co-ordinate u_i. The function G_{jm} is defined as

$$\begin{aligned} G_{jm}(\beta_{01}, \ldots, \beta_{0q}) &= r_j h_j^{(m)}(\beta_{01}, \ldots, \beta_{0q}) \left[\sum_{t=1}^{q}(r_t - 1)h_t(\beta_{01}, \ldots, \beta_{0q}) + 1 \right] \\ &- r_j h_j(\beta_{01}, \ldots, \beta_{0q}) \left[\sum_{t=1}^{q}(r_t - 1)h_t^{(m)}(\beta_{01}, \ldots, \beta_{0q}) \right], \end{aligned}$$

and

$$\pi_{j0}^{\mathrm{T}} = \frac{\lambda_j h_j(\beta_{01}, \ldots, \beta_{0q})}{\sum_{t=1}^{q} \lambda_t h_t(\beta_{01}, \ldots, \beta_{0q}) + \lambda_{q+1}(1 - \sum_{t=1}^{q} h_t(\beta_{01}, \ldots, \beta_{0q}))}$$

denotes the probabilities for category j, under the null model.

Thus we have related the true model parameters to the limiting values of the MLE's under the false model by an equation of the form

$$\boldsymbol{\beta} = \boldsymbol{H}^{-1}\boldsymbol{\beta}^{*} \tag{13}$$

where $\boldsymbol{H}$ is a $q \times q$ matrix with entries depending on the sampling ratios (λ_m/λ_{q+1}), and the intercepts (β_{0m}), $m = 1, \cdots, q$. Equivalently, a knowledge of the disease risk for each category at the baseline value of the covariate x and the sampling rates is necessary to compute the matrix $\boldsymbol{H}$.

Remark 3: As shown in Neuhaus (2002), when $q = 1$, that is, for GLMs for binary data with any general link function g, and $h = g^{-1}$, $\frac{\partial}{\partial \beta_1} H(\beta_1)|_{\beta_1=0}$ simplifies to

$$\frac{g^{(1)}(\pi_0)\pi_0(1 - \pi_0)}{g^{(1)}(\mu_0)\mu_0(1 - \mu_0)},$$

where

$$\begin{aligned} \pi_0 &= \frac{rh(\beta_{01})}{(r-1)h(\beta_{01}) + 1}, \\ \mu_0 &= h(\beta_{01}) \quad \text{and,} \quad g^{(1)}(\pi_0) = \frac{\partial g(\pi)}{\partial \pi}|_{\pi=\pi_0}. \end{aligned}$$

This bias factor could be greater than or less than one depending on the sampling ratio $r = \lambda_1/\lambda_2$, the link function, and the baseline disease risk.

Since the sampling rates and baseline disease risks are typically unknown for a given study, it is potentially difficult to adopt a bias correction strategy based on the expression in (13). The purpose of this note is to study this bias analytically and present a clear illustration through the following data example.

4 Illustration through real data example

The data example is based on the large ongoing Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial at the National Cancer Institute, USA (Gohagan et al., 2000; Hayes et al., 2005). The association between tobacco smoking and colorectal adenoma and hyperplastic polyps in this trial has been documented in Ji et al (2006), and we use the same dataset. Data is available on patients with sigmoidoscopy screening of the left side of the distal colon. Patients are classified into three disease states based on the number of adenomas detected on the left side (1=sigmoidoscopy negative, 2=single adenoma, 3=multiple adenoma). We consider a subjects's cigarette smoking behavior (0=never and 1=ever, which includes both former and current smokers) as the only risk factor X. After deleting subjects with missing observations, we have complete information on 47364 subjects in the trial. The cohort data is represented by the following frequency table

Adenoma	1	2	3
Smoking			
0	20420	1234	329
1	22397	2213	771
Total	42817	3447	1100

In view of the natural ordering of the disease states, one may be inclined to fit one of the most popular models for ordered categorical outcomes, namely, the cumulative logit model (Agresti, 2002) given by,

$$\text{logit}[P(Y \leq m|X)] = \beta_{0m} + \beta_m X, m = 1, \cdots, q = K - 1. \tag{14}$$

Instead of the popular proportional odds structure, we do allow separate covariate effects (β_m) for each cumulative logit as that model appears to be more reasonable in the current context. This model is also known as the partial proportional odds model (Peterson and Harrell, 1990). We first analyze the available data on the whole cohort of 47364 subjects using the above cumulative logit model with smoking history as the risk factor of interest. The fitted model is given by,

$$\text{logit}[P(Y \leq 1|X)] = 2.570 - 0.554X \qquad \text{logit}[P(Y \leq 2|X)] = 4.187 - 0.724X. \tag{15}$$

The results suggest that the smokers are less likely to have no adenoma (versus more than one adenoma) and less likely to have single or no adenoma (versus multiple adenoma) than the non-smokers. Both the cumulative log odds ratio parameters are statistically significant ($P < 0.001$). We can consider these fitted values as the 'TRUE' values of the parameters, as obtained via a prospective study of the full cohort.

Suppose we now take a retrospective sample from the given cohort, conditional on the multiple adenoma category and then analyze the retrospective data by the cumulative logit model, ignoring the sampling design. Note that the cumulative logit model does not have a multiplicative intercept structure as required by Theorem 1 for prospective-retrospective equivalence, thus the estimates of β_1 and β_2 obtained by this analysis of the retrospectively collected data will be typically different from the ones obtained in (15). The difference in magnitude of the two estimates will reflect the resultant bias. We furnish an empirical estimate of the bias factor by first taking repeated retrospective samples from the cohort under a given sampling design (with fixed sampling rates for each category) and then calculating the ratio of the mean of the resultant estimates with the "true" estimate obtained in (15). We compare this estimated bias with the bias computed by using our analytical approximation formula as given in Section 3, under the same design and parameter setting. The numerical results are collected in Table 1, whereas the analytical details specific to the cumulative logit model are available in Appendix A.3. Table 1 clearly brings out the fact that with multiple disease categories, ignoring the sampling design may provide quite inaccurate point estimate of disease-exposure association depending on the sampling rates. We also notice that our analytical approximation is remarkably close to the empirical estimate of the bias factor.

Because of the special logistic structure of the cumulative logit model in terms of the cumulative probabilities, it can be noted from Table 1 and also Appendix A.3 that whenever $\lambda_2 = \lambda_3$, an unbiased estimate of β_1 can be obtained, though the estimate of β_2 remain biased. Only in the event of $\lambda_1 = \lambda_2 = \lambda_3$, both the estimates of β_1 and β_2 are unbiased.

Figure 1 plots the bias factor (β_m^*/β_m, $m = 1, 2$) as obtained by our analytical formulae, when 1500 controls ($Y = 1$) are selected from the 42817 controls in our cohort, and the sampling rates for the outcome categories $Y = 2$ and $Y = 3$ vary freely from 0 to 1. The values of the intercept parameters are set at the estimates obtained in (15). One can note that under this setting, the estimate of β_1 is inflated, whereas the estimate of β_2 is deflated. The bias seems to be more severe for β_2 for a wide range of sampling rates, whereas the bias in β_1 is significant for small values of λ_2 or small values of λ_3 (< 0.2).

Figure 2 represents one of the common designs used in practice, when one includes half/all available cases in the case-control sample. Since in both of the designs, $\lambda_2 = \lambda_3$, the estimate of β_1 is unbiased. The bias factor for β_2 is plotted as a function of λ_1, the sampling rate for the controls and one can notice that the plotted curve crosses the vertical axis at 1 (reflecting no bias) only when $\lambda_1 = \lambda_2 = \lambda_3$. The figure also indicates that sampling 20-30% controls is sufficient to reduce much of the bias under such a sampling design, with a baseline disease risk as noticed in the colorectal adenoma data. If one has prior information on the baseline disease risks from past historical data, and a prospective model is implemented, the bias approximation could be used to evaluate possible sampling strategies for a given study.

5 Concluding Remarks

In this note, we consider the problem of fitting multivariate generalized linear models for categorical outcomes under an outcome dependent sampling scheme. We first provide a rigorous characterization result for the link functions which allow prospective and retrospective equivalence and then provide an approximation to the bias incurred by ignoring the sampling scheme. The characteri-

zation illustrates that for categorical outcomes, prospective-retrospective equivalence of likelihood inference in terms of the regression parameters do not hold beyond the generalized multinomial logit links. Although for binary outcomes, similar issues have been investigated thoroughly, results of this nature have not previously been collected in the literature for a general categorical outcome variable. The findings imply that direct prospective approaches which consider flexible non-parametric modeling of link functions for categorical outcomes, are not appropriate under outcome dependent sampling scheme unless some additional supplementary information is included (Scott and Wild, 1986). The real data example based on the PLCO trial, where case-control samples are selected from a prospective cohort, is reflective of how many of the case-control studies are carried out in practice. We study the bias under some common sampling designs one may implement in a real investigation. Though we illustrate the results with the partial proportional odds model, there are other commonly used models for polytomous outcome, like the continuation-ratio logit model (Agresti, 2002), which models logit of $P(Y = j|Y \geq j, \boldsymbol{x})$, does not fall in the generalized multinomial logit class. Since this link function lies somewhere intermediate between the multinomial and the cumulative logit links, it will be another interesting link function to investigate. The purpose of this note is to leave the reader with an analytical and practical understanding of the bias mechanism for multicategory outcomes, when common prospective models are fitted by ignoring an outcome dependent sampling process.

ACKNOWLEDGEMENTS: The authors will like to thank Nilanjan Chatterjee for many helpful comments. The research of Bhramar Mukherjee was partially supported by NSA Young Investigator Grant H98230-06-1-0033.

6 References

1. Agresti, A. (2002). *Categorical Data Analysis*, Second Edition. John Wiley, New York.

2. Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In: B. N. Petrov and F. Cszaki, eds., Second International Symposium on Information

Theory, 267–281. Budapest: Akademiai Kiadó.

3. Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society*, Ser B **32**, 283–301.

4. Breslow, N.E. and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, **75**, 11-20.

5. Breslow, N.E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics*, **4**, 457-468.

6. Breslow, N.E. and Holubkov, R. (1997a). Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *Journal of the Royal Statistical Society, B*, **59**, 447-461.

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

8. Chatterjee, N. (2004). A two-stage regression model for epidemiological studies with multivariate disease classification data. *Journal of the American Statistical Association*, **99**, 127-138.

9. Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Second Edition, Springer.

10. Gohagan J. K, Prorok P. C., Hayes R. B., Kramer B. S. (2000). Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Project Team. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Controlled Clinical Trials.* **21(6 Suppl)**, 273S–309S.

11. Hayes R. B., Sigurdson A., Moore L., Peters U., Huang W. Y., Pinsky P., Reding D., Gelmann E. P., Rothman N., Pfeiffer R. M., Hoover R. N., Berg C. D. (2005). Methods for

etiologic and early marker investigations in the PLCO trial. *Mutation Research* **592**, 147–154.

12. Huber, P. J. (1967). The Behavior of Maximum-likelihood Estimates under Non-standard Conditions. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley CA: University of California Press: 221-33.

13. Ji, B-T, Weissfeld, J.L., Chow, W-H, Huang, W-Y, Schoen, R.E., Hayes, R. B. (2006). Tobacco smoking and colorectal hyperplastic and adenomatous polyps. *Cancer Epidemiology, Biomarkers and Prevention*, **15**, 897-901.

14. Kagan, A. (2001). A note on the logistic link function. *Biometrika* **88**, 599–601.

15. Manski, C.F. and McFadden, D. (1981). *Structural Analysis of Discrete Data with Applications.* MIT Press, Cambridge.

16. McCullagh, P. and Nelder, J.A. (1999). *Generalized Linear Models*, 2nd Edition. Chapman & Hall, New York.

17. Neuhaus, J. M. (2002). Bias due to ignoring the sample design in case-control studies. *Australian & New Zealand Journal of Statistics*, **44**, 285–293.

18. Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, **88**, 843-855.

19. Peterson, B. and Harrell, F. E. (1990). Partial proportion odds models for ordinal response variables. *Applied Statistics*, **39**, 205-217.

20. Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, **8**, 1087-1114.

21. Pfeffermann, D. and Sverchov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhya B*, **61**, 166-186.

22. Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403-411.

23. Scott, A.J. and Wild, C.J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society, B*, **48**, 170-182.

24. Scott, A.J. and Wild, C.J. (1991). Fitting logistic models in stratified case-control studies. *Biometrics*, **47**, 497-510.

25. Scott, A.J. and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, **84**, 57-71.

26. Wang, C.Y., Wang, S. and Carroll, R.J. (1997). Estimation in choice-based sampling with measurement error and bootstrap analysis. *Journal of Econometrics*, **77**, 65-86.

27. White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1-25.

28. Wild, C.J. (1991). Fitting prospective regression models to case-control data. *Biometrika*, **78**, 705-717.

7 Appendix

A.1 Proof of Theorem 1: We first establish the necessity part of Theorem 1, i.e., (3) implies (4). Let $Y_i = m$ for all i, that all individuals are selected from the m-th response category. (i.e., $y_{im} = 1$ for all $i = 1, \ldots, n$ and $y_{ij} = 0$ for all $j \neq m$ and $i = 1, \ldots, n$. Then the equality in (3) becomes

$$\prod_{i=1}^{n} \frac{\lambda_m h_m(u_{i1}, \ldots, u_{iq})}{\sum_{j=1}^{q} \lambda_j h_j(u_{i1}, \ldots, u_{iq}) + \lambda_{q+1} \left(1 - \sum_{j=1}^{q} h_j(u_{i1}, \ldots, u_{iq})\right)}$$

$$= \prod_{i=1}^{n} h_m(u_{i1} + \theta_1(\boldsymbol{\lambda}), u_{i2} + \theta_2(\boldsymbol{\lambda}), \ldots, u_{iq} + \theta_q(\boldsymbol{\lambda}))$$

Since $u_{i1}, \ldots, u_{iq}$ for $i = 1, \ldots, n$ are free variables with range $\mathcal{R}$, this implies

$$\begin{aligned} & \frac{\lambda_m h_m(u_1, \ldots, u_q)}{\sum_{j=1}^{q} \lambda_j h_j(u_1, \ldots, u_q) + \lambda_{q+1}\left(1 - \sum_{j=1}^{q} h_j(u_1, \ldots, u_q)\right)} \\ = \; & h_m(u_1 + \theta_1(\boldsymbol{\lambda}), u_{i2} + \theta_2(\boldsymbol{\lambda}), \ldots, u_q + \theta_q(\boldsymbol{\lambda})), \end{aligned} \tag{16}$$

By dividing the numerator and denominator of LHS of (16) by $(1 - \sum_{j=1}^{q} h_j(u_1, \ldots, u_q))$, we have

$$\frac{\lambda_m \tilde{h}_m(u_1, \ldots, u_q)}{\sum_{j=1}^{q} \lambda_j \tilde{g}_j(u_1, \ldots, u_q) + \lambda_{q+1}} = h_m(u_1 + \theta_1(\boldsymbol{\lambda}), \ldots, u_q + \theta_q(\boldsymbol{\lambda})) \tag{17}$$

Where $\tilde{h}_m(u_1, \ldots, u_q) = \lambda_m h_m(u_1, \ldots, u_q)/(1 - \sum_{j=1}^{q} h_j(u_1, \ldots, u_q))$.

Summing both sides of (17) over m and subtracting from 1, we have

$$\frac{\lambda_{q+1}}{\sum_{m=1}^{q} \lambda_m \tilde{h}_m(u_1, \ldots, u_q) + \lambda_{q+1}} = 1 - \sum_{m=1}^{q} h_m(u_1 + \theta_1(\boldsymbol{\lambda}), \ldots, u_q + \theta_q(\boldsymbol{\lambda})) \tag{18}$$

Dividing (17) by (18), and then taking logarithms on each side, we have

$$\log \tilde{h}_m(u_1, \ldots, u_q) + \log\left(\frac{\lambda_m}{\lambda_{q+1}}\right) \;\; = \;\; \log \tilde{h}_m(u_1 + \theta_1(\boldsymbol{\lambda}), \ldots, u_q + \theta_q(\boldsymbol{\lambda})) \tag{19}$$

The above equation (19), is of the form,

$$A_m(u_1, \ldots, u_q) + B_m(\boldsymbol{\lambda}) = A_m(u_1 + \theta_1(\boldsymbol{\lambda}), \ldots, u_q + \theta_q(\boldsymbol{\lambda})),$$

where $A_m = \tilde{h}_m$ and $B_m(\boldsymbol{\lambda}) = \log(\lambda_m/\lambda_{q+1})$.

Let $\boldsymbol{u} = (u_1, \cdots, u_q)'$ and $\boldsymbol{v} = [\boldsymbol{\theta}(\boldsymbol{\lambda})] = (\boldsymbol{\theta}(\lambda_1), \cdots, \boldsymbol{\theta}(\lambda_q))'$. We may rewrite $B_m(\boldsymbol{\lambda}) = B_m(f^{-1}(\boldsymbol{\theta}(\boldsymbol{\lambda}))) = B_m(f^{-1}(\boldsymbol{v}))$, where $f : \boldsymbol{\lambda} \to \boldsymbol{\theta}(\boldsymbol{\lambda})$ is a one to one and onto mapping according to Theorem 1, then the above equation can be written in the form,

$$A_m(\boldsymbol{u}) + \tilde{B}_m(\boldsymbol{v}) = A_m(\boldsymbol{u} + \boldsymbol{v}),$$

where $\tilde{B}_m = B_m \circ f^{-1}$.

We will now need the following lemma.

Lemma 1: Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be $q \times 1$ vectors and A, B be continuous functions from $\boldsymbol{R}^q \to \boldsymbol{R}$ such that,

$$A(\boldsymbol{u}) + B(\boldsymbol{v}) = A(\boldsymbol{u} + \boldsymbol{v}) \quad \forall \; \boldsymbol{u}, \boldsymbol{v}, \tag{20}$$

Then,

$$A(\boldsymbol{u}) = \boldsymbol{c}'\boldsymbol{u} + d.$$

Proof: By (20), we have, for any set of vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$,

$$\begin{aligned} A(\boldsymbol{u}+\boldsymbol{v}+\boldsymbol{w}) &= A(\boldsymbol{u}) + B(\boldsymbol{v}+\boldsymbol{w}) \text{ and also,} \\ A(\boldsymbol{u}+\boldsymbol{v}+\boldsymbol{w}) &= A(\boldsymbol{u}+\boldsymbol{v}) + B(\boldsymbol{w}) = A(\boldsymbol{u}) + B(\boldsymbol{v}) + B(\boldsymbol{w}). \end{aligned}$$

Therefore,

$$B(\boldsymbol{v}+\boldsymbol{w}) = B(\boldsymbol{v}) + B(\boldsymbol{w}).$$

By the above property of B, for every rational number r, and vector $\boldsymbol{u}$, we have $B(r\boldsymbol{u}) = rB(\boldsymbol{u})$. Implying the linearity of B (recall that B is continuous), i.e., $B(\boldsymbol{u}) = \boldsymbol{c}'\boldsymbol{u}$, for some vector $\boldsymbol{c}$. Thus by (20), we have,

$$A(\boldsymbol{u}) = A(\boldsymbol{0}) + B(\boldsymbol{u}) = \boldsymbol{c}'\boldsymbol{u} + A(\boldsymbol{0}) = \boldsymbol{c}'\boldsymbol{u} + d$$

where $A(0) = d$, is some scalar. Therefore, $A(\boldsymbol{u})$ is linear in $\boldsymbol{u}$. By the relationship $B(\boldsymbol{v}) = A(\boldsymbol{v}) - A(0)$, it follows that $B(\boldsymbol{v}) = \boldsymbol{c}'\boldsymbol{v}$.

Returning to the proof of Theorem 1, applying Lemma 1 directly to (19), exponentiating and normalizing, we have,

$$h_m(\boldsymbol{u}) = \frac{\exp(\boldsymbol{c}_m'\boldsymbol{u} + d_m)}{1 + \sum_{l=1}^{q} \exp(\boldsymbol{c}_l'\boldsymbol{u} + d_l)}, \tag{21}$$

Letting $\tilde{B}_m(\boldsymbol{v}) = B(\boldsymbol{v})$ in Lemma 1, it also follows that,

$$\log\left(\frac{\lambda_m}{\lambda_{q+1}}\right) = \boldsymbol{c}_m'\boldsymbol{\theta}(\boldsymbol{\lambda}).$$

Translating in terms of the model parameters, we have, $\boldsymbol{c}_m'\boldsymbol{u} = \sum_{j=1}^{q} c_{mj}u_j = \sum_{j=1}^{q} c_{mj}(\beta_{0j} + \boldsymbol{x}'\boldsymbol{\beta}_j) = \beta_{0m}^* + \boldsymbol{x}'\boldsymbol{\beta}_m^*$. Thus, $h_m(\boldsymbol{x})$ is a response function with multiplicative intercept and odds structure and we have the necessity part of Theorem 1.

The sufficiency part follows by simple algebra, plugging in a response function with multiplicative intercept and odds structure in (3) and verifying that the result holds .

A.2 The details of the Taylor's approximation

In the following we suppress the suffix l in H_l. By first order Taylor's expansion, we have,

$$\begin{aligned} H(\beta_1, \ldots, \beta_q) &\approx H(0, \ldots, 0) + \sum_{j=1}^{q} \beta_j \frac{\partial}{\partial \beta_j} H(\beta_1, \ldots, \beta_q) \mid_{(0,\ldots,0)} \\ &= \sum_{j=1}^{q} \beta_j \frac{\partial}{\partial \beta_j} H(\beta_1, \ldots, \beta_q) \mid_{(0,\ldots,0)} \end{aligned}$$

Recall that, $H(\beta_1, \cdots, \beta_q) = g(\boldsymbol{\pi}^{\mathrm{T}}(x)) - g(\boldsymbol{\pi}^{\mathrm{T}}(x+1))$. The derivative of g can be obtained as,

$$\begin{aligned} & \frac{\partial}{\partial \beta_m} g(\pi_1^{\mathrm{T}}[\beta_{01} + \beta_1 x, \ldots, \beta_{0q} + \beta_q x], \ldots, \pi_q^{\mathrm{T}}[\beta_{01} + \beta_1 x, \ldots, \beta_{0q} + \beta_q x]) \\ = & \sum_{j=1}^{q} \frac{\partial}{\partial \pi_j^{\mathrm{T}}} g[\pi_1^{\mathrm{T}}, \ldots, \pi_q^{\mathrm{T}}] \times \frac{\partial}{\partial \beta_m} \pi_j^{\mathrm{T}}(\beta_{01} + \beta_1 x, \ldots, \beta_{0q} + \beta_q x) \\ = & \sum_{j=1}^{q} \frac{\partial}{\partial \pi_j^{\mathrm{T}}} g[\pi_1^{\mathrm{T}}, \ldots, \pi_q^{\mathrm{T}}] \times \frac{\partial}{\partial u_m} \pi_j^{\mathrm{T}}(u_1, \ldots, u_q) \times x \end{aligned}$$

By taking the difference of two such derivatives at $x+1$ and x, we evaluate the derivative of H as

$$\frac{\partial}{\partial \beta_m} H(\beta_1, \ldots, \beta_q) = \sum_{j=1}^{q} \frac{\partial}{\partial \pi_j^{\mathrm{T}}} g[\pi_1^{\mathrm{T}}, \ldots, \pi_q^{\mathrm{T}}] \times \frac{\partial}{\partial u_m} \pi_j^{\mathrm{T}}(u_1, \ldots, u_q), \tag{22}$$

where $u_m = \beta_{0m} + \beta_m x$. Let

$$g^{(j)}(\pi_1, \ldots, \pi_q) = \frac{\partial}{\partial \pi_j} g(\pi_1, \ldots, \pi_q)$$

We can write the derivative of π_j^T as

$$\begin{aligned} \frac{\partial}{\partial u_m} \pi_j^{\mathrm{T}}(u_1, \ldots, u_q) &= \frac{\partial}{\partial u_m} \left[\frac{\lambda_j h_j(u_1, \ldots, u_q)}{\sum_{t=1}^{q} \lambda_t h_t(u_1, \ldots, u_q) + \lambda_{q+1}(1 - \sum_{t=1}^{q} h_t(u_1, \ldots, u_q))} \right] \\ &= \frac{\partial}{\partial u_m} \left[\frac{r_j h_j(u_1, \ldots, u_q)}{\sum_{t=1}^{q} (r_t - 1) h_t(u_1, \ldots, u_q) + 1} \right], \end{aligned} \tag{23}$$

where

$$\begin{aligned} r_j &= \text{sampling ratio of } Y = j \text{ to the baseline group of } Y = q+1 \\ &= \frac{\lambda_j}{\lambda_{q+1}} \end{aligned}$$

The derivative in (23) becomes

$$\frac{\partial}{\partial u_m} \left[\frac{r_j h_j(u_1, \ldots, u_q)}{\sum_{i=1}^{q} (r_t - 1) h_t(u_1, \ldots, u_q) + 1} \right] = \frac{G_{jm}(u_1, \ldots, u_q)}{[\sum_{t=1}^{q} (r_t - 1) h_t(u_1, \ldots, u_q) + 1]^2},$$

where

$$\begin{aligned} G_{jm}(u_1,\ldots,u_q) &= r_j h_j^{(m)}(u_1,\ldots,u_q)\left[\sum_{t=1}^{q}(r_t-1)h_t(u_1,\ldots,u_q)+1\right] \\ &- r_j h_j(u_1,\ldots,u_q)\times\left[\sum_{t=1}^{q}(r_t-1)h_t^{(m)}(u_1,\ldots,u_q)\right] \end{aligned}$$

Hence we arrive at our expressions in (12).

A.3. Derivatives for the cumulative logit model

For simplicity of expressions, let us consider $q = 2$, as in the PLCO data example. To translate the cumulative logit model into the MVGLM set-up using the notations followed in the paper, we have,

$$\begin{aligned} \pi_1(x) &= h_1(\beta_{01}+\beta_1 x, \beta_{02}+\beta_2 x) = \frac{\exp(\beta_{01}+\beta_1 x)}{1+\exp(\beta_{01}+\beta_1 x)} \\ \pi_2(x) &= h_2(\beta_{01}+\beta_1 x, \beta_{02}+\beta_2 x) \\ &= \frac{\exp(\beta_{02}+\beta_2 x)}{1+\exp(\beta_{02}+\beta_2 x)} - \frac{\exp(\beta_{01}+\beta_1 x)}{1+\exp(\beta_{01}+\beta_1 x)} \end{aligned}$$

and the link functions are given by,

$$\begin{aligned} g_1(\pi_1,\pi_2) &= \log\left(\frac{\pi_1}{1-\pi_1}\right) \\ g_2(\pi_1,\pi_2) &= \log\left(\frac{\pi_1+\pi_2}{1-(\pi_1+\pi_2)}\right) \end{aligned}$$

Plugging these particular expressions in (13) we have the bias approximation in (11) as

$$\begin{bmatrix} \frac{\partial}{\partial\beta_1}H_1(\beta_1,\beta_2)\,|_{(0,0)} & \frac{\partial}{\partial\beta_2}H_1(\beta_1,\beta_2)\,|_{(0,0)} \\ \frac{\partial}{\partial\beta_1}H_2(\beta_1,\beta_2)\,|_{(0,0)} & \frac{\partial}{\partial\beta_2}H_2(\beta_1,\beta_2)\,|_{(0,0)} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \beta_1^* \\ \beta_2^* \end{bmatrix}.$$

The derivative components of the matrix are given by,

$$\begin{aligned} \frac{\partial}{\partial\beta_1}H_1(\beta_1,\beta_2)\,|_{(0,0)} &= \frac{\exp(\beta_{02})\lambda_2+\lambda_3}{\exp(\beta_{02})\lambda_2+\lambda_3+\exp(\beta_{01})(\lambda_3-\lambda_2)} \\ \frac{\partial}{\partial\beta_2}H_1(\beta_1,\beta_2)\,|_{(0,0)} &= \frac{\exp(\beta_{02})(1+\exp(\beta_{01}))(\lambda_3-\lambda_2)}{(1+\exp(\beta_{02}))(\exp(\beta_{02})\lambda_2+\lambda_3+\exp(\beta_{01})(\lambda_3-\lambda_2))} \\ \frac{\partial}{\partial\beta_1}H_2(\beta_1,\beta_2)\,|_{(0,0)} &= \frac{\exp(\beta_{01})(1+\exp(\beta_{02}))(\lambda_1-\lambda_2)}{(1+\exp(\beta_{01}))(\exp(\beta_{02}+\beta_{01})\lambda_1+\exp(\beta_{02})\lambda_2+\exp(\beta_{01})(\lambda_1-\lambda_2))} \\ \frac{\partial}{\partial\beta_2}H_2(\beta_1,\beta_2)\,|_{(0,0)} &= \frac{\exp(\beta_{02})(\lambda_1\exp(\beta_{01})+\lambda_2)}{(\exp(\beta_{02}+\beta_{01})\lambda_1+\exp(\beta_{02})\lambda_2+\exp(\beta_{01})(\lambda_1-\lambda_2))}. \end{aligned}$$

Table 1: Estimates of Bias factor under different designs when n_m individuals are sampled from disease category $Y = m$ from the PLCO cohort, $m = 1, 2, 3$. There are 42817 controls with no adenoma ($Y = 1$), 3447 cases with single adenoma ($Y = 2$) and 1100 cases with multiple adenoma ($Y = 3$) among the 47364 subjects we considered in our sampling frame. Under each design, 1000 replicates of the retrospective sample are generated. A cumulative logit model as described in (14) is fitted to each retrospective sample. The empirical estimate of the bias factor for each parameter is calculated by computing the ratio of the mean of the 1000 cumulative odds ratio estimates to the true prospective estimates. The true values of the model parameters are the estimates obtained by analyzing the cohort data: $\beta_{01} = 2.57$, $\beta_{02} = 4.18$, $\beta_1 = -0.554$ and $\beta_2 = -0.724$.

Design	Empirical estimate of bias factor β_1^*/β_1	Estimate obtained by bias approximation formula	Empirical estimate bias factor β_2^*/β_2	Estimate obtained by bias approximation formula
$n_1 = 1000, n_2 = n_3 = 500$	1.12	1.11	0.82	0.85
$n_1 = 1000, n_2 = 500, n_3 = 1000$	1.20	1.19	0.83	0.85
$n_1 = 1000, n_2 = 1000, n_3 = 500$	1.04	1.04	0.72	0.74
$n_1 = 1000, n_2 + n_3 = 1000$	1.00	1.00	0.76	0.79
$n_1 = 1500, n_2 = n_3 = 1000$	1.04	1.12	0.79	0.81
$n_1 = 1500, n_2 = 500, n_3 = 1000$	1.21	1.19	0.89	0.90
$n_1 = 1500, n_2 = 1000, n_3 = 500$	1.05	1.04	0.79	0.81
$n_1 = 4500, n_2 = 3447, n_3 = 1100$	1.00	1.00	0.76	0.79

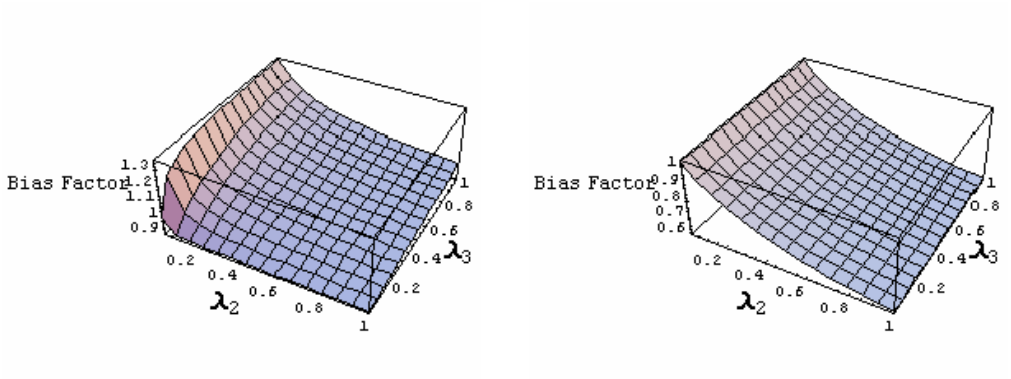

Figure 1: The figure on the left represents the bias in estimating the true parameter β_1 by the cumulative logit model where the Bias Factor plotted on the vertical axis denotes the ratio β_1^* / β_1. The figure on the right represents the bias in estimating the true parameter β_2 by the cumulative logit model where the Bias Factor plotted on the vertical axis denotes the ratio β_2^* / β_2. The other two axes in both plots represent the sampling rates for disease categories 2 and 3 (λ_2 and λ_3 respectively). The sampling rate for controls, namely, λ_1 is fixed at 1500/42817. The intercept for category 1 and category 2 are set at 2.57 and 4.18 in accordance with the analysis of the multiple adenoma data.

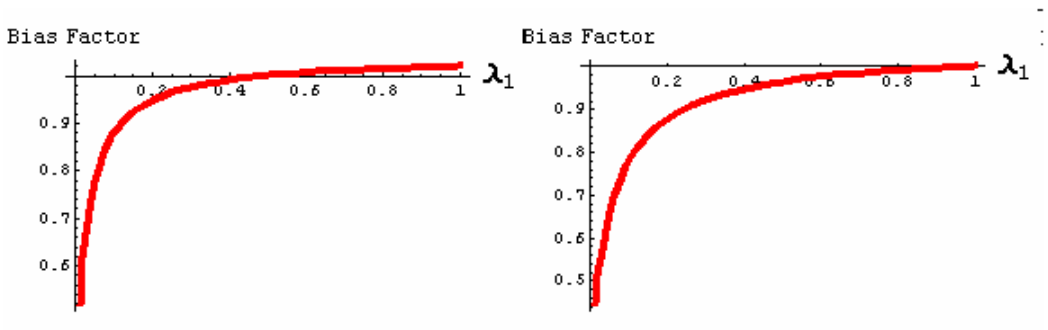

Figure 2: The two figures represent the bias in estimating the true parameter β_2 by the cumulative logit model where the Bias Factor plotted on the vertical axis denotes the ratio β_2^*/β_2. The Bias Factor is plotted as a function of λ_1, the sampling rate for the controls. The figure on the left represents the design when you select half of the available observations in categories 2 and 3, i.e. $\lambda_2=\lambda_3=0.5$, whereas the figure on the right side represents the design when you select ALL available cases, i.e. $\lambda_2=\lambda_3=1$. Note that whenever $\lambda_2=\lambda_3$, under the cumulative logit model, the estimate of β_1 is unbiased, thus we only examine the estimate of β_2.