Statistica Sinica **21** (2011), 231-253

AN EM COMPOSITE LIKELIHOOD APPROACH FOR MULTISTAGE SAMPLING OF FAMILY DATA

Y. Choi[1,2] and L. Briollais[2,3]

[1] *University of Western Ontario,* [2] *Mount Sinai Hospital*
and [3] University of Toronto

Abstract: Multistage sampling of family data is a common design in the field of genetic epidemiology, but appropriate methodologies for analyzing data collected under this design are still lacking. We propose here a statistical approach based on the composite likelihood framework. The composite likelihood is a weighted product of individual likelihoods corresponding to the sampling strata, where the weights are the inverse sampling probabilities of the families in each stratum. Our approach is developed for time-to-event data and can handle missing genetic covariates by using an Expectation-Maximization algorithm. A robust variance estimator is employed to account for the dependence of individuals within families. Our simulation studies have demonstrated the good properties of our approach in terms of consistency and efficiency of the genetic relative risk estimate in the presence of missing genotypes and under different multistage sampling designs. Finally, an application to a familial study of early-onset breast cancer shows the interest of our approach. While it confirms the important effect of the genes BRCA1 and BRCA2 in these families, it also shows that incorrect inference can be made about this effect if the sampling design is not properly taken into account.

Key words and phrases: Composite likelihood, EM algorithm, family data, missing genotype, multistage sampling.

1. Introduction

The concept of multistage sampling is not new in epidemiology and genetic epidemiology; it has been used for example to investigate the association between a rare disease and a rare exposure (White (1982)). In familial genetic studies, multistage sampling permits the allocation of resources to families that are the most informative for a given objective while allowing population-based inference using a design-corrected estimator (Whittemore and Halpern (1997)). Typically, in the first stage, a simple random sample is drawn from the population and then stratified according to some easily measured covariates. In the subsequent stages a random subset of previously selected units is sampled for more detailed observation, with a unit's sampling probability determined by its covariates as observed in the previous stages. While multistage designs can be quite efficient in certain situations, they also raise numerous statistical challenges. A first difficulty with multistage design is the estimation of the parameters of interest when the sampling design is not ignorable. A second problem, that often occurs when collecting family data, is to deal with missing genetic covariates and correlated observations within families (or clusters).

The objective of this paper is to estimate a regression coefficient associated with a known gene mutation when the outcome is the time to an event and the sampling design is not ignorable. A composite likelihood approach for time-to-event data is proposed to provide a design-corrected estimator of the gene mutation effect (Sections 3 and 4). In family studies, mutation status is often missing for some of the family members. To infer the mutation status in these individuals, we used an Expectation-Maximizaion (EM) algorithm where the missing mutation statuses are estimated from their conditional expectation given the observed ages of onset (or censored times) and mutation statuses of other family members. Moreover, a robust variance estimator is proposed to account for the dependence of individuals within families (Section 5). Our simulation study investigates the properties of our proposed EM composite likelihood approach in the presence of missing genotype data, and under different multistage sampling strategies (Section 6). In Section 7, we illustrate our approach through an application to a familial study of early-onset breast cancer. A two-stage design was used to sample the families and our goal is to estimate the effect of gene mutations in two genes (BRCA1 and BRCA2) on time to breast cancer. Finally, we investigate an optimal sampling strategy based on the results obtained from our application (Section 8) and conclude with possible extensions of this work.

2. The Multistage Sampling Design

Our goal is to estimate a parameter θ in the probability density function of a random vector of observations y. We assume that the sample space of y can be divided into K disjoint strata, $S_1, \ldots, S_K$, and that the sampling units in each stratum provide different amount of information about θ. An efficient design could therefore consist in the two stages:

- Stage 1: draw a random sample $y_1, \ldots, y_N$ from the population and only observe the stratum to which each y_i belongs; let the sample be $S_k^1, k = 1, \ldots, K$.
- Stage 2: for each stratum S_k^1 with sample size N_k, draw a random subsample of size n_k independently with probability $p_k = P(y \in S_k^2 | y \in S_k^1)$, where p_k represents the probability that a unit from the stratum k of the first stage is selected in the second stage and all units in each stratum are sampled with the same sampling probability. Observe the y values of the observations

sampled in the second stage, and denote the sample from the second stage by $S^2 = \cup_{k=1}^{K} S_k^2$.

The two-stage design can be generalized to involve more stages. The first two stages would be similar to those of the two-stage design, except that at Stage 2 we do not observe the y directly but rather the finer substrata they belong to, and a random sample is drawn from these substrata. This process can be repeated several times until the last stage in which the y are observed directly. In practice, however, there could be some loss in efficiency in adding too many stages and designs with more than three stages are relatively rare, at least in the field of genetic epidemiology (Whittemore and Halpern (1997)).

A variant of this two-stage design consists of collecting family data in Stage 2. For example, case patients (and sometime population controls) are selected in the first stage and asked about the prevalence of a disease outcome in their family and, in the second stage, case patients with or without particular characteristics are subsampled with different sampling probabilities and are used to obtain extended-family histories. The sampling probability assigned to each family could depend on the case patient's (i.e. the proband's) genetic risk, age, and ethnicity and are computed according to some optimality criteria for a particular question of interest (Whittemore and Halpern (1997)).

Considering that our primary sampling units are families with several individuals, we denote by n_k the number of families sampled in stratum k at Stage 2, N_k the total number of families in stratum k, and S_{kf}^2 a family f $(f = 1, ..., n_k)$ in stratum k sampled from Stage 2. Thus the total sample S^2 at Stage 2 can be expressed as $S^2 = \cup_{k=1}^{K} S_k^2 = \cup_{k=1}^{K} \cup_{f=1}^{n_k} S_{kf}^2$. Assuming that a stratum is identified for all the primary sampling units, the full likelihood under a variety of sampling schemes has been shown to be (Lawless, Kalbfleisch, and Wild (1999))

$$L_F(y;\theta) = \prod_{k=1}^{K} \Big\{ \prod_{f=1}^{n_k} L_{kf}(y;\theta) \Big\} w_k(\theta)^{N_k - n_k},$$

where

$$L_{kf}(y;\theta) = \prod_{i \in S_{kf}^2} f(y_i;\theta)$$

is the likelihood of family f in stratum k, i is an individual in this family, and

$$w_k(\theta) = \int_{S_k^2} L_{kf}(y;\theta)dy$$

is the sampling probability for the families in stratum k.

Several estimation methods have been proposed to estimate the parameter θ without having to compute the full likelihood L_F, and it has been shown that some gain in efficiency can be obtained by including information about the stratum sizes (Lawless, Kalbfleisch, and Wild (1999)). However, in certain situations it might be difficult to compute the stratum-specific sampling probabilities $w_k(\theta)$, and the likelihood method can be sensitive to the misspecification of these probabilities. Alternatively, some weighted pseudo-likelihood methods have been proposed for problems involving response-selective observations (Kalbfleisch and Lawless (1988); Whittemore and Halpern (1997)) and missing data (Little and Rubin (1987)). The rationale of the weighted pseudo-likelihood approach is to consider only the completely observed units and weight their contribution inversely proportional to their probability of selection (Wild (1991); Whittemore and Halpern (1997)) to give the following log weighted pseudo-likelihood function

$$\ell_W(y;\theta) = \sum_{k=1}^{K} \tilde{p}_k^{-1} \sum_{f=1}^{n_k} \log\{L_{kf}(y;\theta)\},$$

where $\tilde{p}_k$ is the estimated sampling probability for family f in stratum k and assumed the same for all the families in this stratum. For basic stratified sampling, the use of $\tilde{p}_k = n_k/N_k$ provides an unbiased estimation of the parameter θ. For variable probability sampling, the use of $\tilde{p}_k = p_k$ leads to an unbiased estimator, but taking $\tilde{p}_k = n_k/N_k$ can yield more efficient estimates (Lawless, Kalbfleisch, and Wild (1999)).

3. A Composite Likelihood Formulation

The weighted pseudo-likelihood described above can be thought of as a composite likelihood based on independent contributions of each family to the likelihood. Composite likelihood has been proposed for estimation involving complex likelihood functions and is obtained by removing some terms in the complex full likelihood with the objective that the removed part is not very informative for estimating the parameter of interest, and the resulting loss of efficiency remains acceptable (Lindsay (1988)). If we consider a parametric statistical model $f(y;\theta), y \in \mathcal{Y} \subseteq \mathbb{R}^n, \theta \in \Theta \subseteq \mathbb{R}^d$, and a set of measurable events $\{\mathcal{A}_k; k = 1, ..., K\}$, then a composite likelihood is the weighted product of the likelihoods corresponding to each single event. The log composite likelihood has the general form

$$\ell_C(y;\theta) = \sum_{k=1}^{K} w_k \ell_k(y;\theta)$$

where, in the context of multistage design, the events correspond to the sampling in the different strata, $\ell_k(y;\theta) = \sum_{f=1}^{n_k} \sum_{i \in S_{kf}^2} \log\{f(y_i;\theta)\}$, and $w_k = \tilde{p}_k^{-1}$.

This formulation allows one to use the asymptotic theory developed for the maximum composite likelihood estimator. In particular, the observations do not need to be independent, as originally assumed for weighted pseudo-likelihood (Wild (1991)), and when they are correlated (i.e. individuals within a family in our context), a robust variance estimator can be used instead of the naive variance estimator. The maximum composite likelihood estimator, $\hat{\theta}$, is obtained by maximizing the log composite likelihood, $\ell_C(y;\theta)$, or by solving the composite score equations $U(\theta) = 0$ with

$$U(\theta) = \sum_{k=1}^{K} w_k U_k(\theta) = \sum_{k=1}^{K} w_k \frac{\partial \ell_k(y;\theta)}{\partial \theta}.$$

3.1. Robust variance estimator

As noted in Lindsay (1988) or Varin (2008), the maximum composite likelihood estimator $\hat{\theta}$ is consistent and asymptotically normally distributed. The limiting normal distribution has mean θ and variance matrix

$$H(\theta)^{-1} J(\theta) H(\theta)^{-1} ,$$

where

$$H(\theta) = E\left\{\frac{\partial U(\theta)}{\partial \theta}\right\} \quad \text{and} \quad J(\theta) = \mathrm{Var}\left\{U(\theta)\right\}. \tag{3.1}$$

An empirical estimator for $H(\theta)$ arises by dropping the expectation and replacing the unknown θ with its estimator $\hat{\theta}$,

$$\hat{H}(\theta) = \sum_{k=1}^{K} w_k \frac{\partial U_k(\theta)}{\partial \theta}|_{\theta=\hat{\theta}} ,$$

and, as we have $J(\theta) = \mathrm{Var}\left\{U(\theta)\right\} = \mathrm{E}\left[U(\theta)U(\theta)^\top\right]$, an empirical estimator for $J(\theta)$ at $\hat{\theta}$ is obtained as

$$\hat{J}(\theta) = \frac{1}{K} \sum_{k=1}^{K} w_k^2 U_k(\hat{\theta}) U_k(\hat{\theta})^\top .$$

3.2. Correction for ascertainment

In family-based multistage designs, it is often that some strata are not sampled at all, such as families of unaffected individuals (e.g. control proband) or families with affected individuals (e.g. case proband) above or below a certain age at onset, the proband being the individual from whom the family is ascertained. As an example, we consider in our application (see Section 7) a study that enrolled only case probands under 40 years old. A correction for ascertainment should be applied by computing a conditional likelihood where the conditioning corresponds to the event A_f that the proband in family f selected in stage 1 meets certain criteria (Choi, Kopciuk, and Briollais (2008)). We then express $L_f(y;\theta)$ as the product of conditional probabilities of family members observed for family f given the covariates x and the proband having the event A_f as

$$L_f(y;\theta) = \prod_{i \in S^2_{kf}} \frac{P(y_i|x_i)}{P(y_p \in A_f|x_p)}, \tag{3.2}$$

where index p represents the proband. We give more details about the family-specific composite likelihood in the next section when the response is the time to an event.

4. Application To Time to Event Data

In the following we are interested in modeling a time-to-event response, for example time to cancer, from our two-stage sampling design where we use family data at the second stage. In this context, design-based estimators have been proposed for the Cox's proportional hazard (PH) model (Cox (1972)) by Binder (1992) and Lin (2000), and recently extended by Boudreau and Lawless (2006) to stratified and clustered survival data. Boudreau and Lawless (2006) used a weighted estimating function to obtain a consistent estimator of the regression parameters, denoted by β. A partial likelihood approach is used when the sampling design is non-ignorable since the standard likelihood function may not yield a consistent estimator of β. An additional difficulty in our context is the correction for ascertainment described above. Because we need to compute a conditional likelihood to take into account the ascertainment of the families, we found that it would be challenging to do so in the Cox PH model setting. Instead, we develop a design-based composite likelihood estimation using a parametric model for the ages at onset data. We denote by T_i and x_i, respectively, the time of onset and the vector of covariates for individual i in family f and stratum k sampled in Stage 2 of the design ($i \in S^2_{kf}$). We assume that T follows the Weibull model that has the survival and hazard functions

$$S(t|x) = e^{-\{\lambda(t-t_0)\}^{\rho} e^{x^\top \beta}},$$
$$h(t|x) = \lambda\rho\{\lambda(t-t_0)\}^{\rho-1} e^{x^\top \beta},$$

where t_0 is the minimum time (age) at which an event can occur, x is a vector of the covariates of interest, β is a corresponding vector of regression parameters, and λ and ρ are the scale and shape parameters of the Weibull distribution, respectively. Hereafter x represents the gene mutation status (with individuals carrying the mutation coded 1 and non-carriers coded 0) with β a corresponding regression parameter.

Consider a response time $y_i \geq 0$ corresponding to the age at onset t_i if individual i is affected, or the current age a_i otherwise. The likelihood contribution for an individual i in family f and stratum k follows from (3.2) with

$$f(y_i \mid x_i) = h(t_i|x_i)^{\delta_i} S(t_i|x_i) \left\{ \frac{1 - S(t_i|x_i)}{h(t_i|x_i)S(t_i|x_i)} \right\}^{\nu_i}, \tag{4.1}$$

and $P(y_p \in A_f|x_p) = 1 - S(a_p|x_p)$, where δ_i indicates the affection status, $\nu_i = 1$ if the age of onset is not reported for affected individual i but only his/her current age (age at examination) is known, 0 otherwise, and a_p and x_p denote the age of examination and mutation status of the proband, respectively.
The log composite likelihood for a family f is

$$\begin{aligned} \ell_f(y;\theta) = & \sum_{i \in S_{kf}^2} (\delta_i - \nu_i) \log\left[\lambda\rho\{\lambda(t_i - t_0)\}^{\rho-1}\right] \\ & + \sum_{i \in S_{kf}^2} (\delta_i - \nu_i) x_i \beta \\ & + \sum_{i \in S_{kf}^2} (\nu_i - 1)\{\lambda(t_i - t_0)\}^{\rho} e^{x_i\beta} \\ & + \sum_{i \in S_{kf}^2} \nu_i \log\left[1 - e^{-\{\lambda(t_i - t_0)\}^{\rho} e^{x_i\beta}}\right] \\ & - \log\left[1 - e^{-\{\lambda(a_p - t_0)\}^{\rho} e^{x_p\beta}}\right]. \end{aligned} \tag{4.2}$$

Estimating baseline risk parameters (λ, ρ) from mutation carrier families might be problematic since noncarrier individuals in these families can exhibit higher risk than in the general population (Begg (2002)). To circumvent this problem, we adopted a two-stage estimation procedure where the baseline parameters are estimated in a first stage using mutation non-carrier families only, and the β parameter is estimated in a second stage using both mutation carrier and non-carrier families. Therefore, we considered the parameter β as the parameter of interest and the other parameters (λ, ρ) as nuisance parameters. Then, we estimate the parameter β using a profile likelihood approach by inserting the estimates of $\gamma = (\lambda, \rho)$ into the log composite likelihood function in (4.2) as a function of β only.

For fixed γ, we find the maximum composite likelihood estimate of β as $\hat{\beta}(\gamma)$. Then, we fix β and estimate γ. By iterating these two steps until convergence, we obtain maximum profile likelihood estimates $\hat{\beta}$ and $\hat{\gamma}$. To avoid computational complications, assuming that γ is fixed at $\hat{\gamma}$, the maximum composite likelihood estimator $\hat{\beta}$ based on two-stage estimation procedure of the profile likelihood has asymptotic properties of normal distribution with mean β and variance of a form, $H(\beta)^{-1}J(\beta)H(\beta)^{-1}$, where $H(\beta)$ and $J(\beta)$ are estimated at $\hat{\beta}$ for fixed $\hat{\gamma}$,

$$\hat{H}(\beta)=\sum_{k=1}^{K} w_k \sum_{f=1}^{n_k} \frac{\partial U_f\{\beta(\hat{\gamma})\}}{\partial\beta}|_{\beta=\hat{\beta}} \text{ and } \hat{J}(\beta)=\sum_{k=1}^{K} w_k^2 \sum_{f=1}^{n_k} U_f\{\hat{\beta}(\hat{\gamma})\}U_f\{\hat{\beta}(\hat{\gamma})\}^\top .$$

Here, $U\{\beta(\gamma)\}$ and $\partial U_f\{\beta(\gamma)\}/\partial\beta$ represent the first and second derivatives of the profile log-likelihood with respect to β at a fixed value of γ, respectively; detailed expressions are presented in Appendix I. The asymptotic normality of the two-stage estimator was also shown in the context of correlated survival data in Proposition 3.1 of Andersen (2005). As pointed out by an associate editor, this variance estimator might underestimate the asymptotic variance; however, our simulation studies showed little difference between this estimator and the asymptotic variance estimator.

5. An EM Algorithm for Missing Genotype Covariates

Family data often include some missing information, in particular, missing genotypes. In our simulation study and data application (e.g. see Sections 6 and 7), we consider situations where mutation statuses are partially missing in the family. In the presence of missing genotype information, we estimate the disease risk associated with a known gene mutation in the families. Suppose a vector of genetic covariates g_f in family f consist of observed genotypes g_{fo} and missing genotypes g_{fm} and a vector of time-of-onset responses y_f (also called"phenotypes" in genetics) with no missing data. To infer the unobserved genotypes in the family, we implement an EM algorithm (Dempster, Laird, and Rubin (1977)) to estimate the parameters in our composite likelihood approach. The EM algorithm is an iterative procedure that computes the maximum likelihood estimates (MLEs) in the presence of missing data. At each iteration, it takes two steps–the expectation and maximization steps. In our situation, the expectation of the complete data (y_f, g_f) is taken with respect to the conditional distribution of missing genotypes g_{fm} given observed data (y_f, g_{fo}), and current estimates of θ and then the parameter estimates are updated by maximizing the likelihood function using the estimate of missing data in the expectation step. These steps iterate until convergence to obtain the MLEs, since the algorithm is guaranteed to increase the likelihood at each iteration. The following details the application of the EM algorithm for our proposed composite likelihood approach.

E-step:
Let θ be the vector of unknown parameters and $\theta^{(j)}$ denote their values at the end of the jth iteration. The Q function is obtained as the conditional expectation of the log-likelihood function given the observed information (y_f, g_{fo}) at $\theta^{(j)}$, $f = 1, \ldots, n$:

$$\begin{aligned}
Q(\theta|\theta^{(j)}) &= \sum_{k=1}^{K} w_k \sum_{f=1}^{n_k} Q_f(\theta|\theta^{(j)}), \\
Q_f(\theta|\theta^{(j)}) &= E_{\theta^{(j)}}[\ell_f(\theta)|y_f, g_{fo}] \\
&= \sum_{i \in S_{kf}^2} (\delta_i - \nu_i)(\log[\lambda\rho\{\lambda(t_i - t_0)\}^{\rho-1}] + \beta E_{\theta^{(j)}}[g_{im}|y_f, g_{fo}]) \\
&\quad + \sum_{i \in S_{kf}^2} (\nu_i - 1)\{\lambda(t_i - t_0)\}^{\rho} E_{\theta^{(j)}} \left[e^{x_i^\top \beta}|y_f, g_{fo}\right] \\
&\quad + \sum_{i \in S_{kf}^2} \nu_i E_{\theta^{(j)}} \left[\log(1 - e^{-\{\lambda(t_i - t_0)\}^{\rho} e^{x_i \beta}})|y_f, g_{fo}\right] \\
&\quad - \log(1 - e^{-\{\lambda(a_p - t_0)\}^{\rho} e^{x_p \beta}}).
\end{aligned}$$

Thus, the unobserved genotype g_{im} for individual i in family f comes into the complete data log-likelihood using the conditional expectation given its observed responses and genotypes in the family, which can be expressed as

$$\begin{aligned}
&E_{\theta^{(j)}}[g_{im}|y_f, g_{fo}] \\
&\quad = P_{\theta^{(j)}}(g_{im} = 1|y_f, g_{fo}) \\
&\quad = \frac{P_{\theta^{(j)}}(y_f|g_{im} = 1, g_{fo}) P(g_{im} = 1|g_{fo})}{P_{\theta^{(j)}}(y_f|g_{im} = 1) P(g_{im} = 1|g_{fo}) + P_{\theta^{(j)}}(y_f|g_{im} = 0) P(g_{im} = 0|g_{fo})},
\end{aligned}$$

where $P(g_{im}|g_{fo})$ is the conditional probability of mutation carrier status of i given the observed genotypes in family f, and obtained based on Mendelian transmission probabilities for individuals with parents in the family or based on the mutation frequency in the population for individuals without parents in the family, and $P_{\theta^{(j)}}(y_f|g_{im}, g_{fo})$ has the form given in (4.1). In addition,

$$\begin{aligned}
E_{\theta^{(j)}}[h(g_{im})|y_f, g_{fo}] = {} & h(g_{im} = 1) P_{\theta^{(j)}}(g_{im} = 1|y_f, g_{fo}) \\
& + h(g_{im} = 0) P_{\theta^{(j)}}(g_{im} = 0|y_f, g_{fo}),
\end{aligned}$$

where $h(x)$ is any function of x.

M-step: $\theta^{(j+1)}$ is found by maximizing $Q(\theta|\theta^{(j)})$ with respect to θ.

The maximum composite likelihood estimates are obtained by iterating these two steps until convergence to update at each iteration the parameter values that maximize the expectation of the complete data log-likelihood.

5.1. Robust variance estimators for the EM algorithm

The variance estimator for composite likelihood should be modified for the facts that missing genotypes have been estimated using the EM algorithm. The observed information matrix in the EM algorithm was suggested by Louis (1982). Let $U(\theta)$ and $B(\theta)$ denote the score vector and the negative of the associated matrix of second derivatives for the complete data, respectively, and $U^*(\theta)$ and $B^*(\theta)$ be the same vector and matrix for the incomplete data. Then the observed information matrix can be expressed as

$$I_o(\theta) = E_\theta[B(\theta)|g_o, y_o] - E_\theta[U(\theta)U^\top(\theta)|g_o, y_o] + U^*(\theta)U^{*\top}(\theta), \tag{5.1}$$

where g_o and y_o denote the vectors of observed genotypes and responses from data. At the maximum composite likelihood estimate of θ, because of the convergence of the EM algorithm, U^* is zero. Thus the observed information matrix can be obtained as the first two terms on the right hand side of (5.1) that arise from the complete data log-likelihood analysis. The first term is evaluated as

$$E_\theta[B(\theta)|g_o, y_o] = E_\theta\left[-\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta^\top}|g_o, y_o\right].$$

We used the modified observed information matrix from the EM algorithm to get the variance estimator for the composite likelihood,

$$\mathrm{Var}(\theta) = I_o(\theta)^{-1}J(\theta)I_o(\theta)^{-1},$$

where $J(\theta)$ is the expected information matrix which can be also written in terms of the variance of the score vector for the conditional distribution given observed data as

$$\begin{aligned} J(\theta) &= \mathrm{Var}[U(\theta)] \\ &= \mathrm{Var}[\mathrm{E}\{U(\theta)|g_o, y_o\}] + \mathrm{E}[\mathrm{Var}\{U(\theta)|g_0, y_o\}] \\ &= \mathrm{E}_\theta[U(\theta)U^\top(\theta)|g_o, y_o] \\ &= \sum_{k=1}^{K} w_k^2 \sum_{f=1}^{n_k} \mathrm{E}_\theta\left[U_f(\theta)U_f(\theta)^\top|g_o, y_o\right], \end{aligned}$$

and I_o is the observed information matrix obtained from the EM algorithm, which can be expressed as $I_o(\theta) = E_\theta[B(\theta)|g_o, y_o] - J(\theta)$.

Thus, the robust variance of $\hat{\theta}$ can be estimated by

$$\hat{\text{Var}}(\hat{\theta}) = I_o(\hat{\theta})^{-1} J(\hat{\theta}) I_o(\hat{\theta})^{-1}.$$

Recall that this robust variance estimator can also account for familial correlation since we did not model explicitly the dependency between family members not due to the presence of a major gene.

5.2. Statistical properties

It has been recently shown that the EM algorithm for composite likelihood retains the important theoretical properties of the classical full likelihood EM algorithm (Gao and Song (2009)). These properties are (i) the proposed EM algorithm for composite likelihood retains the ascent property (i.e. the log-likelihood of the observed data is non-decreasing over the sequence of updated estimates $\beta^{(r)}$), (ii) it is a fixed point algorithm converging to a stationary point, and (iii) the convergence rate of the new algorithm depends on the curvature of the composite likelihood function surface. In addition, we prove in Appendix II that the parameter estimator is consistent given our specific ascertainment-corrected likelihood function for family data. Note that this consistency is different from algorithmic convergence of EM algorithm proved by Gao and Song (2009).

6. Simulation Study

We performed Monte Carlo simulations to investigate the properties of our novel EM composite likelihood approach, especially the bias and efficiency in the presence of missing genotype information and under different multistage sampling strategies. We considered a single stage sampling (no oversampling involved) and two 2-stage sampling designs where high-risk (HR) families were oversampled compared to low-risk (LR) families.

6.1. Family data generation

We generated families with three generations following the principle described in Choi, Kopciuk, and Briollais (2008). In brief, family members' ages at examination were first generated using a normal distribution with mean age 65 for the first generation and 45 for the second generation, with variance fixed at 2.5 years for both. The third generation had an average of 20 years difference with variance 1 year from the second generation. To generate the genotypes, the proband's genotype of a major gene was determined conditional on her/his affection status, assuming Hardy-Weinberg equilibrium (HWE) with fixed population allele frequencies. The proband was required to be affected by disease before his/her age at examination. Given the proband's genotypes, the genotypes of the other family members were then determined using HWE and Mendelian transmission probabilities calculated with Bayes' formula. Once we simulated the age at examination and genotype information for all family members, the time-to-onset of individual i was then simulated from the Cox's PH model with Weibull baseline,

$$h(t_i|g_i) = h_0(t_i)\exp(\beta x_i),$$

where x_i indicates if the individual i is a carrier of disease mutation gene, and the baseline hazard was assumed to follow the Weibull distribution with $h_0(t) = \lambda\rho\{\lambda(t-20)\}^{\rho-1}$.

The proband's age at onset was generated conditioning on the fact that the proband was affected before his(her) age at examination, a_p. For the rest of family members, their times to onset were generated unconditionally. We also assumed the minimum age at onset was 20 years of age and the maximum age for followup was 90 years of age. Finally, the affection status, δ_i, for individual i was determined by comparing the age at onset, t_i, and age at examination, a_i: $\delta_i = 1$ if $t_i < a_i$, and 0 otherwise.

6.2. Simulation study design

Generating ages of onset data was based on a Weibull distribution with scale (λ) and shape (ρ) parameters set at 3.4 and 0.01, which leads to a cumulative risk of 9% in the non-carrier group by age 70. These parameters were close to those observed in our data analysis (See Section 7 below). The β coefficient for the major gene effect varied between 1 and 3, and the missing genotype rates considered were 0%, 10%, 25% and 50%.

Three sampling designs were considered: a single stage design with no oversampling, which corresponds to a proportion of high-risk families of 15% (design 1), and two 2-stage sampling designs with oversampling of high-risk families whose proportion was increased to 30% (design 2) and 50% (design 3), respectively, where high risk families are defined as having two or more affected individuals. For each choice of log genetic relative risk (given by β), we fixed the minor allele frequency of the major gene in order to have the same proportion of high risk families. The allele frequency was set at 1%, 0.4%, 0.15%, 0.07%, and 0.03% for β=1, 1.5, 2, 2.5, and 3, respectively, corresponding to 15% of high risk families in the one-stage design. For the sampling design with 30% and 50% of high risk families, we sampled all high risk families with probability 1 and the low risk families with probability 41% and 18%, respectively. For each parameter combination, we performed runs of 500 simulations, each run including 1,000 families, and obtained the maximum composite likelihood estimates based on our proposed EM composite likelihood approach.

Table 1. Accuracy and precision of the EM composite likelihood estimation of log relative risk (β) of the major gene effect subject to various missing genotype rates in the a single stage design (no oversampling).

Allele	True	0% Missing			10% Missing			25% Missing			50% Missing		
Freq.	β	Bias	SE[1]	MSE[2]	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
1%	1.00	0.01	0.30	0.09	0.01	0.33	0.11	0.01	0.42	0.17	0.00	0.64	0.41
0.4%	1.50	0.01	0.31	0.10	0.00	0.34	0.12	-0.01	0.43	0.18	0.01	0.67	0.45
0.15%	2.00	0.02	0.34	0.11	0.01	0.37	0.14	0.02	0.46	0.21	0.03	0.68	0.46
0.07%	2.50	0.03	0.33	0.11	0.03	0.36	0.13	0.03	0.46	0.22	0.04	0.69	0.47
0.03%	3.00	0.03	0.34	0.12	0.03	0.38	0.14	0.04	0.47	0.23	0.05	0.87	0.75

[1] robust standard error

[2] mean square error

6.3. Simulation results

The results of our simulation study are summarized in Table 1 for the single stage sampling and in Table 2 for the 2-stage sampling designs, where we present the average bias, the robust standard error (SE), and the mean square error (MSE) of the log relative risk (β) of the major gene for each simulated scenario. For the 2-stage sampling designs, we also compare the design-corrected and uncorrected estimates in terms of bias and precision in Table 2.

For the single stage sampling design, displayed in Table 1, the average bias was negligible (absolute value less than 0.05) regardless of the gene effect and missing rate considered; the magnitude of the bias was always much smaller than the standard error. As we would expect, the precision of the parameter β increased (standard error decreased) with the allele frequency and inversely with the genetic effect and the proportion of missing data.

For the 2-stage sampling designs, shown in Table 2, our design-corrected estimates outperformed the design-uncorrected estimates in accuracy. The magnitudes of the bias for the design-corrected estimates were almost negligible while the design-uncorrected estimates were subject to a severe bias and appeared to overestimate β in most situations. The standard error of the design-corrected estimate increased with the genetic risk effect except in a few cases corresponding to a high rate of missing data and large β values. We also observed a trend toward larger standard errors with higher proportion of missing genotypes in the family, as expected, and this remained true for the various values of β and the different sampling designs. This pattern of SEs was also reflected in the MSE values. The MSE measures the tradeoff between bias and precision, defined as $\mathrm{Var}(\hat{\beta}) + \mathrm{Bias}(\hat{\beta}, \beta)^2$ where $\mathrm{Bias}(\hat{\beta}, \beta) = \hat{\beta} - \beta$. In most situations, the MSE values increased with the missing rate of genotypes and the log genetic risk effect (β). In terms of design efficiency, we noticed that both the MSE values and SE estimates were slightly larger in design 3 (50% HR families) than in design

Table 2. Comparison of the design-corrected and uncorrected estimates of log relative risk (β) of the major gene effect subject to various missing genotype rates in the multistage sampling designs.

HR	True		0% Missing			10% Missing			25% Missing			50% Missing		
families	β	Design	Bias	SE[1]	MSE[2]	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
30%	1.00	Corr	0.01	0.23	0.05	0.02	0.26	0.07	0.02	0.34	0.11	0.03	0.57	0.33
	1.00	Uncorr	0.69	0.21		0.73	0.23		0.81	0.28		0.92	0.40	
	1.50	Corr	0.02	0.25	0.06	0.02	0.28	0.08	0.01	0.37	0.14	0.02	0.65	0.43
	1.50	Uncorr	0.63	0.21		0.64	0.23		0.68	0.28		0.76	0.39	
	2.00	Corr	0.04	0.28	0.08	0.04	0.32	0.10	0.03	0.42	0.18	0.04	0.80	0.65
	2.00	Uncorr	0.55	0.22		0.56	0.24		0.58	0.29		0.61	0.40	
	2.50	Corr	0.02	0.29	0.08	0.03	0.32	0.10	0.04	0.42	0.18	0.04	0.81	0.66
	2.50	Uncorr	0.41	0.21		0.41	0.23		0.42	0.27		0.44	0.38	
	3.00	Corr	0.07	0.30	0.09	0.05	0.33	0.11	0.08	0.41	0.17	0.04	0.66	0.44
	3.00	Uncorr	0.31	0.22		0.30	0.23		0.30	0.28		0.29	0.38	
50%	1.00	Corr	0.04	0.21	0.05	0.03	0.25	0.06	0.05	0.38	0.15	0.05	1.33	1.78
	1.00	Uncorr	1.18	0.16		1.24	0.17		1.36	0.20		1.54	0.26	
	1.50	Corr	0.04	0.24	0.06	0.04	0.29	0.09	0.03	0.49	0.24	0.04	1.91	3.65
	1.50	Uncorr	1.05	0.16		1.07	0.17		1.13	0.20		1.22	0.26	
	2.00	Corr	0.05	0.28	0.08	0.06	0.34	0.12	0.05	0.62	0.39	0.06	1.72	2.95
	2.00	Uncorr	0.87	0.16		0.88	0.17		0.90	0.20		0.93	0.27	
	2.50	Corr	0.04	0.30	0.09	0.04	0.36	0.13	0.06	0.66	0.44	0.06	1.12	1.26
	2.50	Uncorr	0.64	0.16		0.64	0.17		0.64	0.20		0.65	0.26	
	3.00	Corr	0.06	0.31	0.10	0.06	0.35	0.12	0.07	0.49	0.24	0.10	0.68	0.47
	3.00	Uncorr	0.42	0.16		0.42	0.17		0.41	0.20		0.39	0.28	

[1] robust standard error

[2] mean square error

Table 3. The asymptotic relative efficiency of the two multistage designs compared to the single stage design for estimating log relative risk β

		30% HR families					50% HR families			
β	1	1.5	2	2.5	3	1	1.5	2	2.5	3
0% missing	1.75	1.59	1.46	1.31	1.32	2.09	1.72	1.44	1.21	1.19
10% missing	1.69	1.52	1.35	1.27	1.32	1.82	1.41	1.21	1.03	1.20
25% missing	1.52	1.32	1.20	1.22	1.35	1.21	0.76	0.55	0.50	0.93
50% missing	1.25	1.05	0.72	0.72	1.73	0.23	0.12	0.16	0.37	1.62

2 (30% HR families), especially when the samples involved a large amount of missing genotypes.

Based on these results, we further studied the relative performance of our three different designs based on their asymptotic relative efficiency (ARE). The ARE of one design A_1 to another A_0 is given by the ratio of the inverse asymptotic variances of the parameter β. Here, we evaluated the relative efficiencies of designs 2 and 3 compared to design 1. Our results clearly indicated that the ARE depended on the genetic model considered and the proportion of missing genotypes (Table 3). More specifically, with no missing genotype data, the two 2-stage sampling designs provided more efficiency than the single stage design. Compared to each other, the relative efficiency of the 2-stage designs varied with the value of β. For genetic models with β lower than 2.0, a balanced design with 50% HR and 50% LR families was the best strategy; for β higher than 2.0, sampling 30% of HR and 70% of LR families improved the accuracy of $\hat{\beta}$. In the presence of missing genotypes, the pattern of AREs became more complicated. With 10% and 25% of missing genotype data, design 2 was almost always the most efficient; with 50% missing genotype data, the general pattern became unclear. This reflects the difficulty in making inference about β in that situation, as clearly illustrated by the larger SEs.

7. Application to an Early-Onset Breast Cancer Study

We now show an application of our approach to a family study of early-onset breast cancer among BRCA1/2 mutation carriers. The goal is to estimate the genetic relative risk (GRR) associated with mutations in the BRCA1 and BRCA2 genes. The family data were collected from three population-based breast cancer family registries (Ontario, Northern California, and Australia) as a part of the NCI-funded Breast Cancer Family Registries (Breast CRF) initiatives (John et al. (2004)). The Ontario and Northern California registries used a two-stage sampling design while the Australian registry used a one-stage design. In the first stage, affected probands were randomly selected from the cancer registries and in the second stage, probands and their relatives were sampled with different sampling probabilities depending on their family history, ethnic origin, and age. The sampling criteria can be summarized into two categories: high risk and low risk, and only the low risk families had sampling probabilities lower than one (John et al. (2004)). In this study we focus only on the early breast cancer families whose probands were affected before the age 40.

7.1 The data

A total of 1,505 early breast cancer families was identified by the three registries but only 974 of them with a known mutation status for either BRCA1 or BRCA2 were used in our analyses. For BRCA1 analysis, we used 924 families (including 98 mutation positive and 826 mutation negative families) after exclusion of the BRCA2 positive families in order to remove any possible confounding in the baseline risk estimation. This number breaks down into 334, 248, and 342 families from Australia, Ontario, and Northern California, respectively. Similarly, 876 families were used for BRCA2 analysis (including 50 mutation positive and 826 mutation negative families). This number breaks down into 321, 225, and 330 families from Australia, Ontario, and Northern California, respectively, after exclusion of the BRCA1 positive families. The estimation of the baseline survivor function was based on the 826 BRCA1 and BRCA2 negative families.

We recall that for basic stratified sampling or variable probability sampling, the use of $\tilde{p}_k = n_k/N_k$ provides an efficient estimation of the parameter of interest θ. However, because we did not have a good estimate of N_k we used $\tilde{p}_k = p_k$, where the p_k were known and fixed by design. The Ontario and Northern California registries under-sampled the low-risk families, less informative in the study of the segregation of BRCA1/2 mutations, and those families were assigned a sampling probability lower than 1. All the high-risk families in these two registries had sampling probabilities of 1, as did all families in the Australian registry. The low-risk families consisted of 13 out of 248 families in the Ontario registry with inclusion probability of 0.25, and 56 out of 342 families in the Northern California registry, with sampling probabilities between 0.046 and 0.416. These sampling probabilities were meant to achieve some optimality criteria regarding the estimation of the genetic effect of interest (Siegmund, Whittemore, and Thomas (1999); John et al. (2004)). The inverse sampling probabilities were used as weights in the composite likelihood expression to compute a design-corrected GRR.

7.2. The Statistical analyses

We compared the design-corrected and uncorrected estimates of the log GRR associated with BRCA1 or BRCA2 mutations in the presence of missing genotypes. For the design-corrected GRR estimates, the robust variance accounts for the design effect and the use of family data. For the design-uncorrected GRR, variance estimates were calculated based on both the naive variance estimator assuming independent observations and the robust variance estimator. The results are summarized in Table 4.

The log GRRs of BRCA1 gene mutation associated with early onset breast cancer were estimated at 1.92 and 1.76 with and without design correction, respectively. The corresponding standard error estimates were 0.45 (robust SE) with design correction and 1.14 (robust SE) and 0.23 (naive SE, based on the second derivative of the pseudo loglikelihood) when no design correction was applied. The estimates of the log GRRs associated with BRCA2 were 1.90 and 1.77 with and without design correction and the corresponding robust standard errors were 0.62 and 2.13, respectively, while the naive estimator was 0.24.

These results confirmed the important role of the genes BRCA1 and BRCA2 in early-onset breast cancer and gave an estimate of their effect size in a large sample of families. From a methodological point of view, they also showed that

Table 4. Estimates of the log genetic relative risk (GRR) and corresponding standard error (SE) for BRCA1 and BRCA2 gene mutations associated with early-onset breast cancer.

Sampling Design	BRCA1	BRCA2
No Design Correction	1.76	1.77
Robust SE	(1.14)	(2.13)
Naive SE	(0.23)	(0.24)
Design Corrected	1.92	1.90
Robust SE	(0.45)	(0.62)

assuming an ignorable sampling design can lead to both biased estimates and underestimation of the standard errors, as also shown in our simulations, and therefore wrong hypothesis testing results. The naive SE estimator of the design-uncorrected analysis in Table 4 clearly underestimated the robust SE of the design-corrected analysis. Overall, we found that BRCA1 mutation appeared to have slightly higher relative risk than BRCA2 for early-onset breast cancer ($\hat{\beta}$=1.92 for BRCA1, 1.90 for BRCA2). The robust variance estimates played two important roles in our data analysis, first taking into account the sampling design and second, the residual familial correlation not due to BRCA1/2 gene mutations.

8. Design Optimality

To construct an optimal design we first determine some optimal weights for each stratum and then decide the optimal sample sizes accordingly. The optimal weighting problem was discussed by Lindsay (1988) who obtained optimal weights in a way that maximizes the information over a class of estimating functions. Let w be the vector of weights, S the vector of component scores, and U be the score function based on the full likelihood. Then, the optimal weights satisfy

$$\min_w \mathrm{E}_\beta (U - w^\top S)^2,$$

and are given by

$$w_{opt} = [\mathrm{Var}(S)]^{-1}\mathrm{E}(US),$$

with $\mathrm{E}(US) = E(S^2)$, where S^2 denotes the vector whose elements are the squared elements of S, and $\mathrm{Var}(S)$ is a block matrix where the size of each block depends on the size of the stratum.

Consider the problem of determining the optimal weights for estimating the GRR associated with BRCA1 mutations. As in our data setting, the optimal weights were determined separately for the Ontario and Northern California registries, assuming that each registry used only two sampling strata (high-risk and low-risk families) with high-risk families sampled with probability 1. The Australian registry, which used only a one-stage design, was not considered for this problem. The optimal sampling weights for the strata in each registry are proportional to the variance of the GRR estimate in each stratum (see the definition of w_{opt} above). The robust variances for the GRR estimates in high-risk and low-risk families were 4.04 and 1.12, respectively, in the Ontario registry and 0.38 and 0.28 in the Northern California registry. As the sampling probabilities are inversely proportional to these variances, the sampling probability of one for the high-risk families leads to a sampling probability of the low-risk families of 0.28 in the Ontario registry and 0.72 in Northern California. For Ontario, this was very close to the actual sampling probability but, for the Northern California registry, sampling more low-risk families could have increased the efficiency of the BRCA1 GRR estimate. The reason could be that low-risk families contribute to the estimation of the baseline survival function and, because it is correlated with the GRR, they also contribute to the GRR efficiency (Choi, Kopciuk, and Briollais (2008)).

9. Concluding Remarks

Our work shows the interest of the composite likelihood framework for analyzing data collected under a multistage sampling design. This is to our knowledge the first such application. Using an appropriate weighting of individual composite likelihoods corresponding to the different sampling strata allows one to obtain consistent parameter estimates while the use of a robust variance estimator provides an advantageous way of accounting for the use of family data and sampling design. To model time to onset data, we used a parametric model which offers some flexibility in correcting for the ascertainment bias of the families, a well-known problem in genetic epidemiological studies (Choi, Kopciuk, and Briollais (2008)). A discussion about optimal weights is also provided. Finally, the composite likelihood framework is extended through the use of an EM algorithm to account for missing genetic covariates. The composite likelihood methodology has also been used in family studies where the family likelihood is decomposed into a product over pairs of relatives where each pair has a weight that depends only on the family size (Andersen (2004)). The likelihood has the form

$$\ell_C(y;\theta) = \sum_{j=1}^{n} w_j \sum_{(i,h)\in\mathcal{G}_j} \ell_{ih}(y;\theta),$$

where $\mathcal{G}_j$ represents the set of possible pairs for family j, $\ell_{ih}(y;\theta)$ is the log-likelihood for a pair (i,h) in family j, and w_j is a weight that depends on the family size. The log-likelihood for a pair of relatives can be specified with a copula function for paired survival data (Andersen (2004); Choi and Matthews (2005)). This approach could lead to more efficient estimation than our method because it models directly the association between family members. In our case, we do not explicitly model association within families but used a robust variance estimator instead. Our method can be extended to include the estimation of some additional residual familial correlations besides the effects of BRCA1 and BRCA2. We have recently proposed various likelihood formulations using gamma frailty models in the context of familial studies that could account for familial residual correlations (Choi and Briollais (2010)). In particular, a pairwise likelihood formulation that we proposed could be combined with Andersen's pairwise composite likelihood and lead to a more general framework to analyze family data under a multistage design. We are also planning to expand our work to estimate the baseline hazard function parameters jointly with the genetic relative risk parameters, so that cumulative risks associated with specific gene mutations could be estimated under our approach.

Finally, our results confirm the important role of the genes BRCA1 and BRCA2 in early-onset breast cancer families. Some additional results also showed evidence for a possible additional major gene besides BRCA1 and BRCA2 in these families, but these results need some further confirmation. Also we cannot exclude that the cumulative risks associated with BRCA1 and BRCA2 modified by other genetic or non-genetic risk factors, however these risk factors are still not very well known (Antoniou and Chevenix-Trench (2010)). The emergence of genome-wide association studies raises some hope that common genetic variants could be identified as modifiers of BRCA1 and BRCA2 (Antoniou and Chevenix-Trench (2010)) and statistical approaches such as the one proposed here could help in this investigation when data are collected under a multistage design.

Acknowledgement

This research was supported by a grant from the Institutes of Genetics and Population and Public Health of the Canadian Institutes of Health Research (Grant # 110053), an Interdisciplinary Health Research Team award from the Canadian Institutes of Health Research (Grant # 43821), the Natural Sciences and Engineering Research Council of Canada, and a fellowship from the Canadian Breast Cancer Foundation Ontario Chapter. This work was also supported by the United States National Cancer Institute, National Institutes of Health under RFA # $CA - 95 - 003$ as part of the Breast Cancer Family Registries (Breast CFR), and through cooperative agreements with Breast CFR and principal investigators from Cancer Care Ontario (U01 CA69467), Northern California Cancer Center (U01 CA69417) and University of Melbourne (U01 CA69638).

The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFRs, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CFRs.

Appendix I

The first and second derivatives of the profile log-likelihood, based on (4.2), with respect to β at a fixed value of γ are respectively,

$$U\{\beta(\gamma)\} = \sum_{k=1}^{K} w_k \sum_{f=1}^{n_k} U_f\{\beta(\gamma)\} = \sum_{k=1}^{K} w_k \sum_{f=1}^{n_k} \frac{\partial \ell_f(y;\theta)}{\partial \beta}\,,$$

where

$$\begin{aligned} U_f\{\beta(\gamma)\} = & \sum_{i\in S^2_{kf}} (\delta_i - \nu_i) x_i \\ & + \sum_{i\in S^2_{kf}} (\nu_i - 1)\{\lambda(t_i - t_0)\}^{\rho} e^{x_i^\top \beta} x_i \\ & + \sum_{i\in S^2_{kf}} \frac{\nu_i \{\lambda(t_i - t_0)\}^{\rho} e^{x_i^\top \beta} x_i e^{-\{\lambda(t_i-t_0)\}^{\rho} e^{x_i^\top \beta}}}{1 - e^{-\{\lambda(t_i-t_0)\}^{\rho} e^{x_i^\top \beta}}} \\ & - \frac{\{\lambda(a_p - t_0)\}^{\rho} e^{x_p^\top \beta} x_p e^{-\{\lambda(a_p-t_0)\}^{\rho} e^{x_p^\top \beta}}}{1 - e^{-\{\lambda(a_p-t_0)\}^{\rho} e^{x_p^\top \beta}}}, \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 \ell(\beta(\gamma))}{\partial \beta \partial \beta^\top} = & \sum_{k=1}^{K} w_k \sum_{f=1}^{n_k} \frac{\partial U_f\{\beta(\gamma)\}}{\partial \beta}\,, \\ \frac{\partial U_f\{\beta(\gamma)\}}{\partial \beta} = & \sum_{i\in S^2_{kf}} (\nu_i - 1)\{\lambda(t_i - t_0)\}^{\rho} e^{x_i^\top \beta} x_i x_i^\top \\ & + \sum_{i\in S^2_{kf}} \frac{\nu_i A}{[1 - e^{-\{\lambda(t_i-t_0)\}^{\rho} e^{x_i^\top \beta}}]^2} - \frac{B}{[1 - e^{-\{\lambda(a_p-t_0)\}^{\rho} e^{x_p^\top \beta}}]^2}, \end{aligned}$$

$$\begin{aligned} A = & \{\lambda(t_i - t_0)\}^{\rho_0} x_i {x_i}^\top e^{x_i^\top \beta} e^{-\{\lambda(t_i-t_0)\}^{\rho} e^{x_i^\top \beta}} \times \\ & [1 - \{\lambda(t_i - t_0)\}^{\rho} e^{x_i^\top \beta} - e^{-\{\lambda(t_i-t_0)\}^{\rho} e^{x_i^\top \beta}}], \\ B = & \{\lambda(a_p - t_0)\}^{\rho} x_p x_p^\top e^{x_p^\top \beta} e^{-\{\lambda(a_p-t_0)\}^{\rho} e^{x_p^\top \beta}} \times \\ & [1 - \{\lambda(a_p - t_0)\}^{\rho} e^{x_p^\top \beta} - e^{-\{\lambda(a_p-t_0)\}^{\rho} e^{x_p^\top \beta}}]. \end{aligned}$$

Appendix II

We show that the score function of our proposed composite likelihood is unbiased under the usual regularity conditions. Therefore, the maximum composite likelihood estimator is still consistent for the true parameter.

We rewrite our log ascertainment-corrected likelihood function derived from (3.2) into two parts for non-probands and probands. Let i index the non-probands and p index the probands. Then the log composite likelihood function for the observed data y can be written, assuming y's are independent given genotypes G, as

$$\ell(\theta; y) = \sum_i w_i \left\{\delta_i \log f(y_i|G_i) + (1-\delta_i) \log S(y_i|G_i)\right\} + \sum_p w_p \log \left\{\frac{f(a_p|G_p)}{F(a_p|G_p)}\right\},$$

where $f(y|G), F(y|G)$, and $S(y|G)$ are the density, cumulative density, and survivor functions of response time y given genotype G, respectively, w represents the weight and a_p represents the age at examination for proband as probands were affected before their ages at examination.

We let $U(\theta; y) = \partial\ell(\theta; y)/\partial\theta$ be the score statistic based on the observed data y. In the presence of missing genotypes, the conditional expectation of the score statistic is obtained given the observed data y, as

$$\begin{aligned} E_\theta[U(\theta; y)] &= E_\theta\left[\frac{\partial}{\partial\theta}\ell(\theta; y)\right] \\ &= \sum_i \left\{w_i \delta_i E_\theta\left[\frac{\partial}{\partial\theta}\log f(y_i|G_i)\right] + w_i(1-\delta_i)E_\theta\left[\frac{\partial}{\partial\theta}\log S(y_i|G_i)\right]\right\} \\ &\quad + \sum_p w_p \left\{E_\theta\left[\frac{\partial}{\partial\theta}\log f(a_p|G_p)\right] - E_\theta\left[\frac{\partial}{\partial\theta}\log F(a_p|G_p)\right]\right\}. \end{aligned}$$

We show that $E_\theta[U(\theta; y)] = 0$ at the composite MLE using the facts that

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial\theta}\log f(y_i|G_i)\right] &= 0, \\ E_\theta\left[\frac{\partial}{\partial\theta}\log f(Y_i|G_i)|Y_i > z_i\right] &= \int_{z_i}^{\infty} \frac{\frac{\partial}{\partial\theta}f(y_i|G)}{f(y_i|G)}\frac{f(y_i|G_i)}{S(z_i|G_i)}dy_i = \frac{\frac{\partial}{\partial\theta}S(z_i|G_i)}{S(z_i|G_i)} \\ &= \frac{\partial}{\partial\theta}\log S(z_i|G_i), \\ E_\theta\left[\frac{\partial}{\partial\theta}\log S(z_i|G_i)\right] &= 0, \\ E_\theta\left[\frac{\partial}{\partial\theta}\log f(y_p|G_p)|Y_p < a_p\right] &= \int_{\infty}^{a_p} \frac{\frac{\partial}{\partial\theta}f(y_p|G_p)}{f(y_p|G_p)}\frac{f(y_p|G_p)}{F(a_p|G_p)}dy_p = \frac{\frac{\partial}{\partial\theta}F(a_p|G_p)}{F(a_p|G_p)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial\theta}\log F(a_p|G_p),\\
E_\theta\left[E_\theta\left[\frac{\partial}{\partial\theta}\log f(y_p|G_p)|Y_p < a_p\right]\right] &= E_\theta\left[\frac{\partial}{\partial\theta}\log F(a_p|G_p)\right],\\
E_\theta\left[\frac{\partial}{\partial\theta}\log f(y_p|G_p)\right] &= E_\theta\left[\frac{\partial}{\partial\theta}\log F(a_p|G_p)\right].
\end{aligned}
$$

The results follow on interchanging the operations of expectation and differentiation. It is assumed in the above that regularity conditions hold for this interchange.

References

Andersen, E. W. (2004). Composite likelihood and two-stage estimation in family studies. *Biometrics* **5**, 15-30.

Andersen, E. W. (2005). Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis* **11**, 333-350.

Antoniou, A. C. and Chevenix-Trench, G. (2010). Common genetic variants and cancer risk in Mendelian cancer syndromes. *Curr Opin Genet Dev* **20**, 299-307.

Begg, C. B. (2002). On the use of familial aggregation in population-based case probands for calculating penetrance. *J Natl Cancer Inst.* **94**, 221-6.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139-147.

Boudreau, C. and Lawless, J. F. (2006). Survival analysis based on proportional hazards model and survey data. *Canad. J. Statist.* **34**, 203-216.

Choi, Y.-H. and Briollais, L. (2010). A frailty model-based approach to estimating the age-dependent penetrance function of a gene mutation using family-based study designs. In *JSM Proceedings*, Biopharmaceutical Section. American Statistical Association, Alexandria, VA.

Choi, Y.-H., Kopciuk, K. A. and Briollais, L. (2008). Estimating disease risk associated with mutated genes in family-based designs. *Hum. Hered.* **66**, 238-251.

Choi, Y.-H. and Matthews, D. E. (2005). Accelerated life regression modeling of dependent bivariate time-to-event data. *Canad. J. Statist.* **33**, 449-464.

Cox, D. R. (1972). Regression models and life tables (with discussions). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Gao, X. and Song, P. X.-K. (2009). Composite likelihood EM algorithm with applications to multivariate hidden Markov model. *COBRA Preprint Series.* http://biostats.bepress.com/cobra/ps/art61

John, E. M., Hopper, J. L., Beck, J. C., Knight, J. A., Neuhausen, S. L., Senie, R. T., Ziogas, A., Andrulis, I. L., Anton-Culver, H., Boyd, N., Buys, S. S., Daly, M. B., O'Malley, F. P., Santella, R. M., Southey, M. C., Venne, V. L., Venter, D. J., West, D. W. , Whittemore, A. S., Seminara, D. and Breast Cancer Family Registry. (2004). The breast cancer family registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res.* **6**, R375-89.

Kalbfleisch, J. D. and Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statist. Med.* **7**, 149-160.

Lawless, J. F., Kalbfleisch, J. D. and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. Roy. Statist. Soc. Ser. B* **61**, 413-438.

Lin, D. Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* 87, 37-47.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 221-239.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Louis, T. A. (1982). Finding observed information matrix using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 226-233.

Siegmund, K., Whittemore, A. S. and Thomas, D. C. (1999). Multistage sampling for disease family registries. *J. Natl. Cancer Inst. Monogr.* **26**, 43-8.

Varin, C. (2008). On composite marginal likelihoods. *Adv. Stat. Anal.* **92**, 1-28.

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

Whittemore, A. S. and Halpern, J. (1997). Multi-stage sampling designs in genetic epidemiology. *Statist. Medicine* **16**, 153-167.

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

Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada.

E-mail: Yun-Hee.Choi@schulich.uwo.ca

Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada

Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada.

E-mail: laurent@lunenfeld.ca

(Received November 2009; accepted September 2010)