A generalized estimating equation approach to longitudinal conditional logistic regression

by
RADU CRAIU
Department of Statistics, University of Toronto
100 St. George Street,
Toronto, Ontario, M4S 3G3, Canada
craiu@utstat.toronto.edu

&

THIERRY DUCHESNE
Département de mathématiques et de statistique,
Université Laval, Québec
Québec, G1K 7P4, Canada
duchesne@mat.ulaval.ca

Technical Report No. 0509 August 25, 2005

TECHNICAL REPORT SERIES

University of Toronto

Department of Statistics

A generalized estimating equation approach to longitudinal conditional logistic regression

BY RADU V. CRAIU

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

craiu@utstat.toronto.edu

AND THIERRY DUCHESNE

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

duchesne@mat.ulaval.ca

SUMMARY

The method of generalized estimating equations developed by Liang and Zeger (1986, *Biometrika*, **73**, 13-22) has been used increasingly for modelling repeated measurements of a categorical variable. In this paper we develop an approach based on generalized estimating equations for conditional logistic regression. The repeated measurements data consist of strata grouped in clusters. Within each stratum the number of cases is fixed and we assume that all the strata within a given cluster are correlated. The dependence structure is complicated in that the correlations between observations within one stratum have a known form while the correlations between observations in different strata are unknown so a working correlation approach is needed. We implement the quasi-likelihood under independence criterion (QIC) of Pan (2001, *Biometrics*, **57**, 120-125) to select from a number of possible models.

Some key words: Case-control logistic regression, conditional logistic regression, generalized estimating equations, quasi-likelihood, quasi-likelihood criterion, working correlation.

1. Introduction

In many scientific investigations researchers are interested in exploring possible relationships between the characteristics and environment of an individual and a categorical response. Particularly in the case of a binary response for which one of the values is "rare", 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 models has been intensive in biostatistics and in particular in epidemiology. Many such studies require the close matching of control and case subjects. 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. Recently, Arbogast and Lin (2004) have proposed goodness-of-fit tests for the elements of matched case-control logistic regression model.

However, in many cases the control variables between stratum are highly correlated (a stratum is defined as a set of matched subjects containing a fixed number of controls and cases) and the specific structure of the dependence is not known. As discussed by Longford (1994), a possible approach is provided by the theory of generalized estimating equations (GEE) of Liang & Zeger (1986). Miller, Davis & Landis (1993) consider GEE for multinomial data. In the context where we have one case per stratum, their techniques would be applicable; however, if the number of cases per stratum is fixed to a number higher than one, then the joint distribution of the responses is not a multinomial. Park & Kim (2004) have considered estimating equations for longitudinal case-control logistic regression. Their approach consists in a marginal specification of the mean that does not consider fully the longitudinal structure of the data.

In this paper we consider estimation for the conditional logistic regression model based on unbiased conditional GEE (i.e., we use the mean conditional on the stratum sums in the estimating equations). The GEE is a fitting approach as we allow for a (unknown) dependence structure within clusters of strata. The paper is organized as follows. In §2, we introduce the notation and the model and we derive the maximum likelihood score equations under the assumption of independence between clusters. We then derive GEE to allow for robust inference in the presence of within-cluster correlation in §3. Issues relating to practical implementation, such as the choice of a working correlation structure or the choice of a best regression model are discussed in §4 where the QIC, an analogue of the AIC for GEE (Pan, 2001), for this model is derived.

2. Likelihood score estimating equations for conditional logistic regression

Consider that we have K independent individuals/clusters under study. For the cth individual/cluster, suppose that we observe $S^{(c)}$ strata. For each stratum $\mathcal{S}_j^{(c)}$, $c = 1, \ldots, K$, $j = 1, \ldots, S^{(c)}$, we observe $\boldsymbol{Y}_j^{(c)} = (Y_{j1}^{(c)}, \ldots, Y_{jN_j^{(c)}}^{(c)})^T$, a vector of $\{0, 1\}$ responses and $\boldsymbol{X}_j^{(c)} = (\boldsymbol{x}_{j1}^{(c)\,T}, \ldots, \boldsymbol{x}_{jN_j^{(c)}}^{(c)})^T$, a matrix of covariates. We suppose that the number of responses whose value is 1 in the (c, j)th 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).

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 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T$.

As in Hosmer & Lemeshow (1989), 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

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

It is well known (Hosmer and Lemeshow, 1989, equation (7.3)) that the likelihood, conditional on $\sum_{i=1}^N Y_i = m$, under the model described by (1) is given by

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

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 $\sum_{j=1}^N v_{lj} = m$.

A result that will be useful when we consider working correlation structures in §3 is the following:

Lemma 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):*

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

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

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

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

Therefore without loss of generality, from hereon we shall only work with

$$L(\boldsymbol{\beta}) \equiv L_{(-1)}(\boldsymbol{\beta}) = \frac{\exp\left\{\sum_{i=2}^N \boldsymbol{\beta}^T \boldsymbol{x}_i^* Y_i\right\}}{\sum_{l=1}^{\binom{N}{m}} \exp\left\{\sum_{i=2}^N \boldsymbol{\beta}^T \boldsymbol{x}_i^* v_{li}\right\}}, \tag{4}$$

where $\boldsymbol{x}_i^* = \boldsymbol{x}_i - \boldsymbol{x}_1$.

2.2. Likelihood score estimating equations

We now derive the likelihood score equations under the assumption that conditionally on the stratum sums and the covariates, the observations from different strata are uncorrelated, i.e., $\mathrm{cov}(Y_{j,i}^{(c)}, Y_{j',i'}^{(c')} | \sum_i Y_{j,i}^{(c)} = m_j^{(c)}, \sum_i Y_{j',i}^{(c')} = m_{j'}^{(c')}, \boldsymbol{X}^*) = 0$, as long as $(c, j) \neq (c', j')$. Under this assumption, we simply need to derive the log-likelihood and likelihood score function for the single strata and sum them up over all strata to obtain the global likelihood score equation. We therefore again consider a single stratum and drop the $(c)/j$ superscript/subscript. We first need an expression for the conditional means.

Lemma 2 *Let* $\mu_i = E[Y_i | \sum_{j=1}^N Y_j = m, \boldsymbol{X}^*]$. *Then*

$$\mu_i = \frac{\sum_{l=1}^{\binom{N}{m}} v_{li} \exp\left\{\sum_{k=2}^N \boldsymbol{\beta}^T \boldsymbol{x}_k^* v_{lk}\right\}}{\sum_{l=1}^{\binom{N}{m}} \exp\left\{\sum_{k=2}^N \boldsymbol{\beta}^T \boldsymbol{x}_k^* v_{lk}\right\}}. \tag{5}$$

Proof. 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

$$\begin{aligned}
\mu_i &= P\left(Y_i = 1 \middle| \sum_{j=1}^N Y_j = m, \boldsymbol{X}^*\right) \\
&= \frac{\frac{\exp(\theta + \boldsymbol{\beta}^T \boldsymbol{x}_i^{*\,T})}{1+\exp(\theta+\boldsymbol{\beta}^T \boldsymbol{x}_i^*)} \sum_{l=1}^{\binom{N-1}{m-1}} \prod_{h \neq i} \frac{\exp\{(\theta+\boldsymbol{\beta}^T \boldsymbol{x}_h^*)\tilde{v}_{lh}^{(i)}\}}{1+\exp\{(\theta+\boldsymbol{\beta}^T \boldsymbol{x}_h^*)\tilde{v}_{lh}(i)\}}}{\sum_{l=1}^{\binom{N}{m}} \prod_{i=1}^N \left[\exp(\theta + \boldsymbol{\beta}^T \boldsymbol{x}_i^* v_{li}) / \{1 + \exp(\theta + \boldsymbol{\beta}^T \boldsymbol{x}_i^*)\}\right]} \\
&= \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}_i^*} \sum_{l=1}^{\binom{N-1}{m-1}} \exp\left(\sum_{\substack{h=1 \\ h \neq i}}^N \boldsymbol{\beta}^T \boldsymbol{x}_h^* \tilde{v}_{lh}(i)\right)}{\sum_{l=1}^{\binom{N}{m}} \exp\left(\sum_{h=1}^N \boldsymbol{\beta}^T \boldsymbol{x}_h^* v_{lh}\right)}.
\end{aligned} \tag{6}$$

After applying Lemma 1, the denominators of (5) and (6) are the same. To show that the numerators are equal, notice that in (5), 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 1, is exactly the numerator of (6). □

Since the conditional covariances between two responses in a same stratum will also be needed, the following expectation will be useful:

Lemma 3 *Let* $\mu_{ij} = E[Y_i \cdot Y_j | \sum_{j=1}^N Y_j = m, \boldsymbol{X}^*]$. *Then*

$$\mu_{ij} = \frac{\sum_{l=1}^{\binom{N}{m}} v_{li} v_{lj} \exp\left\{\sum_{k=2}^N \boldsymbol{\beta}^T \boldsymbol{x}_k^* v_{lk}\right\}}{\sum_{l=1}^{\binom{N}{m}} \exp\left\{\sum_{k=2}^N \boldsymbol{\beta}^T \boldsymbol{x}_k^* v_{lk}\right\}}. \tag{7}$$

Proof. The proof is similar to the proof of (5) since

$$E\left[Y_i \cdot Y_j \middle| \sum_{k=1}^N Y_k = m, \boldsymbol{X}^*\right] = P\left(Y_i = 1, Y_j = 1 \middle| \sum_{k=1}^N Y_k = m, \boldsymbol{X}^*\right)$$

and the right hand term can be expanded as in (6). □

We can now write the likelihood score for $\boldsymbol{\beta}$:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=2}^N \boldsymbol{\beta}^T \boldsymbol{x}_i^* \boldsymbol{Y}_i - \ln \sum_{l=1}^{\binom{N}{m}} \exp\left(\sum_{h=2}^N \boldsymbol{\beta}^T \boldsymbol{x}_h^* v_{lh}\right) \\ \Rightarrow \boldsymbol{U}(\boldsymbol{\beta}) &= \sum_{i=2}^N \left\{ \boldsymbol{x}_i^* Y_i - \frac{\sum_{l=1}^{\binom{N}{m}} v_{li} \boldsymbol{x}_i^* \exp\left(\sum_{h=2}^N \boldsymbol{\beta}^T \boldsymbol{x}_h^* v_{lh}\right)}{\sum_{l=1}^{\binom{N}{m}} \exp\left(\sum_{h=2}^N \boldsymbol{\beta}^T \boldsymbol{x}_h^* v_{lh}\right)} \right\} \\ &= \sum_{i=2}^N \boldsymbol{x}_i^* \{Y_i - \mu_i(\boldsymbol{\beta})\} = \boldsymbol{X}^{*T} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}, \end{aligned} \tag{8}$$

where $\mu_i(\boldsymbol{\beta}) \equiv \mu_i$ emphasizes the fact that the conditional mean of Y_i depends on $\boldsymbol{\beta}$ and where $\boldsymbol{\mu}(\boldsymbol{\beta}) = \{\mu_2(\boldsymbol{\beta}), \ldots, \mu_N(\boldsymbol{\beta})\}^T$. Under the assumption of no correlation between strata, we have that the global likelihood score equations that are given by

$$\boldsymbol{U}_{indep}(\boldsymbol{\beta}) = \sum_{c=1}^K \sum_{j=1}^{S^{(c)}} \boldsymbol{U}_j^{(c)}(\boldsymbol{\beta}) = \boldsymbol{0}, \tag{9}$$

with $\boldsymbol{U}_j^{(c)}(\boldsymbol{\beta})$ given by (8), would be valid and efficient. However, we might have strata that are correlated. We will tackle this problem via the GEE approach.

Hence it is useful to rewrite (9) in a form more suitable for implementing the GEE. To do so, first define $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_K^T)^T$, with $\boldsymbol{Y}^{(c)\,T} = (\boldsymbol{Y}_1^{(c)\,T}, \ldots, \boldsymbol{Y}_{S^{(c)}}^{(c)\,T})^T$ for each $c = 1, \ldots, K$ and where $\boldsymbol{Y}_j^{(c)} = (Y_{j2}^{(c)}, \ldots, Y_{jN_j^{(c)}}^{(c)})^T$ 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 $E[\mathbf{Y} \,|\, \sum Y, \boldsymbol{X}^*]$ and $E[\boldsymbol{Y}^{(c)} | \sum Y, \boldsymbol{X}^*]$, respectively.

Proposition 1 *Let $\boldsymbol{D}^{(c)} = \partial \boldsymbol{\mu}^{(c)}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}^T$ 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 $\boldsymbol{V}^{(c)\,Indep} = Var[\mathbf{Y}^{(c)} | \sum Y, \boldsymbol{X}^*]$. Then*

$$\boldsymbol{U}_{Indep}(\boldsymbol{\beta}) = \sum_{c=1}^{K} \boldsymbol{D}^{(c)\,T} \left(\boldsymbol{V}^{(c)\,Indep} \right)^{-1} \{ \boldsymbol{Y}^{(c)} - \boldsymbol{\mu}^{(c)}(\boldsymbol{\beta}) \}. \tag{10}$$

Proof. We have to show that $\boldsymbol{D}^{(c)\,T} \left(\boldsymbol{V}^{(c)\,Indep} \right)^{-1} = \boldsymbol{X}^{*(c)\,T}$ or, equivalently, that $\boldsymbol{X}^{*(c)\,T} \boldsymbol{V}^{(c)\,Indep} = \boldsymbol{D}^{(c)\,T}$. 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 (5) and (7), respectively. Let us now calculate the element in position (i, j) of $\boldsymbol{X}^{*\,T} \boldsymbol{V}^{Indep}$. 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 $\boldsymbol{X}^{*T}$ (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

$$\begin{aligned}
\sum_l x_{li}^* V_{lj} &= \sum_l x_{li}^*(\mu_{lj} - \mu_l\mu_j) \\
&= \sum_l x_{li}^* \left[\sum_h v_{hl}v_{hj}w_h(\boldsymbol{\beta}) - \left\{\sum_k v_{kl}w_k(\boldsymbol{\beta})\right\}\left\{\sum_g v_{gj}w_g(\boldsymbol{\beta})\right\}\right], \quad (11)
\end{aligned}$$

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

$$w_h(\boldsymbol{\beta}) = \frac{\exp\left(\sum_l \boldsymbol{\beta}^T \boldsymbol{x}_l^* v_{hl}\right)}{\sum_k \exp\left(\sum_l \boldsymbol{\beta}^T \boldsymbol{x}_l^* v_{kl}\right)}.$$

Now all that is left to do is to explicitly compute the (i, j)th element of $\boldsymbol{D}^T$:

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

which is exactly equation (11). □

3. Generalized estimating equations

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. In this case, $\hat{\boldsymbol{\beta}}$ obtained by solving $\boldsymbol{U}_{Indep}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ would still be a consistent estimator of $\boldsymbol{\beta}$, but its asymptotic variance would not be given by the inverse of the corresponding information matrix anymore. Moreover, $\hat{\boldsymbol{\beta}}$ might also suffer from a lack of efficiency. We therefore use GEE to make inferences about $\boldsymbol{\beta}$.

The elements of the true conditional variance matrix of $\boldsymbol{Y}$, say $\boldsymbol{V}$, are of the form

$$\operatorname{cov}(Y_{ji}^{(c)}, Y_{j'i'}^{(c')}) = \begin{cases} 0, & c \neq c' \\ \mu_{ji,ji'}^{(c)} - \mu_{ji}^{(c)}\mu_{ji'}^{(c)}, & c = c', j = j' \\ \rho(Y_{ji}^{(c)}, Y_{j'i'}^{(c)})\sqrt{\mu_{ji}^{(c)}(1-\mu_{ji}^{(c)})\mu_{j'i'}^{(c)}(1-\mu_{j'i'}^{(c)})}, & c = c', j \neq j', \end{cases} \quad (12)$$

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.

To use GEE, we first have to specify a working correlation structure, i.e., specify the values for $\rho(Y_{ji}^{(c)}, Y_{j'i'}^{(c)})$. For instance, if we put $\rho(Y_{ji}^{(c)}, Y_{j'i'}^{(c)}) = 0$, then we obtain the conditional variance under independent strata $\boldsymbol{V}^{Indep}$ from §2. To generate correlation structures, let $\boldsymbol{A}^{(c)} = \boldsymbol{V}^{(c)\,Indep}$ and let $\boldsymbol{A}^{(c)\,1/2}$ be such that $[\boldsymbol{A}^{(c)\,1/2}]^T\boldsymbol{A}^{(c)\,1/2} = \boldsymbol{V}^{(c)\,Indep}$.

The block diagonal structure of the matrix $\boldsymbol{V}^{(c)\,Indep}$ under the independence assumption makes it necessary to detail the way one can incorporate directly a general working correlation matrix as is customary done in GEE analyses. Since $\boldsymbol{A}^{(c)}$ is not diagonal it results that $[\{\boldsymbol{A}^{(c)}\}^{1/2}]^T \neq \{\boldsymbol{A}^{(c)}\}^{1/2}$. 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 (12). Let us drop the superscript (c) for a moment and let $\boldsymbol{V}^{Indep}$ be the variance-covariance matrix under independance. Then $\boldsymbol{V}^{Indep}$ is block diagonal, say

$$\boldsymbol{V}^{Indep} = \begin{pmatrix} \boldsymbol{B}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{B}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{B}_S \end{pmatrix}.$$

Now since the $\boldsymbol{B}_i$'s are themselves covariance matrices, we can write them as $\boldsymbol{B}_i = (\boldsymbol{B}_i^{1/2})^T \boldsymbol{B}_i^{1/2}$ and it's easy to see that we can write $\boldsymbol{V}^{Indep}$ as

$$\boldsymbol{V}^{Indep} = (\boldsymbol{A}^{1/2})^T \boldsymbol{I} \boldsymbol{A}^{1/2}, \tag{13}$$

where $\boldsymbol{A}^{1/2}$ is a block diagonal matrix with blocks $\boldsymbol{B}_1^{1/2}, \ldots, \boldsymbol{B}_S^{1/2}$ and $\boldsymbol{I}$ is the identity matrix. We can now easily create new covariance matrices $\boldsymbol{V}^{Dep}$ by replacing $\boldsymbol{I}$ in (13) by some correlation matrix $\boldsymbol{R}$. We must, however, impose a restriction on $\boldsymbol{R}$ in order to preserve the correlation structure within strata. More specifically, $\boldsymbol{R}$ must be of the following form:

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{R}_{1,2} & \cdots & \boldsymbol{R}_{1,S} \\ \boldsymbol{R}_{1,2} & \boldsymbol{I} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{R}_{S-1,S} \\ \boldsymbol{R}_{1,S} & \cdots & \boldsymbol{R}_{S-1,S} & \boldsymbol{I} \end{pmatrix}, \tag{14}$$

where the $\boldsymbol{I}$ represent identity matrices and the $\boldsymbol{R}_{ij}$ are correlation matrices.

For a fixed choice of correlation structure the generalized estimating equations are given by

$$U(\beta) = \sum_{c=1}^{K} D^{(c)\,T} V^{(c)\,-1} \{ Y^{(c)} - \mu^{(c)}(\beta) \}. \tag{15}$$

As usual, (15) reduces to (9) when the working correlation matrices $R^{(c)}$ are the identity matrices, i.e., when $V^{(c)} = V^{(c)\,Indep}$ for each c.

Assuming that the parameters α of the correlation matrices $R^{(c)}$ can be estimated with $\sqrt{K}$-consistency (see §3.2), then from Theorem 2 of Liang and Zeger (1986) we have that for any choice of correlation structure, $\sqrt{K}(\hat{\beta} - \beta)$ is asymptotically multivariate Gaussian with zero mean and variance matrix

$$\begin{aligned} V_G &= \lim_{K\to\infty} K \left(\sum_{c=1}^{K} D^{(c)\,T} V^{(c)\,-1} D^{(c)} \right)^{-1} \left\{ \sum_{c=1}^{K} D^{(c)\,T} V^{(c)\,-1} Var(Y^{(c)}) V^{(c)\,-1} D^{(c)} \right\} \\ &\quad \times \left(\sum_{c=1}^{K} D^{(c)\,T} V^{(c)\,-1} D^{(c)} \right)^{-1}, \end{aligned} \tag{16}$$

which is consistently estimated by replacing $Var(Y^{(c)})$ by $\{Y^{(c)} - \mu^{(c)}(\beta)\}\{Y^{(c)} - \mu^{(c)}(\beta)\}^T$ and α and β by their estimates in (16). Pan (2001) suggests an alternate estimator of the variance of $\hat{\beta}$. In our setup and with an equal number of cases and controls in all strata, Pan's estimator is obtained by replacing the estimator $\{Y^{(c)} - \mu^{(c)}(\beta)\}\{Y^{(c)} - \mu^{(c)}(\beta)\}^T$ of $Var(Y^{(c)})$ in (16) by

$$\left\{ A^{(c)\,1/2} \right\}^T \left(\sum_{c=1}^{K} \left\{ A^{(c)\,-1/2} \right\}^T S^{(c)} S^{(c)\,T} A^{(c)\,-1/2} \right) A^{(c)\,1/2}, \tag{17}$$

where $S^{(c)} = Y^{(c)} - \mu^{(c)}(\hat{\beta})$.

In some applications, the number of clusters, K, may be small relative to the number of observations per cluster. In such situations, Liang, Zeger & Qaqish (1992) suggest that GEE2 should be the method of choice. However, as outlined in Harbin

& Hilbe (2002, p. 105), robustness to misspecification of the correlation structure is then lost. Since correctly specifying a correlation structure for responses from different strata in conditionnal logistic regression is arguably a difficult task, we defer the treatment of the current problem using GEE2 to a future communication.

3.2. Inferences about correlation parameters

We expect that the random vectors within one cluster, $c \in \{1, \ldots, C\}$, $(Y_{si}^{(c)}, \mathbf{X}_{si}^{(c)})$ are exchangeable, for $i \in \{1, \ldots, N\}$, $s \in \{1, \ldots, S^{(c)}\}$. In light of this we assume that

$$\mathrm{corr}(Y_{si}^{(c)}, Y_{s'i'}^{(c)} | \mathbf{X}_{si}^{(c)}, \mathbf{X}_{si}^{(c)}) = g(\mathbf{X}_{si}^{(c)}, \mathbf{X}_{s'i'}^{(c)})$$

where g is the same for all $i \neq i'$, $s \neq s'$. In order to define the working correlation matrix, we need to specify the form of g.

For each cluster c, we define $\boldsymbol{\eta}_{(si)(s'i')}^{(c)}(\boldsymbol{\alpha}) = \mathrm{corr}(Y_{si}^{(c)}, Y_{s'i'}^{(c)})$ for all $s \neq s'$ and any i, i'. In the models considered by us it is reasonable to assume that a subvector of the covariates, say $\tilde{\boldsymbol{X}}$, can be used to model the correlation structure as follows.

$$\boldsymbol{\eta}_{(si)(s'i')}^{(c)}(\boldsymbol{\alpha}) = \frac{exp\{|\tilde{\boldsymbol{X}}_{s,i}^{c} - \tilde{\boldsymbol{X}}_{s',i'}^{c}|^T \boldsymbol{\alpha}\} - 1}{exp\{|\tilde{\boldsymbol{X}}_{s,i}^{c} - \tilde{\boldsymbol{X}}_{s',i'}^{c}|^T \boldsymbol{\alpha}\} + 1},$$

where $|\tilde{\boldsymbol{X}}_{s,i}^{c} - \tilde{\boldsymbol{X}}_{s',i'}^{c}|$ is the vector having as components the absolute differences between the components of $\tilde{\boldsymbol{X}}_{s,i}^{c}$ and $\tilde{\boldsymbol{X}}_{s',i'}^{c}$.

To estimate $\boldsymbol{\alpha}$ define the Pearson residuals

$$r_{ji}^{(c)} = \frac{y_{ji}^{(c)} - \mu_{ji}^{(c)}(\hat{\boldsymbol{\beta}})}{\sqrt{\mu_{ji}^{(c)}(\hat{\boldsymbol{\beta}})\{1 - \mu_{ji}^{(c)}(\hat{\boldsymbol{\beta}})\}}}. \tag{18}$$

It can be observed that

$$E[r_{si}^{(c)} \cdot r_{s'i'}^{(c)}] \approx \boldsymbol{\eta}_{(si)(s'i')}^{(c)}(\boldsymbol{\alpha})$$

so we can estimate $\boldsymbol{\alpha}$ through the regression of $\log\left(\frac{r_{si}^{(c)} \cdot r_{s'i'}^{(c)}+1}{r_{si}^{(c)} \cdot r_{s'i'}^{(c)}-1}\right)$ on $|\tilde{\boldsymbol{X}}^c_{s,i} - \tilde{\boldsymbol{X}}^c_{s',i'}|$. The approach described is similar to that of Liang and Zeger (1986).

4. Model selection criterion: The QIC

We now consider two aspects of model selection, namely the choice of a working correlation structure and the choice of the covariates that should be part of the linear predictor $\boldsymbol{\beta}^T\boldsymbol{x}$. For the latter, stepwise selection procedures with p-values based on the asymptotic normality of $\hat{\boldsymbol{\beta}}$ and the robust variance estimate based on (16) are one way to proceed. However, in some fields of application "information-theoretic" model selection criteria (e.g., AIC, BIC) are much preferred to p-value based model selection (Burnham and Anderson, 2002). Recently, Pan (2001) proposed a quasi-likelihood under independence criterion (QIC) that can be viewed as a generalization of the AIC under GEE. This criterion can be used for both working correlation structure and covariate selection. In general, in order to derive the QIC, one needs to compute the quasi-likelihood corresponding to the score equations (15) and a working independence matrix $V^{(c)} = V^{(c)\ Indep}$. However, our derivations in §3 imply that, under assumed independence between strata, the quasi-likelihood function is the same as the likelihood function (4). Following the notation of Pan (2001) we denote this likelihood function as $Q(\boldsymbol{\mu}(\beta), \boldsymbol{Y}, I)$, to emphasize that it was computed under the independence assumption.

Now let us use a (possibly different) working correlation matrix $\boldsymbol{R}$. To emphasize their dependence on $\boldsymbol{R}$, let us denote the solution of the GEE (15) by $\hat{\beta}(\boldsymbol{R})$ and its robust sandwich variance estimator by $\hat{\boldsymbol{V}}_G(\boldsymbol{R})$, and let $\hat{\beta}(\boldsymbol{I})$ be the maximum likelihood estimator of $\boldsymbol{\beta}$ under the independence assumption. We also define

$$\boldsymbol{\Omega}_{\boldsymbol{I}} = \sum_{c=1}^{C} \boldsymbol{D}^{(c)T} \boldsymbol{V}^{(c)\,Indep} \boldsymbol{D}^{(c)} = -\frac{\partial^2 Q(\boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{Y}, \boldsymbol{I})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\boldsymbol{I})},$$

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

$$QIC(\boldsymbol{R}) = -2Q(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}(\boldsymbol{R})), \boldsymbol{Y}, \boldsymbol{I}) + 2\text{trace}(\boldsymbol{\Omega}_{\boldsymbol{I}} \hat{\boldsymbol{V}}_G(\boldsymbol{R})). \tag{19}$$

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 $QIC(I)$ is the best criterion for covariate selection, while $QIC(R)$ can be used to choose among correlation structures.

5. Discussion

In this paper we have derived GEE inference for the conditional logistic regression model. An example of application of this method is given in Fortin et al., (2005), where the authors attempt to detect the factors that influence elk movements. More generally, the techniques developed here are applicable to any situation where conditional logistic regression is used and where there is reason to suspect that correlation is present among matched sets (strata).

Acknowledgement

The authors are thankful to Daniel Fortin for fruitful discussions and for having brought the methodological problem to them. Financial support from the Natural

Sciences and Engineering Research Council of Canada (both authors) and of the Fonds québécois de la recherche sur les sciences et les technologies (second author) is gratefully acknowledged.

References

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

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

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

Fortin, D., Beyer, H.L., Boyce, M.S., Smith, D.W., Duchesne, T. & Mao, J.S. (2005). Wolves influence elk movements: Behavior shapes a trophic cascade in Yellowstone National Park. *Ecology*, **86**, 1320-1330.

Hardin, J.W. & Hilbe, J.M. (2003). *Generalized Estimating Equations.* Chapman & Hall, Boca Raton.

Hosmer, D.W. & Lemeshow, S. (1989). *Applied logistic regression.* New York: Wiley.

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

Liang, K.-Y., Zeger, S.L. & Qaqish, B. (1992). Multivariate regression analyses for categorical data. *J. R. Statist. Soc. B*, **54**, 3-40.

LONGFORD, N.T. (1994). Logistic regression with random coefficients. *Comput. Statist. Data Anal.*, **17**, 1-15.

MILLER, M.E., DAVIS, C.S. & LANDIS, J.R. (1993). The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics*, **49**, 1033-1044.

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

PARK, E. & KIM, Y. (2004). Analysis of longitudinal data in case-control studies. *Biometrika*, **91**, 321-330.

SCOTT, A.J. & WILD, C.J. (1986). Fitting logistic models under case-control and choice based sampling. *J.R.Statist. Soc. B*, **48**, 170-182.

ZEGER, S.L. & LIANG, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121-130.