Inference methods for the conditional logistic regression model with longitudinal data

Radu V. Craiu[1], **Thierry Duchesne***[2], and **Daniel Fortin**[3]

[1] Department of Statistics, University of Toronto, 100 St. George Street, Toronto, Ontario, M5S 3G3, Canada

[2] Département de mathématiques et de statistique, Université Laval, Québec, Québec, G1K 7P4, Canada

[3] Département de biologie, Université Laval, Québec, Québec, G1K 7P4, Canada

Received 15 November 2004, revised 30 November 2004, accepted 2 December 2004
Published online 3 December 2004

Summary

This paper considers inference methods for case-control logistic regression in longitudinal setups. The motivation is provided by an analysis of plains bison spatial location as a function of habitat heterogeneity. The sampling is done according to a longitudinal matched case-control design in which, at certain time points, exactly one case, the actual location of an animal, is matched to a number of controls, the alternative locations that could have been reached. We develop inference methods for the conditional logistic regression model in this setup, which can be formulated within a generalized estimating equation (GEE) framework. This permits the use of statistical techniques developed for GEE-based inference, such as robust variance estimators and model selection criteria adapted for non-independent data. The performance of the methods is investigated in a simulation study and illustrated with the bison data analysis.

Key words: Akaike information criterion (AIC), case-control logistic regression, estimating equations, generalized estimating equations, quasi-likelihood under independence criterion (QIC), retrospective sampling, robust sandwich estimators.

1 Introduction

In many scientific investigations in ecology and health, researchers are interested in exploring possible relationships between the characteristics of the individual's environment and a binary response. In some cases, one of the values of the response might be “rare” or the sets of covariates corresponding to both values of the response must be matched. This is the case in the experiment that motivated this work, the analysis of the spatial distribution of plains bison (Latin name *bison bison bison*) as a function of the characteristics of the environment. Bison are located every hour, and each observed location is matched to other locations that could have been reached during the 1-hour time interval. In such setups, a conditional or case-control design is one in which sampling is stratified on the values of the response variable itself.

The study of case-control data has been intensive in biostatistics. Many such studies require the close matching of control and case subjects, and there is a vast literature on conditional logistic regression (a classical reference is Breslow and Day (1980)). Scott and Wild (1986) compare the ad-hoc methods generated by sample survey techniques to a likelihood based approach in the case of conditional logistic regression. More recently, Fay et al. (1998) derived sandwich variance estimators for conditional logistic regression models and Fay and Graubard (2001) have studied how small sample adjustments to these robust variance estimates could be applied, whereas Arbogast and Lin (2004) have proposed goodness-of-fit tests for such models in non-longitudinal setups.

* Corresponding author: e-mail: duchesne@mat.ulaval.ca, Phone: 001 418 656 5077, Fax: 001 418 656 2817.

In many situations the common assumption that observations from different matched sets are independent may be unreasonable. For example we can think of a study where many matched sets could come from the same cluster (e.g., hospital or family), or subjects may be followed over time (as in the bison data analysis of Section 5). As discussed by Longford (1994), a possible approach is provided by the theory of generalized estimating equations (GEE) of Liang and Zeger (1986). Fay and Graubard (2001) use an approach based on GEE and offer corrections for inferences when the sample size is small and/or the range of covariates varies with clusters (unbalanced data). Under a certain specific longitudinal case-control design where cases are observed at all time points and controls at some predetermined subset of the time points, Park and Kim (2004) consider estimating equations with independence working correlation. Their approach consists in a marginal specification of the mean that is unconditional on the value of the response (prospective), but they show that with an independence working correlation, their estimating equation is conditionally (retrospectively) unbiased. However, some longitudinal case-control sampling schemes do not fall within the sampling designs that they investigated, and this is the case with our data on plains bison, whose sampling design we now describe.

We consider estimation for the conditional logistic regression model under a case-control design where the number of cases and controls per matched set (stratum) is predetermined before sampling and where strata within clusters might be correlated; this type of design is becoming more popular in biological applications, where GPS technology allows longitudinal follow up of subjects' locations. Our goal is to derive population-averaged inference methods for the conditional logistic regression model parameters under this sampling scheme. More precisely, we propose conditional (retrospective) estimating equations that lead to robust inferences about the model parameters in this context. To derive these estimating equations, we show how we can embed this inference problem into the generalized estimating equation (GEE) framework of Liang and Zeger (1986). This will allow us to benefit from some of the good properties of GEE analysis, such as robustness of inferences to misspecification of the working correlation structure. We explain in Section 2 that, for the model and data under study, using a working correlation matrix other than independence induces certain difficulties with the inferences. While this impedes the capacity to "root-n consistently" estimate correlation matrix parameters, we avoid the problem by choosing an independence working correlation and, thus, can still take advantage of the many robust inference tools derived for GEE, such as estimating equations, sandwich variance estimators and model selection criteria (e.g., Pan (2001)).

The paper is organized as follows. In Section 2 we introduce the model and notation and show how GEE can be derived. We tackle specific inference problems in Section 3. The validity of the approach in finite samples is briefly investigated by simulations in Section 4, and we apply the method to an analysis of a dataset on habitat selection by plains bison in Section 5. Section 6 concludes the paper with a discussion and ideas for further work.

2 Estimating equations for the conditional logistic model

Consider that we have K independent individuals/clusters under study. For the cth individual/cluster, suppose that we observe $S^{(c)}$ strata (matched sets). For each stratum $\mathcal{S}_j^{(c)}$, $c = 1, \dots, K$, $j = 1, \dots, S^{(c)}$, we observe $\boldsymbol{Y}_j^{(c)} = (Y_{j1}^{(c)}, \dots, Y_{jN_j^{(c)}}^{(c)})^\top$, a vector of $\{0, 1\}$ responses and $\boldsymbol{X}_j^{(c)} = (\boldsymbol{x}_{j1}^{(c)}, \dots, \boldsymbol{x}_{jN_j^{(c)}}^{(c)})^\top$, an $N_j^{(c)} \times p$ matrix of covariates. We suppose that the number of responses whose value is 1 in the (c, j) stratum is fixed to $m_j^{(c)}$ by study design and our objective is to estimate the effect of the covariate values on the value of the response. Mathematically, $Y_{ji}^{(c)} \in \{0, 1\}\ \forall c, j, i$ and $\sum_i Y_{ji}^{(c)} = m_j^{(c)}$ with $m_j^{(c)}$ a fixed integer value (that may or may not be the same for all strata). In the bison data of Section 5, clusters are the individual animals followed over a time period and a stratum is a set of one visited location ($Y = 1$) matched with several locations that could potentially have been visited at the same time ($Y = 0$).

2.1 Regression model and likelihood under independence

For simplicity, let us first consider a single stratum and drop the superscript (c) and the subscript j. For each observation i in the stratum, we have a p-vector of covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^\top$.

As in (Hosmer and Lemeshow, 2000, Chapter 7), for a given stratum, we suppose that given the covariates $\boldsymbol{X}$ and a stratum-specific random effect θ, $Y_1, \ldots, Y_N$ are independent Bernoulli random variables with

$$\mathrm{P}[Y_i = 1|\theta, \boldsymbol{X}] = \frac{\exp\left(\theta + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)}{1 + \exp\left(\theta + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)}. \tag{1}$$

It is well known (Hosmer and Lemeshow, 2000, Equation 7.4) that the likelihood, conditional on $\sum_{i=1}^N Y_i = m$, under the model described by (1) is proportional to

$$L_{Full}\left(\boldsymbol{\beta} \middle| \sum_{i=1}^N Y_i = m, \theta\right) = L_{Full}(\boldsymbol{\beta}) = \frac{\exp\left(\sum_{i=1}^N \boldsymbol{\beta}^\top \boldsymbol{x}_i Y_i\right)}{\sum_{l=1}^{\binom{N}{m}} \exp\left(\sum_{i=1}^N \boldsymbol{\beta}^\top \boldsymbol{x}_i v_{li}\right)}, \tag{2}$$

where $\sum_{l=1}^{\binom{N}{m}}$ denotes a sum over all N-vectors $\boldsymbol{v}_l$ such that $v_{lj} \in \{0,1\}$ and $\sum_{j=1}^N v_{lj} = m$. Lemma 2.1 shows a reformulation of $L_{Full}(\boldsymbol{\beta})$ that avoids having to deal with singular covariance matrices for the data in one stratum. Note that the proofs of the lemmas and theorem that follow can all be found in Appendix A.

Lemma 2.1 *Let $\boldsymbol{x}_i^{(-j)} = \boldsymbol{x}_i - \boldsymbol{x}_j$. Then, for any choice of $j \in \{1, \ldots, N\}$, the following likelihood is equal to $L_{Full}(\boldsymbol{\beta})$ given by (2):*

$$L_{(-j)}(\boldsymbol{\beta}) = \frac{\exp\left(\sum_{i \neq j} \boldsymbol{\beta}^\top \boldsymbol{x}_i^{(-j)} Y_i\right)}{\sum_{l=1}^{\binom{N}{m}} \exp\left(\sum_{i \neq j} \boldsymbol{\beta}^\top \boldsymbol{x}_i^{(-j)} v_{li}\right)}. \tag{3}$$

In practical terms, Lemma 2.1 means that since the stratum sums are given, for each stratum we can delete one observation without changing anything to the conditional inferences, provided that we correct the covariates of the remaining observations for that of the deleted observation. Therefore, without loss of generality, from hereon we shall only work with $L(\boldsymbol{\beta}) \equiv L_{(-1)}(\boldsymbol{\beta})$ and we set $\boldsymbol{x}_i^* = \boldsymbol{x}_i - \boldsymbol{x}_1$.

2.2 Likelihood score estimating equations

In most case-control studies the data are obtained *conditionally on the response*, i.e. the cases and controls are obtained and then, subsequently, the covariates $\boldsymbol{X}$ are observed. Prentice and Pyke (1979) have shown that for the logistic model unbiased estimates for $\boldsymbol{\beta}$ can be obtained by treating the data as if they were coming from a prospective study, i.e., one in which the items included in the study are sampled unconditionally on the response Y. (However, if the logistic model (1) includes an intercept term α, then, as shown by Scott and Wild (1986), the maximum likelihood estimator for α would be biased.) In addition, as shown by Park and Kim (2004), using "prospective" estimating equations in longitudinal case-control studies may produce biased estimators. In this paper, we choose to base our inferences on "retrospective" estimating equations, i.e., estimating equations that condition on the fact that the case-control sampling fixes the number of cases within each stratum, as this seems to us like a simpler and more natural way to obtain consistent inferences under our sampling design.

We derive the expressions for the conditional mean and variance for the responses assuming that there is only one stratum (we drop the $(c)/j$ superscript/subscript).

Lemma 2.2 *Let* $\mu_i = E[Y_i | \sum_{j=1}^{N} Y_j = m, \boldsymbol{X}]$ *and* $\mu_{ij} = E[Y_i \cdot Y_j | \sum_{j=1}^{N} Y_j = m, \boldsymbol{X}]$. *Then*

$$\mu_i = \frac{\sum_{l=1}^{\binom{N}{m}} v_{li} \exp\left(\sum_{k=2}^{N} \boldsymbol{\beta}^\top \boldsymbol{x}_k^* v_{lk}\right)}{\sum_{l=1}^{\binom{N}{m}} \exp\left(\sum_{k=2}^{N} \boldsymbol{\beta}^\top \boldsymbol{x}_k^* v_{lk}\right)}, \tag{4}$$

$$\mu_{ij} = \frac{\sum_{l=1}^{\binom{N}{m}} v_{li} v_{lj} \exp\left(\sum_{k=2}^{N} \boldsymbol{\beta}^\top \boldsymbol{x}_k^* v_{lk}\right)}{\sum_{l=1}^{\binom{N}{m}} \exp\left(\sum_{k=2}^{N} \boldsymbol{\beta}^\top \boldsymbol{x}_k^* v_{lk}\right)}. \tag{5}$$

Under the assumption of independence between strata, the likelihood score for $\boldsymbol{\beta}$ can be obtained from (3) and reexpressed using (4):

$$l(\boldsymbol{\beta}) = \sum_{i=2}^{N} \boldsymbol{\beta}^\top \boldsymbol{x}_i^* \boldsymbol{Y}_i - \ln \sum_{l=1}^{\binom{N}{m}} \exp\left(\sum_{h=2}^{N} \boldsymbol{\beta}^\top \boldsymbol{x}_h^* v_{lh}\right) \tag{6}$$

$$\Rightarrow \boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=2}^{N} \left\{ \boldsymbol{x}_i^* Y_i - \frac{\sum_{l=1}^{\binom{N}{m}} v_{li} \boldsymbol{x}_i^* \exp\left(\sum_{h=2}^{N} \boldsymbol{\beta}^\top \boldsymbol{x}_h^* v_{lh}\right)}{\sum_{l=1}^{\binom{N}{m}} \exp\left(\sum_{h=2}^{N} \boldsymbol{\beta}^\top \boldsymbol{x}_h^* v_{lh}\right)} \right\}$$

$$= \sum_{i=2}^{N} \boldsymbol{x}_i^* \{Y_i - \mu_i(\boldsymbol{\beta})\} = \boldsymbol{X}^{*\top} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}, \tag{7}$$

where $\boldsymbol{Y} = (Y_2, \ldots, Y_N)^\top$ and $\boldsymbol{\mu}(\boldsymbol{\beta}) = \{\mu_2(\boldsymbol{\beta}), \ldots, \mu_N(\boldsymbol{\beta})\}^\top$. Under the assumption of no correlation between strata, we have that the global likelihood score equations that are given by

$$\boldsymbol{U}_{indep}(\boldsymbol{\beta}) = \sum_{c=1}^{K} \sum_{j=1}^{S^{(c)}} \boldsymbol{U}_j^{(c)}(\boldsymbol{\beta}) = \mathbf{0}, \tag{8}$$

with $\boldsymbol{U}_j^{(c)}(\boldsymbol{\beta})$ given by (7), would be valid and efficient. However, as we might have strata that are correlated, we want to derive inference methods that are more robust to between-stratum correlation.

It is useful to rewrite (8) in a form more suitable for implementing the GEE. To do so, define $\boldsymbol{Y} = (\boldsymbol{Y}^{(1)\top}, \ldots, \boldsymbol{Y}^{(K)\top})^\top$, with $\boldsymbol{Y}^{(c)\top} = (\boldsymbol{Y}_1^{(c)\top}, \ldots, \boldsymbol{Y}_{S^{(c)}}^{(c)\top})^\top$ for each $c = 1, \ldots, K$ and where $\boldsymbol{Y}_j^{(c)} = (Y_{j2}^{(c)}, \ldots, Y_{jN_j^{(c)}}^{(c)})^\top$ is the $(N_j^{(c)} - 1)$-vector of binary responses (without the first observation) for the observations in the jth stratum of the cth cluster. Let $\boldsymbol{\mu}(\boldsymbol{\beta})$ and $\boldsymbol{\mu}^{(c)}(\boldsymbol{\beta})$ denote $E[\boldsymbol{Y} | \sum Y, \boldsymbol{X}^*]$ and $E[\boldsymbol{Y}^{(c)} | \sum Y, \boldsymbol{X}^*]$, respectively. The following theorem states that (8) can be written in the usual GEE form with working independence correlation structure.

Theorem 2.3 *Let* $\boldsymbol{D}^{(c)} = \partial \boldsymbol{\mu}^{(c)}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}^\top$ *be the* $\{\sum_{j=1}^{S^{(c)}} (N_j^{(c)} - 1)\} \times p$ *matrix of the derivatives of the conditional mean vector for the cth cluster with respect to each element of* $\boldsymbol{\beta}$. *Let* $\boldsymbol{V}^{(c)\ Indep} = Var[\boldsymbol{Y}^{(c)} | \sum Y, \boldsymbol{X}^*]$. *Then*

$$\boldsymbol{U}_{Indep}(\boldsymbol{\beta}) = \sum_{c=1}^{K} \boldsymbol{D}^{(c)\top} \left(\boldsymbol{V}^{(c)\ Indep}\right)^{-1} \{\boldsymbol{Y}^{(c)} - \boldsymbol{\mu}^{(c)}(\boldsymbol{\beta})\}. \tag{9}$$

Theorem 2.3 allows us to easily generalize the inference. Let us now suppose that conditionally on the stratum sums and the covariates, there can be within cluster correlation between responses, but that responses from different clusters are still uncorrelated. Note that this more complex correlation scheme would not occur by simply letting all strata in a cluster share a common random intercept, as conditioning on the stratum sums suppresses this random effect in the conditional likelihood. As a matter of fact, consider a single cluster of S strata with a prospective logistic model with a cluster-level random effect, i.e., given $\Theta = \theta$, suppose that Y_{si}, $s = 1, \dots, S$, $i = 1, \dots, n_s$ are independent Bernoulli's with P$[Y_{si} = y_{si}|\theta, \boldsymbol{x}_{si}]$ given by the logit model with linear predictor $\theta + \boldsymbol{\beta}^\top \boldsymbol{x}_{si}$. Then the conditional likelihood will be proportional to P$[Y_{si} = y_{si},\ s = 1, \dots, S,\ i = 1, \dots, n_s|\boldsymbol{X}, Y_{s\bullet},\ s = 1, \dots, S]$, where $Y_{s\bullet}$ is the sum of the Y's in stratum S. Under these model assumptions, this probability is

$$\int \prod_{s=1}^{S} \frac{\exp\left\{\sum_{i=1}^{n_s} y_{si}(\theta + \boldsymbol{\beta}^\top \boldsymbol{x}_{si})\right\}}{\sum_{l=1}^{\binom{n_s}{m_s}} \exp\left\{\sum_{i=1}^{n_s} v_{li}(\theta + \boldsymbol{\beta}^\top \boldsymbol{x}_{si})\right\}}\, dF_\Theta(\theta),$$

which simplifies to the usual conditional likelihood with independent strata. Thus, in this section we consider cases where correlation may be induced among strata in some more complex manner. For instance, in the analysis of the bison data presented in Section 5, this would mean that if the only source of correlation among strata is a bison-level random effect, then likelihood-based methods should yield valid inferences, while this might not be the case if there are other sources of correlation (e.g., spatio-temporal, as we expect that locations with similar characteristics, i.e. similar covariate values, are more likely to be selected by an animal in consecutive time periods).

2.3 Working correlation matrices

We now describe the difficulties encountered when modeling a correlation matrix in this setup. Let us consider the working correlation matrix for all the responses in one cluster. The elements of the "true" conditional variance matrix of $\boldsymbol{Y}$, say $\boldsymbol{V}$, are of the form

$$\text{cov}(Y_{ji}^{(c)}, Y_{j'i'}^{(c')}) = \begin{cases} 0, & c \neq c' \\ \rho(Y_{ji}^{(c)}, Y_{j'i'}^{(c)})\sqrt{\mu_{ji}^{(c)}(1-\mu_{ji}^{(c)})\mu_{j'i'}^{(c)}(1-\mu_{j'i'}^{(c)})}, & c = c', \end{cases} \tag{10}$$

where $\rho(Y_{ji}^{(c)}, Y_{j'i'}^{(c)})$ represents the conditional correlation between $Y_{ji}^{(c)}$ and $Y_{j'i'}^{(c)}$ given the stratum sums and the covariates. In order to specify a working correlation structure we need to specify values for $\rho(Y_{ji}^{(c)}, Y_{j'i'}^{(c)})$. For instance, if we put $\rho(Y_{ji}^{(c)}, Y_{j'i'}^{(c)}) = 0$ when $j \neq j'$, then we can recover the conditional variance under independent strata, $\boldsymbol{V}^{(c)\ Indep}$. The working correlation matrix must be constructed so that it allows for correlation between responses from different strata that are in the same cluster while still preserving the correlation structure among responses from the same strata as given by (10). This can be done by reexpressing $\boldsymbol{V}^{(c)} = \text{cov}(\boldsymbol{Y}^{(c)})$ as a product of the form $\boldsymbol{V}^{(c)\ \boldsymbol{R}} = \boldsymbol{A}^{(c)1/2}\boldsymbol{R}\boldsymbol{A}^{(c)1/2\top}$, with $\boldsymbol{A}^{(c)1/2}$ such that $\boldsymbol{V}^{(c)Indep} = \boldsymbol{A}^{(c)1/2}\boldsymbol{A}^{(c)1/2\top}$ and by choosing the proper block diagonal matrix $\boldsymbol{R}$.

Unfortunately, inference about the parameters in $\boldsymbol{R}$ is a more complex issue. Indeed, the parameters of $\boldsymbol{R}$ are constrained since the stratum sums are fixed. To see this, we can assume without loss of generality that there are only two strata in a cluster with n observations per stratum and $m = 1$ for both. Suppose that we denote the responses in cluster one $\mathbf{y} = (y_1, \dots, y_n)$ and the responses in cluster two are $\mathbf{y}' = (y'_1, \dots, y'_n)$. Put corr$(y_i, y'_j) = \rho_{ij}$. Then, if we drop one observation from each stratum, say y_1 and y'_1, we obtain

$$\sum_{j=2}^{n} \rho_{ij}\left\{\sqrt{\mu_i(1-\mu_i)\mu'_j(1-\mu'_j)} + \mu_i\mu'_j\right\} = \mu_i - P(y_i = 1, y'_1 = 1) \leq \mu_i$$

and

$$\sum_{i=2}^{n} \rho_{ij}\left\{\sqrt{\mu_i(1-\mu_i)\mu'_j(1-\mu'_j)} + \mu_i\mu'_j\right\} = \mu_j - P(y_1 = 1, y'_j = 1) \leq \mu'_j.$$

So whichever way we choose to model the correlation matrix, we must ensure that the above constraints are satisfied. However, many reasonable choices for the correlation structure cannot match the aforementioned constraints. For instance, it is possible to show that an exchangeable correlation structure does not satisfy the constraints, unless $\rho = 0$. We also found difficult to implement a GEE2 type of approach since all optimization must be done subject to a set of nontrivial constraints.

For these reasons, in the present paper we restrict ourselves to the working independence structure and the GEE (9). This still allows us to fulfill our objective, namely, making valid inferences on $\boldsymbol{\beta}$, as we outline in the next section and observe in the simulation study of Section 4.

3 Inferences

Under working independence and using Theorem 2 from Liang and Zeger (1986), we have that $\sqrt{K}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically multivariate Gaussian with zero mean and variance matrix consistently estimated by

$$\begin{aligned}\hat{\boldsymbol{V}}_G &= \left(\sum_{c=1}^{K} \boldsymbol{D}^{(c)\,\top}\left(\boldsymbol{V}^{(c)\;Indep}\right)^{-1}\boldsymbol{D}^{(c)}\right)^{-1}\left[\sum_{c=1}^{K} \boldsymbol{D}^{(c)\,\top}\left(\boldsymbol{V}^{(c)\;Indep}\right)^{-1}\{\boldsymbol{Y}^{(c)}-\boldsymbol{\mu}^{(c)}(\boldsymbol{\beta})\}\right.\\ &\times \left.\{\boldsymbol{Y}^{(c)}-\boldsymbol{\mu}^{(c)}(\boldsymbol{\beta})\}^{\top}\left(\boldsymbol{V}^{(c)\;Indep}\right)^{-1}\boldsymbol{D}^{(c)}\right]\left(\sum_{c=1}^{K} \boldsymbol{D}^{(c)\,\top}\left(\boldsymbol{V}^{(c)\;Indep}\right)^{-1}\boldsymbol{D}^{(c)}\right)^{-1}, \quad (11)\end{aligned}$$

with $\boldsymbol{\beta}$ replaced by its estimator. We shall refer to $\hat{\boldsymbol{V}}_G$ given by (11) as the robust variance estimator of $\hat{\boldsymbol{\beta}}$, while the variance estimator obtained by inverting the information matrix corresponding to (6) will be referred to as naive variance and will be denoted $\hat{\boldsymbol{V}}_I$. As the simulation study in Section 4 will show, Wald-type inferences based on $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{V}}_G$ obtained under an independence working correlation exhibit very good finite sample properties. A viable alternative to this approach is proposed by Fay and Graubard (2001) when K is small and/or the distribution of the covariates is such that their possible values are not observed in all clusters. Their approach replaces $\hat{\boldsymbol{V}}_G$ with an alternate estimator and the null distribution of the Wald statistic is modified in consequence.

Let us now consider more specifically the issue of model selection, i.e., the choice of the covariates that should be part of the linear predictor $\boldsymbol{\beta}^{\top}\boldsymbol{x}$. For this purpose, we can again use Wald-type methods via stepwise selection procedures with p-values based on the asymptotic normality of $\hat{\boldsymbol{\beta}}$ and the robust variance $\hat{\boldsymbol{V}}_G$ given by (11). However, in some fields of application "information-theoretic" model selection criteria (e.g., AIC, BIC) are often preferred to p-value based model selection (see, for instance, Burnham and Anderson (2002) in the case of biological/ecological applications). Recently, Pan (2001) proposed a quasi-likelihood under independence criterion (QIC) that can be viewed as a generalization of the AIC under GEE, and this with any working correlation structure; this criterion can actually be used for both covariate and working correlation structure selection. In general, in order to derive the QIC, one needs to compute the quasi-likelihood corresponding to the score equations (9) and a working independence matrix $\boldsymbol{V}^{(c)} = \boldsymbol{V}^{(c)\;Indep}$. Our derivations in Section 2 imply that, under assumed independence between strata, the quasi-likelihood function is the same as the log-likelihood function given by (6). Following the notation of Pan (2001), we denote this log-likelihood function $Q[\boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{Y}, I]$, to emphasize that it was computed under a model with mean $\boldsymbol{\mu}(\boldsymbol{\beta})$, data $\boldsymbol{Y}$ and the independence assumption.

Though we recommend to use a working independence correlation structure, the QIC is defined more generally. Consider a working correlation matrix $\boldsymbol{R}$. To emphasize their dependence on $\boldsymbol{R}$, let us denote the solution of the GEE (9) by $\hat{\boldsymbol{\beta}}(\boldsymbol{R})$ and its robust sandwich variance estimator by $\hat{\boldsymbol{V}}_G(\boldsymbol{R})$, and let $\hat{\boldsymbol{\beta}}(\boldsymbol{I})$

be the maximum likelihood estimator of $\boldsymbol{\beta}$ under the independence assumption. We also define

$$\boldsymbol{\Omega}_{\boldsymbol{I}} = \sum_{c=1}^{C} \boldsymbol{D}^{(c)^\top} \boldsymbol{V}^{(c)\ Indep} \boldsymbol{D}^{(c)} = - \left. \frac{\partial^2 Q[\boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{Y}, \boldsymbol{I}]}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\boldsymbol{I})},$$

which is simply the observed information matrix under independence. The QIC criterion is then defined as

$$QIC(\boldsymbol{R}) = -2Q[\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}(\boldsymbol{R})\}, \boldsymbol{Y}, \boldsymbol{I}] + 2\text{trace}\{\boldsymbol{\Omega}_{\boldsymbol{I}} \hat{\boldsymbol{V}}_G(\boldsymbol{R})\}.$$

Notice that if the true correlation structure is independence, then $\hat{\boldsymbol{V}}_G(\boldsymbol{R})$ should be close to the inverse of the information matrix, $\boldsymbol{\Omega}_{\boldsymbol{I}}^{-1}$, and QIC will therefore approach the Akaike information criterion,

$$AIC = -2Q[\boldsymbol{\mu}\{\hat{\boldsymbol{\beta}}(\boldsymbol{I})\}, \boldsymbol{Y}, \boldsymbol{I}] + 2p.$$

Pan (2001) studied the behavior of QIC by simulation. His conclusion is that $QIC(\boldsymbol{I})$ is a good criterion for covariate selection, while $QIC(\boldsymbol{R})$ can be used to choose among correlation structures.

4 Simulations

The purpose of this simulation study is to investigate the validity of inferences based on GEE with working independence with samples of different sizes and with different levels of correlation between strata. We investigate the unbiasedness of regression coefficient and robust variance estimators. Results on the effectiveness of model selection criteria are not reported here in the interest of space, but they are comparable to those obtained by Pan (2001).

We performed simulations under a few variations of two types of models:

Models RI: These are models with cluster-level random intercepts. Specifically, given $\Theta = \theta, Y_1, Y_2, \ldots$ are independent Bernoulli random variables with logit(P$[Y_i = 1|\boldsymbol{x}_i, \Theta = \theta]$) $= \theta + \boldsymbol{\beta}^\top \boldsymbol{x}_i$. We then form each of the S strata of the cluster by randomly sampling the $(Y_i, \boldsymbol{x}_i)$ until we have 1 case and 4 controls. The random effects Θ of each clusters are i.i.d. $N(0, \sigma^2)$.

Models RS: Same as Models RI, but the random intercept is replaced with a random slope in front of x_{i1}.

Remark that under Models RS, the coefficients β_j in front of x_{ij} that we estimate with GEE are not quite the same as the coefficient β_j in the mixed model, as the former are estimates of the marginal (population averaged) effects while the latter are estimates of the conditional (subject specific) effects (McCulloch and Searle, 2000, Chapter 8). More precisely, with two covariates and one random effect, the RS model is the conditional model

$$\text{P}[Y_i = 1|\theta, \boldsymbol{x}_i] = \frac{\exp\{x_{1i}(\theta + \beta_1) + x_{2i}\beta_2\}}{1 + \exp\{x_{1i}(\theta + \beta_1) + x_{2i}\beta_2\}}, \tag{12}$$

where the random effect $\theta \sim N(0, \sigma^2)$ and the covariates $x_{ij} \sim N(0, \sigma_1^2)$ for $j = 1, 2$. The procedure proposed here approximates the marginal probability

$$\text{P}[Y_i = 1|\boldsymbol{x}_i] = \int \frac{\exp\{x_{1i}(\theta + \beta_1) + x_{2i}\beta_2\}}{1 + \exp\{x_{1i}(\theta + \beta_1) + x_{2i}\beta_2\}} \frac{\phi(\theta/\sigma)}{\sigma}\, d\theta, \tag{13}$$

where ϕ is the standard normal density, using

$$\text{P}[Y_i = 1|\boldsymbol{x}_i] = \frac{\exp(x_{i1}\gamma_1 + x_{i2}\gamma_2)}{1 + \exp(x_{i1}\gamma_1 + x_{i2}\gamma_2)}. \tag{14}$$

It is known that (14) can approximate the marginal logistic model with random effects (13) (e.g., Zeger et al., 1988). In addition, while the estimators $\hat{\gamma_1}$ and $\hat{\gamma_2}$ for (14) depend on the particular simulated values of x_1 and x_2, the summaries presented in Table 1 represent averages over 500 different simulated datasets. Therefore, in order to assess the efficiency of the estimates we compare them with a Monte Carlo approximation of $E[\hat{\gamma_1}]$ and $E[\hat{\gamma_2}]$ where the expectation is taken with respect to x_1 and x_2. In Table 1 we present the parameters β_j, $j = 1, 2$ from (12), the Monte Carlo estimates for γ_j obtained using the approach just described and the average GEE estimates produced by solving (9). One should note that, not surprisingly, $|\gamma_j| \leq |\beta_j|$ for $j = 1, 2$ as is the case when parameters for conditional and marginal mixed models are compared. Note that when the true model is not known, one can use the approximation for the values of γ_j given by Zeger et al. (1988, p. 1054), that is based on a Gaussian approximation to the logistic function.

While we have freedom over all the simulation parameters we decided to focus on those we identified as the most important, both with respect to their effect but also as important in practice: 1) the size of the effects, β; 2) the number of clusters, K; 3) the number of strata in each cluster, S; 4) the variance of the random effects, σ^2.

For Models RI, we expect likelihood based methods to perform best, as under these models independence corresponds to the true model. As can be seen from Table 1, this is indeed the case: the coefficient estimates are unbiased and their variance is best estimated by the naive variance estimator, $\hat{\boldsymbol{V}}_I$.

Inference under Models RS yields different results, as under these models the coefficients estimated with GEE are attenuations of the coefficients used in the conditional model. However, we note that on average, coefficient estimates are close to the "theoretical" marginal coefficient γ_j. Also, since under this setup working independence is not the true model, we expect robust methods to outperform their naive counterparts. From Table 1 one can see that an increase in the variance of the random effects makes the attenuation more important. But more importantly, we note that the naive variance estimator is grossly inappropriate in such cases as its average value is nowhere near the sample variance of the estimator. On the other hand, the robust variance estimator seems to perform well even under the Models RS.

5 Application: Analysis of bison data

The objective of this analysis is to investigate the link between the distribution of the plains bison population of Prince Albert National Park ($53°44'$N, $106°40'$W), Saskatchewan (Canada), and the spatial patterns of landscape attributes. Within the park, the landscape is 85% forested, and within this forest matrix are interspersed meadows (10%) and lakes and rivers (5%) (Fortin et al. (2003)). Agricultural lands are found next to the park, and they are potentially accessible to bison. This study of habitat selection was based on the locations of nine female bison equipped with global positioning system (GPS) radio-collars. Bison were located every hour, twice a week from 2 September 2005 to 2 December 2005.

We studied fine-scale habitat selection using the conditional logistic regression model (1), whose linear predictor is referred to as *resource selection function* (RSF). An RSF compares landscape attributes at animal locations to the attributes at random locations (Manly et al. (2002)). Because bison could not travel to every location found within their range during the 1 hour time interval, we used a case-control design in which each visited location ($Y = 1$) was paired to 10 random locations ($Y = 0$). Random locations were drawn within a circular plot (300 m in radius) centered around each observed location. This radius of 300 m captured 85% of the distance moved within 1 hour, and thus should reflect locations that are available to the animal within that time interval. Such case-control designs, where availability is restricted in space, are now commonly used in ecology (Compton et al. (2002), Johnson et al. (2002), Boyce et al. (2003)).

For each bison, strings of 48 locations (one location at every hour for 48 hours) separated by 120 hours were gathered. Eight of the bison yielded fourteen 48-hour strings while one animal only yielded nine strings due to a malfunction of the GPS collar. Since strings of locations were separated by 120 hours and since radio-collared bison were also largely independent from one another (on average, pairs of bison were

Table 1 Average value (sample variance) of regression coefficient and variance estimators under Models RI and RS based on 500 simulations.

Models RI								
$K = 5, S = 40$								
	$\sigma^2 = 0.5$				$\sigma^2 = 2.5$			
Estimator	$\beta_1 = 0.75$		$\beta_2 = 0.5$		$\beta_1 = 0.75$		$\beta_2 = 0.5$	
Coefficients	0.75	(0.016)	0.51	(0.015)	0.76	(0.015)	0.50	(0.014)
Naive var.	0.015		0.014		0.015		0.014	
Robust var.	0.012		0.011		0.012		0.011	
$K = 40, S = 20$								
	$\sigma^2 = 0.5$				$\sigma^2 = 2.5$			
Estimator	$\beta_1 = 0.75$		$\beta_2 = 0.5$		$\beta_1 = 0.75$		$\beta_2 = 0.5$	
Coefficients	0.75	(0.0037)	0.50	(0.0036)	0.75	(0.0036)	0.50	(0.0035)
Naive var.	0.0038		0.0036		0.0038		0.0036	
Robust var.	0.0037		0.0035		0.0037		0.0035	
Models RS								
$K = 5, S = 40$								
	$\sigma^2 = 0.5$				$\sigma^2 = 2.5$			
Parameter (conditional)	$\beta_1 = 0.75$		$\beta_2 = 0.5$		$\beta_1 = 0.75$		$\beta_2 = 0.5$	
MC estimate (marginal)	$\gamma_1 = 0.660$		$\gamma_2 = 0.478$		$\gamma_1 = 0.515$		$\gamma_2 = 0.427$	
Coefficients	0.68	(0.100)	0.48	(0.015)	0.57	(0.295)	0.44	(0.016)
Naive var.	0.015		0.014		0.016		0.014	
Robust var.	0.082		0.012		0.238		0.012	
$K = 40, S = 20$								
	$\sigma^2 = 0.5$				$\sigma^2 = 2.5$			
Parameter (conditional)	$\beta_1 = 0.75$		$\beta_2 = 0.5$		$\beta_1 = 0.75$		$\beta_2 = 0.5$	
MC estimate (marginal)	$\gamma_1 = 0.660$		$\gamma_2 = 0.478$		$\gamma_1 = 0.515$		$\gamma_2 = 0.427$	
Coefficients	0.67	(0.013)	0.48	(0.0033)	0.53	(0.039)	0.43	(0.004)
Naive var.	0.0037		0.0035		0.004		0.003	
Robust var.	0.014		0.0034		0.035		0.003	

located within 100 m from each other 2.7% of the time), we treat the $8 \times 14 + 1 \times 9 = 121$ strings as uncorrelated clusters in the analyses. Based on a classified Landsat TM satellite image in a geographic information system, the study area was separated into seven major habitat classes: agricultural lands (“agric” in Table 2), meadow, conifer stand (“conif” in Table 2), deciduous stand (comprised mostly of aspen), water (including ponds, lakes and rivers), riparian area (“riparian” in Table 2), and road (including hiking trails and gravel roads, all of which are not frequently used by visitors). Habitat classes were coded using 6 dummy variables, with deciduous stands being used as the baseline category. For each observed and random location, we also quantified the proportion of a circular plot (300 m in radius) centered on the locations that was comprised of meadow. Note that habitat classes and proportion of meadow are variable from cluster to cluster but the large number of clusters ensures that covariate effect estimates are based on many observations.

Table 2 Summary of the Conditional logistic regression models (RSF) fitted to the bison data. "PROPMD300" stands for the proportion of meadows within a 300-m radius, while "meadint" represents the meadow×PROPMD300 interaction.

<table>
<tr><th colspan="6">Complete model (AIC=24122.984, QIC=24143.07)</th></tr>
<tr><th>Variable</th><th>$\hat{\beta}$</th><th>$\sqrt{\hat{\boldsymbol{V}}_I}$</th><th>Naive p-value</th><th>$\sqrt{\hat{\boldsymbol{V}}_G}$</th><th>Robust p-value</th></tr>
<tr><td>PROPMD300</td><td>0.644</td><td>0.252</td><td>0.0106</td><td>0.336</td><td>0.0558</td></tr>
<tr><td>meadow</td><td>1.464</td><td>0.065</td><td>&lt;.0001</td><td>0.118</td><td>&lt;.0001</td></tr>
<tr><td>meadint</td><td>-0.904</td><td>0.246</td><td>0.0002</td><td>0.424</td><td>0.0328</td></tr>
<tr><td>conif</td><td>-0.539</td><td>0.061</td><td>&lt;.0001</td><td>0.100</td><td>&lt;.0001</td></tr>
<tr><td>water</td><td>-0.164</td><td>0.093</td><td>0.0793</td><td>0.159</td><td>0.3031</td></tr>
<tr><td>riparian</td><td>-0.566</td><td>0.281</td><td>0.0440</td><td>0.240</td><td>0.0184</td></tr>
<tr><td>agric</td><td>1.213</td><td>0.413</td><td>0.0033</td><td>0.092</td><td>&lt;.0001</td></tr>
<tr><td>road</td><td>0.875</td><td>0.106</td><td>&lt;.0001</td><td>0.163</td><td>&lt;.0001</td></tr>
<tr><th colspan="6">Model without water (AIC=24124.162, QIC=24140.40)</th></tr>
<tr><th>Variable</th><th>$\hat{\beta}$</th><th>$\sqrt{\hat{\boldsymbol{V}}_I}$</th><th>Naive p-value</th><th>$\sqrt{\hat{\boldsymbol{V}}_G}$</th><th>Robust p-value</th></tr>
<tr><td>PROPMD300</td><td>0.581</td><td>0.250</td><td>0.0198</td><td>0.32380</td><td>0.0726</td></tr>
<tr><td>meadow</td><td>1.471</td><td>0.065</td><td>&lt;.0001</td><td>0.11718</td><td>&lt;.0001</td></tr>
<tr><td>meadint</td><td>-0.862</td><td>0.245</td><td>0.0004</td><td>0.42149</td><td>0.0408</td></tr>
<tr><td>conif</td><td>-0.525</td><td>0.061</td><td>&lt;.0001</td><td>0.09861</td><td>&lt;.0001</td></tr>
<tr><td>riparian</td><td>-0.549</td><td>0.281</td><td>0.0505</td><td>0.23879</td><td>0.0215</td></tr>
<tr><td>agric</td><td>1.223</td><td>0.413</td><td>0.0031</td><td>0.08855</td><td>&lt;.0001</td></tr>
<tr><td>road</td><td>0.885</td><td>0.106</td><td>&lt;.0001</td><td>0.16204</td><td>&lt;.0001</td></tr>
</table>

RSF were built using matched case-control logistic regression. Models were then compared on the basis of their AIC and QIC, as well as based on the p-values associated with naive and robust Wald tests.

Results in Table 2 revealed that bison distribution was linked to multiple landscape attributes. The probability of bison occurrence increased in areas surrounded (i.e., within 300 m) by a large proportion of meadows. Relative to deciduous stands, bison selected meadows, agricultural lands, and roads. The strength of meadow selection decreased, however, as the proportion of meadows increased within a radius of 300 m. Also, bison avoided conifer stands and riparian areas. Interpretation of the effect of water distribution on bison spatial patterns varied depending on whether or not the correlation in the successive bison locations was accounted for in inferences. Indeed, the naive Wald test associated with the water coefficient had a p-value of 0.08, whereas the robust equivalent had p-value of 0.30. Also, AIC indicated that an RSF including a negative coefficient for water seems preferable to a model without water, while QIC points towards the model that excludes water. Model comparisons thus illustrate that the consideration of the correlation in the data might lead to different interpretation of factors influencing the probability of animal occurrence in complex landscapes.

6 Discussion

We have considered conditional logistic regression when matched sets can be correlated. We have shown how to rewrite the likelihood score equations in a fashion that allowed for an easy extension to GEE. We have demonstrated that such an approach yields valid inferences. We have also illustrated how, under complex correlation schemes, robust inferences based on GEE with working independence might lead to conclusions that are more accurate than what one would obtain using likelihood-based methods. This fact transpired in the bison data analysis, where the differences between the robust and naive analyses suggest an underlying (possibly spatio-temporal) correlation structure among the matched sets.

In case of small samples, the present treatment could potentially be improved upon by using the modified variance estimator of Fay and Graubard (2001). Also of interest in this case would be the specification of working correlation structures other than independence. In the latter case, further investigation is needed to overcome the difficulties resulting from the constraints as discussed in Section 2.3.

Zeger et al. (1988) propose a GEE-based approach to subject specific inferences. It would be interesting to see if their method could be adapted to our context and to compare the inference thereby obtained to that obtained with maximum likelihood or partial quasi-likelihood.

Acknowledgements We wish to thank the Associate Editor and two anonymous referees for their constructive comments. We are also grateful to Hawthorne Beyer for all his work with the GIS. Dan Frandsen, Lloyd O'Brodovich, and to other staff from Parks Canada, together with Marie-Eve Fortin who helped the acquisition of bison locations. Funding was provided by the Natural Sciences and Engineering Research Council of Canada (all three authors), the Fonds québécois sur la nature et les technologies (TD and DF) and by the Canada Foundation for Innovation, Recherche Québec and Parks Canada Ecological Integrity Innovation Fund (DF).

References

Arbogast, P. G. and Lin, D. Y. (2004). Goodness-of-fit methods for matched case-control studies. *Canadian Journal of Statistics* **32**, 373–386.

Boyce, M. S., Mao, J. S., Merrill, E. H., Fortin, D., Turner, M. G., Fryxell, J. M. and Turchin, P. (2003). Scale and heterogeneity in habitat selection by elk in yellowstone national park. *Ecoscience* **10**, 321–332.

Breslow, N. E. and Day, W. (1980). *Statistical Methods in Cancer research: the analysis of case-control studies*. Oxford University Press, Oxford, UK.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach, 2nd Edition*. Springer-Verlag, New York, NY, USA.

Compton, B. W., Rhymer, J. M. and McCollough, M. (2002). Habitat selection by wood turtles (clemmys insculpta): an application of paired logistic regression. *Ecology* **83**, 833–843.

Fay, M. P. and Graubard, B. I. (2001). Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics* **57**, 1198–1206.

Fay, M. P., Graubard, B. I., Freedman, L. S. and Midthune, D. N. (1998). Conditional logistic regression with sandwich estimators: Application to a meta-analysis. *Biometrics* **54**, 195–208.

Fortin, D., Fryxell, J. M., O'Brodovich, L. and Frandsen, D. (2003). Foraging ecology of bison at the landscape and plant community levels: the applicability of energy maximization principles. *Oecologia* **134**, 219–227.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression, 2nd Edition*. Wiley, New York, NY, USA.

Johnson, C. J., Parker, K. L., Heard, D. C. and Gillingham, M. P. (2002). Movement parameters of ungulates and scale-specific responses to the environment. *Journal of Animal Ecology* **71**, 225–235.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Longford, N. T. (1994). Logistic regression with random coefficients. *Computational Statistics and Data Analysis* **17**, 1–15.

Manly, B. F. J., McDonald, L. L., Thomas, D. L., McDonald, T. L. and Erickson, W. P. (2002). *Resource Selection by Animals: Statistical design and analysis for field studies*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

McCulloch, C. E. and Searle, R. S. (2000). *Generalized, linear and mixed models*. Wiley, New York, NY, USA.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120–125.

Park, E. and Kim, Y. (2004). Analysis of longitudinal data in case-control studies. *Biometrika* **91**, 321–330.

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

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

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

A Proofs

Proof of Lemma 2.1 It is easily seen that for any $j = 1, \ldots, N$,

$$L_{(-j)}(\boldsymbol{\beta}) \quad = \quad \frac{\exp\left\{\sum_{i=1}^N \boldsymbol{\beta}^\top \boldsymbol{x}_i Y_i - \left(\sum_{i=1}^N Y_i\right) \boldsymbol{\beta}^\top \boldsymbol{x}_j\right\}}{\sum_{l=1}^{\binom{N}{m}} \exp\left\{\sum_{i=1}^N \boldsymbol{\beta}^\top \boldsymbol{x}_i v_{li} - \left(\sum_{i=1}^N v_{li}\right) \boldsymbol{\beta}^\top \boldsymbol{x}_j\right\}}.$$

But $\sum_{i=1}^N Y_i = \sum_{i=1}^N v_{li} = m$, which means that $L_{(-j)}(\boldsymbol{\beta}) = L_{Full}(\boldsymbol{\beta}) \times \exp(-m\boldsymbol{\beta}^\top \boldsymbol{x}_j)/\exp(m\boldsymbol{\beta}^\top \boldsymbol{x}_j) = L_{Full}(\boldsymbol{\beta})$.

Proof of Lemma 2.2 Let $\sum_{l=1}^{\binom{N-1}{m-1}}$ denote a sum over all N-vectors $\tilde{\boldsymbol{v}}_l^{(i)}$ such that $\tilde{v}_{lj}^{(i)} \in \{0, 1\}$, $\sum_j \tilde{v}_{lj}^{(i)} = m$ and $\tilde{v}_{li}^{(i)} = 1$. Because the Y_i are binary, we have that

$$\begin{aligned} \mu_i \quad &= \quad \mathrm{P}\left(Y_i = 1 \,\middle|\, \sum_{j=1}^N Y_j = m, \boldsymbol{X}\right) \\ &= \quad \frac{\frac{\exp(\theta+\boldsymbol{\beta}^\top \boldsymbol{x}_i^\top)}{1+\exp(\theta+\boldsymbol{\beta}^\top \boldsymbol{x}_i)} \sum_{l=1}^{\binom{N-1}{m-1}} \prod_{h\neq i} \frac{\exp\{(\theta+\boldsymbol{\beta}^\top \boldsymbol{x}_h)\tilde{v}_{lh}^{(i)}\}}{1+\exp\{(\theta+\boldsymbol{\beta}^\top \boldsymbol{x}_h)\tilde{v}_{lh}(i)\}}}{\sum_{l=1}^{\binom{N}{m}} \prod_{i=1}^N \left[\exp(\theta + \boldsymbol{\beta}^\top \boldsymbol{x}_i v_{li})/\{1 + \exp(\theta + \boldsymbol{\beta}^\top \boldsymbol{x}_i)\}\right]} \\ &= \quad \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i} \sum_{l=1}^{\binom{N-1}{m-1}} \exp\left(\sum_{\substack{h=1 \\ h\neq i}}^N \boldsymbol{\beta}^\top \boldsymbol{x}_h \tilde{v}_{lh}(i)\right)}{\sum_{l=1}^{\binom{N}{m}} \exp\left(\sum_{h=1}^N \boldsymbol{\beta}^\top \boldsymbol{x}_h v_{lh}\right)}. \end{aligned} \tag{15}$$

After applying Lemma 2.1, the denominators of (4) and (15) are the same. To show that the numerators are equal, notice that in (4), the only terms in the sum over l are those with $v_{li} = 1$. This leaves us with a sum over $\binom{N-1}{m-1}$ N-vectors with 1 in position i and whose elements in other positions add up to $m-1$, which, after applying Lemma 2.1, is exactly the numerator of (15). The proof of (5) is similar to the proof of (4) since $E\left[Y_i \cdot Y_j \,\middle|\, \sum_{k=1}^N Y_k = m, \boldsymbol{X}\right] = \mathrm{P}\left(Y_i = 1, Y_j = 1 \,\middle|\, \sum_{k=1}^N Y_k = m, \boldsymbol{X}\right)$ and the latter probability can be expanded as in (15).

Proof of Theorem 2.3 We have to show that $\boldsymbol{D}^{(c)\,\top}\left(\boldsymbol{V}^{(c)\;Indep}\right)^{-1} = \boldsymbol{X}^{*(c)\,\top}$ or, equivalently, that $\boldsymbol{X}^{*(c)\,\top}\boldsymbol{V}^{(c)\;Indep} = \boldsymbol{D}^{(c)\,\top}$. For ease of notation, we drop the superscript (c) for the remainder of this proof. Because two responses in a same stratum are correlated and responses from different strata are uncorrelated, $\boldsymbol{V}^{Indep}$ will be block diagonal. The element in position (i, j) of $\boldsymbol{V}^{Indep}$ will therefore be

$$V_{ij} = \begin{cases} 0, & i \text{ and } j \text{ from different strata} \\ \mu_i(1-\mu_i), & i = j \\ \mu_{ij} - \mu_i\mu_j, & i \neq j, i \text{ and } j \text{ from same stratum,} \end{cases}$$

where the formulas for μ_i and μ_{ij} are given by equations (4) and (5), respectively. Let us now calculate the element in position (i, j) of $\boldsymbol{X}^{*\top}\boldsymbol{V}^{Indep}$. Since $\boldsymbol{V}^{Indep}$ is block diagonal, this element is given by $\sum_l x^*_{li}V_{lj}$, where the sum is over all columns of $\boldsymbol{X}^{*\top}$ (rows of $\boldsymbol{X}^*$) and all rows of $\boldsymbol{V}^{Indep}$ corresponding to observations from the same stratum as that of the element corresponding to column j of $\boldsymbol{V}^{Indep}$. We then get

$$\begin{aligned} \sum_l x^*_{li}V_{lj} &= \sum_l x^*_{li}(\mu_{lj} - \mu_l\mu_j) \\ &= \sum_l x^*_{li}\left[\sum_h v_{hl}v_{hj}w_h(\boldsymbol{\beta}) - \left\{\sum_k v_{kl}w_k(\boldsymbol{\beta})\right\}\left\{\sum_g v_{gj}w_g(\boldsymbol{\beta})\right\}\right], \end{aligned} \tag{16}$$

where $\sum_k$ and $\sum_g$ denote the sum over all $\binom{N}{m}$ possible vectors comprised of $N-m$ zeros and m ones, and where $w_h(\boldsymbol{\beta}) = \exp\left(\sum_l \boldsymbol{\beta}^\top \boldsymbol{x}^*_l v_{hl}\right) / \sum_k \exp\left(\sum_l \boldsymbol{\beta}^\top \boldsymbol{x}^*_l v_{kl}\right)$. Now all that is left to do is to explicitly compute the (i, j)th element of $\boldsymbol{D}^\top$:

$$\begin{aligned} \boldsymbol{D}^\top &= \frac{\partial \mu_j(\boldsymbol{\beta})}{\partial \beta_i} = \frac{\partial}{\partial \beta_i} \frac{\sum_k v_{kj} \exp\left(\sum_l \boldsymbol{\beta}^\top \boldsymbol{x}^*_l v_{kl}\right)}{\sum_k \exp\left(\sum_l \boldsymbol{\beta}^\top \boldsymbol{x}^*_l v_{kl}\right)} \\ &= \left\{\left(\sum_g e^{\sum_l \boldsymbol{\beta}^\top \boldsymbol{x}^*_l v_{gl}}\right)\left(\sum_k v_{kj}\sum_l v_{kl}x^*_{li}e^{\sum_l \boldsymbol{\beta}^\top \boldsymbol{x}^*_l v_{kl}}\right)\right. \\ &\quad \left. - \left(\sum_k v_{kj}e^{\sum_l \boldsymbol{\beta}^\top \boldsymbol{x}^*_l v_{kl}}\right)\left(\sum_g\sum_l v_{gl}x^*_{li}e^{\sum_l \boldsymbol{\beta}^\top \boldsymbol{x}^*_l v_{gl}}\right)\right\}\left(\sum_k e^{\sum_l \boldsymbol{\beta}^\top \boldsymbol{x}^*_l v_{kl}}\right)^{-2} \\ &= \sum_k v_{kj}\sum_l v_{kl}x^*_{li}w_k(\boldsymbol{\beta}) - \left(\sum_k v_{kj}w_k(\boldsymbol{\beta})\right)\left(\sum_g\sum_l v_{gl}x^*_{li}w_g(\boldsymbol{\beta})\right) \\ &= \sum_l x^*_{li}\left[\sum_k v_{kj}v_{kl}w_k(\boldsymbol{\beta}) - \left\{\sum_k v_{kj}w_k(\boldsymbol{\beta})\right\}\left\{\sum_g v_{gj}w_g(\boldsymbol{\beta})\right\}\right], \end{aligned}$$

which is exactly equation (16).