Mathematical Statistics
Stockholm University

A Stochastic EM type algorithm for estimation in data with ascertainment on continuous outcomes

Maria Grünewald and Keith Humphreys

Research Report 2008:5

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2008:5**,
http://www.math.su.se

A Stochastic EM type algorithm for estimation in data with ascertainment on continuous outcomes

Maria Grünewald and Keith Humphreys

April 2008

Abstract

Outcome dependent sampling probabilities can be used to increase efficiency in observational studies. For continuous outcomes appropriate consideration of sampling design in estimating parameters of interest is often computationally cumbersome. In this article we suggest a Stochastic EM type algorithm for estimation. The computational complexity of the likelihood is avoided by filling in missing data so that the full data likelihood can be used. The method is not restricted to any specific distribution of the data and can be used for a broad range of statistical models.

KEY WORDS: Ascertainment, Stochastic EM algorithm, Missing data, Sequential design, Choice-based sampling, Outcome dependent sampling, Genetic epidemiology

Grünewald: Mathematical statistics, Department of Mathematics, Stockholm University, SE-10691 Stockholm, Sweden. mariag@math.su.se. Humphreys: Department of Medical Epidemiology and Biostatistics, Karolinska Institute, PO Box 281, SE-171 77 Stockholm, Sweden

1 Introduction

Most standard statistical tools for analyzing data from observational studies assume that simple random sampling is used. Outcome dependent sampling may however increase study efficiency. The case-control design (Breslow 1982), for example, has been widely used in epidemiology. An attractive feature of the design is that unbiased estimates of relative risks can be obtained by performing statistical analysis on the data using a logistic regression model, as if the data were from a prospective study. More complex sampling designs may further increase efficiency. In the two-stage case-control design some covariate information is recorded on all subjects included in a study (Stage 1) whilst other covariate information, e.g. more expensive covariates, is gathered only on a subset of samples (Stage 2); the probability that the subject is included in Stage 2 is dependent on Stage 1 covariates. There is a large literature dealing with how to analyze outcome dependent, and two-stage, samples when the outcome is categorical, using a pseudo or semi-parametric likelihood (Breslow & Cain 1988, Breslow & Holubkov 1997, Breslow & Chatterjee 1999, Chen 2003). In general, there is less written about how to deal with continuous outcomes, although some of the above mentioned literature does cover the topic. Dichotomizing continuous outcomes and analyzing the data as if it were a case-control outcome will generally lead to loss in efficiency (Vargha et al. 1996).

Outcome dependent sampling based on continuous outcomes is common in genetic epidemiology. Klos & Kullo (2007), for example, compare candidate gene sequences between individuals in the tails (upper and lower 5%) of the high-density lipoprotein cholesterol distributions of different study populations. An ongoing study at the second author's institute is based on a similar study design, where individuals in the upper and lower tertiles of cholesterol distributions are selected from a cohort study of 60 year old men in Stockholm (www.biobanks.se/cardiovascular.htm), for genotyping. This particular study has a *two-stage cohort* design and hence the study base is clearly defined. For such designs outcome variables Y are known for the entire cohort sample – unbiased estimation of regression parameters is possible and computationally straightforward via application of the EM algorithm. Often study bases are instead ill-defined, e.g. hospital-based studies; it is these study designs which we focus on in this article. For example: in the genetic association study of type II diabetes described by Gu et al. (2004), genotyping was carried out for 106 type II diabetes patients, 325 impaired glucose tolerance patients and 497 normal glucose tolerance controls. Analysis were performed to determine genetic association with continuous metabolic syndrome variables, which were used to define the selection categories. Parameter estimation in regression models with continuous outcomes measured in such samples, i.e. obtained under outcome dependent sampling, will be biased unless the ascertainment scheme is accounted for.

In what follows we will use X to denote explanatory variables, Y to denote outcome variables and A (*ascertainment*) to represent a sampling indicator variable signifying that both X and Y are observed. Both X and Y are allowed to be multivariate. For notational simplicity both probability density and mass functions will be denoted by $P(.)$. We assume that the distribution of X, and of Y conditional on X, is parameterized by θ and that the probability of ascertainment is independent of θ given the observed data, that is $P(A = 1|X, Y, \theta) = P(A = 1|X, Y)$. Furthermore, we assume that ascertainment is independent of X conditional on Y, $P(A = 1|X, Y) = P(A = 1|Y)$. Interest is in estimating θ.

Despite there being several solutions to several specific scenarios, the general problem of parameter estimation under complex ascertainment remains of interest (Clayton 2003). The approaches considered in this article are general in the sense that they are not restricted to a particular design. One general way to account for ascertainment is to inflate the data to a representative sample. In this spirit we describe a novel approach for parameter estimation with ascertainment on continuous outcomes Y, which we note can be easily extended to allow for ascertainment to depend on X as well as Y. This algorithm is similar to a Stochastic EM (SEM) algorithm.

Another general approach for correcting for ascertainment is to base inference on the joint likelihood of the data $Z = (X, Y)$, conditioned on A,

$$L(\theta; Z) = P(X, Y|\theta, A = 1) = \frac{P(A = 1|Y)P(X, Y|\theta)}{P(A = 1|\theta)} \tag{1.1}$$

which corresponds to the log likelihood

$$\log(L) \propto \log(P(X, Y|\theta)) - \log(P(A = 1|\theta)).$$

The form of this likelihood generally makes standard likelihood-based estimation computationally difficult. The computational problem arises because

$$P(A = 1|\theta) = \int_Y \int_X P(A = 1|Y)P(X, Y|\theta)dXdY \tag{1.2}$$

is, in many settings, intractable. For continuous Y some examples of such settings are: if X is continuous or a mixture of discrete and continuous variables, or if ascertainment probability is a continuous function of Y, which could be the case in size biased sampling. See Patil (2002) for a description of size biased sampling. When analytical solutions are not available, methods for numerical integration, such as importance sampling, can be used. This approach is investigated here as a comparison to the SEM type algorithm.

An alternative to basing inference on (1.1) is to use the retrospective likelihood $P(X|Y, A)$, utilising the fact that ascertainment probabilities cancel out of the likelihood when $A \amalg X|Y$, leaving $P(X|Y, A) = P(X|Y)$. However, because of the loss of information in conditioning on the non-ancillary statistic Y, the set of parameters describing the relationship between X and Y is generally not identifiable (Liang 1983). Some, but not all of the parameters may however be identifiable. See (Chen 2003) for a discussion of parameter identification in the general odds ratio function.

There are also alternative solutions for particular designs. An attractive approach to estimation for study designs where samples are drawn with different probabilities in different regions of the space of a continuous outcome, Y, is described by Zhou, Haibo et al. (2002). They describe a semi-parametric empirical likelihood approach to analyze data that consists of both a simple random sample and supplement samples from strata that are presumed highly informative based on their values of Y. Features of the approach are that no parametric assumptions are required for covariates and that ascertainment probabilities are not required to be known or estimated. Often it is advantageous not to make parametric assumptions for covariates, although in genetics it can be advantageous (Chen & Chatterjee 2007).

We will first, in Section 2.1, present the SEM type algorithm for use for the ascertainment problem as described above. Two other approaches, a data augmentation method due to Clayton (2003), and a method based on importance sampling are presented in Section 2.2 for comparison. Some examples are presented in Section 3 and an analysis of simulated data using the three different methods is presented in Section 4. The results are discussed in Section 5.

2 Methods

2.1 A SEM type algorithm

Although not fitting fully into the classical framework of missing data problems (Little & Rubin 1987), non-random ascertainment can still be viewed as a missing data problem, with data missing at random (MAR). In missing data problems data is partitioned into observed data, Z^{obs}, and missing data Z^{mis}, and there is typically a well-defined set of subjects for which some variables have missing values, Z^{mis}, while there is partially complete information, Z^{obs}, on each sample unit. In our setting all variables can be viewed as belonging to Z^{mis} when a subject is not ascertained. Nevertheless it is useful to consider algorithms used in missing data problems, such as the Estimation Maximization (EM) algorithm (Dempster et al. 1977) and it's extensions. Wacholder & Weinberg (1994) used the EM algorithm to obtain Maximum Likelihood estimates in case-control studies with complex ascertainment. This approach would encounter computational difficulties if extended to for example the normal distribution. We describe here a simulation based approach, similar to the SEM algorithm, which overcomes this problem. We begin by summarising the SEM algorithm (Section 2.1.1) as background.

2.1.1 The SEM algorithm

Suppose interest is in estimating θ in a parametric model for a complete data likelihood, when data is actually incomplete. The EM algorithm has been used extensively for maximum likelihood estimation in this setting. First starting values for the parameter estimates are chosen, and then the following two steps are iterated: In the E-step the expectation of the complete data is calculated using the parameter values, $\hat{\theta}$, from the previous M-step (or using starting values at the first iteration). In the M-step the maximum likelihood estimates, $\hat{\theta}$, from the complete data, created in the E-step, are calculated. Although in general, if a likelihood has a unique maximum, the EM algorithm converges to that value (Wu 1983), in practice the EM algorithm is prone to convergence to local maxima. The algorithm is therefore sensitive to choice of starting values.

If calculating the expected complete data likelihood in the E-step requires computationally demanding numerical integration one way to side-step the problem is to simulate the missing data, and use the value of the observed mean instead of the calculated expectation. This is the Monte Carlo EM algorithm (Wei & Tanner 1990). Ibrahim et al. (1999) discuss how the Monte Carlo EM algorithm can be used in generalized linear models when data has missing covariates. The algorithm is performed in two steps; in the S-step the missing data is simulated M times and in the M-step maximum likelihood estimates $\hat{\theta}$ are calculated using the combined data set containing observed and simulated data.

The SEM algorithm (Celeux & Diebolt 1985) is a special case of the Monte Carlo EM algorithm with only one simulation step per maximization step (McLachlan & Krishnan 1997). In iteration i the calculations are performed according to the following algorithm:

S-step: Simulate $M = 1$ set of the missing data, Z^{mis}, using current parameter estimates $\hat{\theta}_{i-1}$.

↓

Construct the complete data likelihood using the observed data, Z^{obs}, and the simulated data, Z^{sim}:

$$\begin{aligned} L(\theta; Z^{com}) &= L(\theta; Z^{obs}_{\in A=1}, Z^{mis}_{\notin A=1}) \\ &\approx L(\theta; Z^{obs}_{\in A=1}, Z^{sim}_{\notin A=1}) \\ &= \prod P(Z^{obs}_{\in A=1}, Z^{sim}_{\notin A=1}|\theta) \\ &= \prod P(Z^{obs}_{\in A=1}|\theta) \prod P(Z^{sim}_{\notin A=1}|\theta) \end{aligned} \tag{2.1}$$

↓

M-step: Obtain new parameter estimates $\hat{\theta}_i$ from (2.1) by maximising the expected complete data likelihood.

↓

Repeat: Go to iteration $i+1$ and repeat the steps above.

Application of the SEM algorithm does not result in a single value for a parameter estimate. Instead there is built-in variation, induced by the simulated data, around the estimate, and the result will be similar to that of a stationary Markov Chain Monte Carlo estimator (Gilks et al. 1996). We will use the word convergence to denote convergence in distribution of the sequence of estimates. In analogue with the terminology of Markov Chain Monte Carlo we will use the word *burn-in* to refer to the initial iterations of the chain that should be excluded from analysis in order to ensure that the estimates are produced by the right distribution.

The SEM algorithm has been shown to be useful in a wide range of missing data problems such as time-to-event data with censoring data sets (Ip 1994) and haplotype estimation (Tregouet et al. 2004). The SEM algorithm is known to be more robust to poorly specified starting values than the deterministic EM algorithm (Gilks et al. 1996), which is a highly attractive feature in our setting. One way of estimating parameters in the SEM algorithm is to choose the set of parameter values in the iteration that gives the highest value of the likelihood for the observed data. The likelihood for the observed data may however be so complicated that this is infeasible. A simpler approach is to compute the mean, $\tilde{\theta}$, of the parameter values in the iterations after an appropriate burn-in period. An approximation of the variance of $\tilde{\theta}$ can, according to Gilks et al. (1996), be computed by utilizing the property that the observed data likelihood in the EM algorithm can be specified in terms of the complete data likelihood (Louis 1982), but replacing the theoretical mean and variance with bootstrap estimates (Efron 1992). The bootstrap estimates are obtained as follows: Fill in the missing data with simulated data K times, using $\tilde{\theta}$, to obtain pseudo complete data sets $z_1^{com}, z_2^{com} \dots z_K^{com}$. If the complete data log likelihood is denoted l^{com}, the observed information is

$$-l''^{obs}(\theta, z) = E_\theta[-l''^{com}(\theta, z)] - Cov_\theta[l'^{com}(\theta, z)], \qquad (2.2)$$

where the expectation and covariance are calculated over the K pseudo samples. The covariance matrix is then obtained by taking the inverse of the information matrix as usual. Some details of the variance calculations will be provided in Section 2.1.2.

For further reading on the SEM algorithm see Gilks et al. (1996) and McLachlan & Krishnan (1997).

2.1.2 Applying a SEM type algorithm to the ascertainment problem

We can implement an algorithm similar in spirit to the SEM algorithm for the ascertainment problem. The essential idea is to inflate the ascertained sample to being representative, using simulated observations. The "non-ascertained" component is considered missing and is imputed in an S-step using the parameter estimates from the most recent M-step in a "SEM type" algorithm. Normally when the SEM algorithm is used to fill in missing data there is a fixed sample size and data is filled in for those observations where data is missing. Here we assume that the sample size of the representative data is not known, as it is for the two-stage cohort design, but that only the ascertainment probabilities conditional on data, and the sample size, n^{obs}, of the observed ascertained data, $Z^{obs}_{\in A=1}$. Data is simulated as described below. In the S-step the missing data is filled in by rejection sampling (see for example Gilks et al. 1996), using a reverse ascertainment scheme:

Simulate: Simulate data from the population distribution $P(Z|\hat{\theta})$ and sort the observations into data that would have been ascertained, $Z^{sim}_{\in A=1}$, and data that would not have been ascertained, $Z^{sim}_{\notin A=1}$. Stop when n^{obs}observations from $Z^{sim}_{\in A=1}$ have been obtained.

↓

Reject: Throw out the observations in $Z^{sim}_{\in A=1}$ and keep those in $Z^{sim}_{\notin A=1}$.

The size of the simulated data-set is random, with distribution NegBin(n^{obs}, P), where $P = P_{\hat{\theta}_{i-1}}(A)$. This implies that $E(n^{sim}) = n^{obs}(P^{-1} - 1)$ and $Var(n^{sim}) = n^{obs}(1 - P)/P^2$. In the M-step maximum likelihood estimates are obtained from the likelihood of the real ascertained data combined with the simulated non-ascertained data as described above.

The information matrix is estimated using (2.2), with quantities on the r.h.s. of the equation estimated as

$$\begin{aligned}
\hat{E}_{\theta}[-l''^{com}(\hat{\theta}, z^{com})] &= \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k^{com}} -l''^{com}(\hat{\theta}, z_{ik}^{com}), \\
\widehat{Cov_{\theta}}[l'^{com}(\hat{\theta}, z^{com})] &= \frac{1}{K-1}\sum_{k=1}^{K} S_k^T S_k - \frac{1}{K-1}(\sum_{k=1}^{K} S_k)^T \frac{1}{K-1}\sum_{k=1}^{K} S_k
\end{aligned}$$

respectively, where

$$S_k = \sum_{i=1}^{n_k^{com}} l'^{com}(\hat{\theta}, z_{ik}^{com}),$$

and where z_{ik}^{com} is the i:th element of the pseudo complete data-set generated in step k.

2.2 Importance sampling and Data augmentation

We now summarize two alternative strategies to obtain parameter estimates in data with non-random ascertainment. Both approaches, in common with the SEM type algorithm, are simulation based and use Maximum Likelihood for estimation.

Importance sampling

As mentioned above the difficulty in calculating the likelihood of the ascertained data lies in the integration of (1.2). Importance sampling (Hammersly & Handscomb 1964) is a Monte Carlo method used for numerical integration. The basic idea is to sample from one distribution to obtain the expectation of another. This is advantageous for sampling efficiently but also when drawing samples from the target distribution is difficult. In general terms, for a random variable X which has density $f_1(x)$, the expectation of some function of X, $g(x)$, can be written as

$$\mu = E_{f1}[g(x)] = \int g(x) f_1 dx = \int \frac{f_1}{f_2} g(x) f_2 dx = E_{f_2}[\frac{f_1}{f_2} g(x)]$$

for $f_2 > 0$ whenever the support of f_2 includes that of $gf_1 > 0$. This means that samples can be drawn from f_2 to obtain the expectation of $g(x)$. We can apply the importance sampling technique to approximate (1.2). One way to implement importance sampling in this context is to draw observations from a distribution which has the same parametric form as the target distribution $P(z|\theta)$, but in the place of the unknown θ, use naive guesses of the values of θ, which we call θ^*, in analogy with the starting values for the SEM type algorithm. In this case $P(A = 1|\theta)$ is estimated by noting that

$$P(A = 1|\theta) = \int P(A = 1|z)P(z|\theta)dz = \int [P(A = 1|z, \theta)\frac{P(z|\theta)}{P(z|\theta^*)}]P(z|\theta^*)dz.$$

If we draw $\dot{M}$ observations from $P(z|\theta^*)$ which we denote as $z_1^{sim}, \ldots, z_{\dot{M}}^{sim}$, we can estimate $P(A = 1|\theta)$ by

$$\hat{P}(A = 1|\theta) = \frac{1}{\dot{M}} \sum_{j=1}^{\dot{M}} P(A = 1|z_j^{sim}) \frac{P(z_j^{sim}|\theta)}{P(z_j^{sim}|\theta^*)}. \tag{2.3}$$

As a consequence we can approximate the log likelihood contribution of individual i,

$$\log(L) \propto \log(P(z_i|\theta)) - \log(P(A = 1|\theta)),$$

up to a constant, by replacing $P(A = 1|\theta)$ by (2.3), thereby obtaining

$$\log(P(z_i|\theta)) - \log(\frac{1}{\dot{M}} \sum_{j=1}^{\dot{M}} P(A = 1|z_j^{sim}) \frac{P(z_j^{sim}|\theta)}{P(z_j^{sim}|\theta^*)}). \tag{2.4}$$

Since the approximation of the likelihood is expressed in terms of θ an approximation of the information matrix can be computed as minus the second derivative of the log likelihood as usual.

Data augmentation

Clayton (2003) derives an ascertainment corrected likelihood by using an analogy to the conditional likelihood for matched case-control data. The idea behind this approach is to simulate a number of *pseudo-observations* for each real observation and use these in combination with the real data to build the likelihood. As in the importance sampling method the true parameter values θ are unknown and are substituted by guesses, θ^*.

Given the pseudo-observations the log likelihood contribution of individual i can, up to a constant, be written as

$$\log(P(z_i|\theta)) - \log(\frac{1}{\ddot{M}+1}\sum_{j=1}^{\ddot{M}+1}\frac{P(z_{ij}|\theta)}{P(z_{ij}|\theta^*)}). \tag{2.5}$$

Since an expression for the likelihood is available, parameter estimates can be obtained using maximum likelihood. Variances of these estimates are obtained as usual by calculating the information matrix from the likelihood. The likelihood (2.5) is similar to the likelihood approximated with the importance sampler, (2.4), especially when ascertainment probabilities are 0/1. The essential differences are that

- Data is drawn under non-random ascertainment in (2.5), using the data augmentation method, while it was drawn from the population distribution in (2.4), using the importance sampler.
- The sum in the second term is over the pseudo-observations only in (2.4) while the real observation are also included in (2.5).
- In (2.5) a separate estimate of $P(A = 1)$ is calculated for each real observation while in (2.4) $P(A = 1)$ is calculated only once.

The last of these differences means that while $\dot{M}$ pseudo-observations are produced in the importance sampler, for a sample size of n^{obs} real observations, $\ddot{M} \times n^{obs}$ pseudo-observations are produced in the data augmentation method.

3 Examples

To illustrate the performance of the methods described above perform we will look at two examples. The example data are simulated to allow comparison of the results with true answers. Starting values/parameter guesses,

θ^*, are required for each method. The first example is based on a univariate continuous outcome with non-random ascertainment and is used to illustrate how sensitive, mainly with respect to variability, the different methods are to specification of starting values. The second simulation is based on a more complex example with a multivariate outcome and non-random ascertainment. For this example we found that our method provides valid parameter estimates while the other approaches fail when θ^* is poorly specified. Both examples are based on a single explanatory variable X.

The simulations are inspired by genetic epidemiology, where non-random ascertainment is widely used for the reason that genetic data has traditionally been more expensive to collect than response variable measurements. In particular the outcomes are thought of as traits representing the metabolic syndrome (Agardh et al. 2003). The metabolic syndrome comprises many health related outcomes that can affect each other in complex ways. In our simulation studies we represent only simplified models of the metabolic syndrome, using outcome variables only to represent BMI and plasma glucose level.

In the examples ascertainment probabilities are assumed known. In reality, these quantities will usually have to be either estimated from data, or inferred from external sources, adding an extra source of uncertainty that has not been taken into account here.

3.1 Simulation model i

For the first simulation model we use a single categorical covariate, X, which takes possible values 0,1,2. The model is based on a genetics example. X represents the genotypes (AA, Aa, aa) of a single nucleotide polymorphism (SNP) with alleles A and a and a minor allele frequency of $\exp(\beta_{0X})/(1 + \exp(\beta_{0X})) \approx 0.2$ $(\beta_{0X} = -1.4)$, so that genotypes AA, Aa and aa have approximate population frequencies 0.64, 0.32 and 0.04. The distribution of the univariate outcome, conditional on $X = x$ is Gaussian with mean $\beta_{0Y} + \beta_{XY} \times x$ and variance σ_Y^2. We use values $\beta_{0Y} = 24, \beta_{XY} = 4$ and $\sigma_Y = \sqrt{2}$, chosen so that Y loosely represents BMI. Individuals with a BMI of 30 or more are defined as *obese* (as according to the WHO definition). About 10 percent of the Swedish population in the ages of 25-64 have such a BMI according to the WHO MONICA project (Tolonen et al. 2000). Ascertainment probabilities are dependent on outcome/phenotype values:

$P(A|y \geq 30) = 1$, $P(A|y < 30) = 0.067$, giving approximately equal numbers of obese and non-obese subjects. In this example, due to the simplicity of the model, evaluation of the integral (1.2) would actually be straightforward. Sampling is continued until the total sample size is 300. This sampling procedure is similar to the one used by Gu et al. (2004). The difference is that in our simulation subgroup sample sizes are not fixed, whereas in Gu et al. (2004) they are. In simulating data we generate samples with random subgroup sample sizes, since this corresponds directly to the way the data is analyzed. The asymptotic equivalence of estimators, whether the subgroup sample sizes are regarded as fixed or random, has been discussed by Breslow et al. (2000) in the case-control setting. The simulation model can be represented graphically as

$$X(\text{Genotype}) \rightarrow Y(\text{BMI}) \rightarrow A \ (\text{Ascertainment}).$$

3.2 Simulation model *ii*

In the second simulation we again assume a single categorical covariate, X, representing a SNP genotype. Instead of a single outcome as in Model *i*, we here use two, Y_1 and Y_2, considered to represent BMI and plasma glucose level, respectively. Obesity, measured in terms of BMI, is a *co-morbid* disease of plasma glucose level; BMI is dependent on genotype and, in turn, affects plasma glucose level. The genotype is assumed to have an additive effect on both outcomes, and Y_1 has an additive effect on Y_2. Given $X = x$, Y_1 has distribution $N(\beta_{0Y_1} + \beta_{XY_1} \times x, \sigma_{Y_1})$ and Y_2, given $X = x$ and $Y_1 = y_1$ has distribution $N(\beta_{0Y_2} + \beta_{XY_2} \times x + \beta_{Y_1Y_2} \times y_1, \sigma_{Y_2})$. Parameter values are chosen to represent outcomes accordingly; $\beta_{0Y_1} = 24$, $\beta_{XY_1} = 4$, $\sigma_{Y_1} = \sqrt{2}$, $\beta_{0Y_2} = 3$, $\beta_{XY_2} = 1$, $\beta_{Y_1Y_2} = 1/15$ and $\sigma_{Y_2} = 0.5$. The ascertainment probability is dependent upon both outcomes, as specified in Table 5.1. Model *ii* can be illustrated graphically as

$$\begin{tikzcd}
 & Y_1 \ (\text{BMI}) \arrow[dd] \arrow[dr] & \\
X \ (\text{Genotype}) \arrow[ur] \arrow[dr] & & A \ (\text{Ascertasinment}). \\
 & Y_2 \ (\text{Plasma Glucose}) \arrow[ur] &
\end{tikzcd}$$

4 Results

All simulations were performed using the software R (The R Development Core Team 2001). Simulations were carried out for both correctly specified ”starting values”, $\theta^* = \theta$, and for misspecified values, $\theta^* \neq \theta$, to investigate how the methods perform under both optimal and sub-optimal circumstances. In the analysis below $\ddot{M} = 50$ is used in the data augmentation method and $\dot{M} = 30000$ is used in the importance sampler. For a more detailed discussion on the choice of $\ddot{M}$ and $\dot{M}$, see Grünewald (2004).

4.1 Results for Model i

Parameter estimates and variance estimates when $\theta^* = \theta$

Estimates calculated using $\theta^* = \theta$ are presented in Table 5.2. Naive estimates, calculated by optimizing the likelihood of the data without ascertainment correction, are also presented. Standard errors of the means of the estimates are presented in parentheses. We calculated the variance estimate of Gilks et al. (1996) based on (2.2) for each of the 100 simulations, using $K = 5000$. The means of these standard errors across the 100 simulations are presented in Table 5.3. For comparison standard errors based on observed variability in simulations are also presented. The standard errors calculated using (2.2) were also used to construct 95% confidence intervals around the estimates obtained using the SEM type algorithm.

The three methods all provide estimates which are appropriately corrected for ascertainment, while naive estimates are biased. A comparison with the observed standard errors for the SEM type algorithm, indicates that the standard errors calculated using (2.2) give appropriate approximations of the variation in the estimates. Of the 95% confidence intervals constructed for the SEM type algorithm, the empirical coverage probabilities based on 100 simulations were 0.94, 0.95, 0.96 and 0.96 for β_{0X}, β_{0Y}, β_{XY} and σ_Y respectively.

It is worth noting that the Gilks et al. (1996) method of calculating standard errors does not take into account the chain length of the SEM, so it is advisable to run a long chain to avoid variability that is unaccounted for. The chain length in Model i was 2000. It is also worth noting that any observed differences in variability of estimates between methods in Table 5.2

should be interpreted with caution since the variance of estimates based on the data augmentation method and the importance sampler depend on $\ddot{M}$ and $\dot{M}$, respectively, and the variance of estimates based on using the SEM type algorithm depend on chain length.

Parameter estimation when θ^* is poorly specified

To investigate the behavior of the methods under poorly specified θ^* simulations were run for $\beta^*_{XY} = 0$ and 2, while remaining starting values were specified as their true parameter value counterparts. The running time of the SEM type algorithm was longer when θ^* was misspecified, to allow for convergence. As for Markov Chain Monte Carlo simulations, an appropriate burn in period has to be identified. When the algorithm has converged to a distribution around the parameter estimates, the standard errors of the estimates after burn in are the same as for correctly specified starting values. In our simulations the SEM type algorithm always converged and gave the same parameter estimates for poorly specified θ^* as for correctly specified θ^*, as presented in the right-hand column of Table 5.2.

When the data augmentation method was run with poorly specified θ^* parameter estimates were unbiased but had large standard errors, as can be seen in Table 5.5. As Clayton (2003) suggests, running a moderate amount of iterations of the data augmentation method improves the performance when θ^* is poorly specified. That is, the standard error estimates become smaller, approaching values that would be obtained if true/population parameter values were used as starting values.

In Table 5.4 parameter estimates and standard errors of mean estimates for the importance sampler are presented. The importance sampler yields incorrect parameter estimates. For example, the mean estimate of $\hat{\beta}_{0X}$ was 3.213, with a standard error of 0.208; for a true value of -1.4. The inflation of the standard errors appears to be more pronounced in the importance sampler than in the data augmentation method. The effect of the misspecification on the standard errors is not linear. It is worth noting that the standard error of $\hat{\beta}_{XY}$ actually seems to be larger for slightly misspecified β^*_{XY} than for more severely misspecified β^*_{XY}. Since the importance sampler estimator is claimed to be unbiased it may seem surprising that the parameter estimates in the example are biased. A condition for the importance sampler is that the sampling distribution f_2 should be positive whenever $gf_1 > 0$. This condition is fulfilled in the simulations above, but when θ^* is misspecified f_2 may be so small in some regions where gf_1 is large, that no observations are actually sampled. The performance of the method may be improved by a better choice of sampling distribution, for example by using a mixture distribution (Hesterberg 1995).

4.2 Results for Model *ii*

When using $\theta^* = \theta$ for Model *ii* the importance sampler, the data augmentation method and the SEM type algorithm all gave reasonable estimates.

Algorithms were also run for poorly specified θ^*. The value of θ^* was ($\beta^*_{0X} = 0$, $\beta^*_{0Y_1} = \beta_{0Y_1}$, $\beta^*_{XY_1} = 0$, $\sigma^*_{Y_1} = \sigma_{Y_1}$, $\beta^*_{0Y_2} = \beta_{0Y_2}$, $\beta^*_{XY_2} = 0$, $\beta^*_{Y_1Y_2} = 0$, $\sigma^*_{Y_2} = \sigma_{Y_2}$). This value was chosen to investigate the robustness of the methods under extreme misspecification of θ^* in a complex model. As can be seen from Table 5.6 under this value of θ^* neither the data augmentation method nor the importance sampler obtain adequate parameter estimates. To investigate whether iterating the data augmentation method compensates for poorly specified θ^* a few exploratory runs were made. However, even after several iterations, the estimates behaved erratically, and we did not observe convergence towards true parameter values.

The SEM type algorithm did converge to appropriate parameter estimates but took longer to converge than it did in Model *i*. A run of the SEM type algorithm on a single data set is shown in Figure 5.1. Since the algorithm is run on a data-set, which in itself contains some uncertainty, convergence will be seen towards the data parameter values corrected for ascertainment, rather than towards the true/population parameter values.

5 Conclusions

In this paper we have presented an algorithm that can be used to correct for ascertainment. The computational complexity of the likelihood under ascertainment is avoided by filling in missing data so that the full data likelihood can be used. An advantage of the method is that it is not restricted to any specific statistical model -some of the traditional methods to correct for ascertainment handle only specific sampling schemes/statistical models. Also, the complexity of the ascertainment scheme hardly affects the complexity of the calculations, since the ascertainment probabilities are used only when simulating data, and not in the likelihood.

For well specified starting values the SEM type algorithm, as well as the two other methods investigated, perform well. For poorly specified starting values the SEM type algorithm seems to perform better than the other methods, when only a single iteration is used, both with regards to bias and to variability of the estimates.

The SEM type algorithm was slower to run than the other two methods discussed. The speed of the algorithm may be a problem if ascertainment probability is low for some portion of data, since large sets of data will then have to be filled in. An alternative to sampling the whole set of missing data is to simulate only a portion of the data and weigh up the likelihood contribution of the simulated data. Data can for example be simulated as above until n^{obs}/q observations from $Z^{sim}_{\in A=1}$ are obtained. Too small values of n^{obs}/q will however cause too large variability in parameter estimates. Ripatti et al. (2002) suggest a rule for increasing the number of samples in a Monte Carlo EM algorithm when approaching convergence. The basic idea of altering the number of samples when approaching the estimate could be used also in our setting. If the size of the missing data is small it is of course also possible to choose $M > 1$, giving an algorithm similar to the Monte Carlo EM.

The approaches presented here demand prior knowledge of sampling probabilities given the data. These probabilities are often not known and approximations may have to be made using, for example, registry data or prior knowledge about disease occurrence. For ill-defined study designs sensitivity analysis may be informative, with ascertainment probabilities as sensitivity parameters. The results described here may also be sensitive to distributional assumptions, especially if the ascertainment probability is low, so that a large proportion of data is filled in. The outcomes in the examples are assumed to be normally distributed given genotype scores, but since real data often do not follow standard distributions, nonparametric extensions of the method would be of interest. It is not possible to check distributional assumptions using standard procedures such as normal QQ-plots since the ascertained data is not assumed to follow the distribution in the population. When missing data is filled in, checks of distributional assumptions can be misleading since the combined data is a mixture of data from the population distribution and data simulated according to the distributional assumptions. If distributional assumptions are to be checked custom made methods have to be constructed.

Other sampling strategies, e.g. two-stage designs, where some information is retained on all individuals, may in some cases be handled by a slightly modified version of the algorithm. This sampling scheme is on the other hand more similar to the classical missing data setting, with observations missing at random (MAR), and methods such as multiple imputation (Little & Rubin 1987) may be useful.

Funding

K.H. gratefully acknowledges support from the Swedish Research Council (grant number 523-2006-972).

References

Agardh, E., Ahlbom, A., Andersson, T., Efendic, S., Grill, V., Hallqvist, J., Norman, A. & Ostenson, C. (2003), 'Work stress and low sense of coherence is associated with type 2 diabetes in middle-aged Swedish women.', *Diabetes Care* **26**(3), 719–24.

Breslow, N. (1982), 'Design and analysis of case-control studies', *Annual Review of Public Health* **3**(1), 29–5.

Breslow, N. & Cain, K. (1988), 'Logistic regression for two-stage case-control data', *Biometrika* **75**, 11–20.

Breslow, N. E. & Chatterjee, N. (1999), 'Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis', *Applied Statistics* **48**(4), 457–468.

Breslow, N. E. & Holubkov, R. (1997), 'Maximum likelihood estimation of logistic regression parameters under two- phase, outcome-dependent sampling', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**(2), 447–461.

Breslow, N., Robins, J. M. & Wellner, J. (2000), 'On the semi-parametric efficiency of logistic regression under case-control sampling', *Bernoulli* **6**(3), 447–455.

Celeux, G. & Diebolt, J. (1985), 'The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem', *Computational Statistics [Formerly: Computational Statistics Quarterly]* **2**, 73–82.

Chen, H. Y. (2003), 'A note on the prospective analysis of outcome-dependent samples', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65, Part 2**, 575–584.

Chen, J. & Chatterjee, N. (2007), 'Exploiting hardy-weinberg equilibrium for efficient screening of single SNP associations from case-control studies', *Human Heredity* **63**(34), 196–204.

Clayton, D. (2003), 'Conditional likelihood inference under complex ascertainment using data augmentation.', *Biometrika* **90**(4), 976–981.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **39**, 1–37.

Efron, B. (1992), Missing data, imputation and the bootstrap, Technical report, Division of Biostatistics, Stanford University.

Gilks, W. R., Richardson, S. & Speigelhalter, D. J. (1996), *Markov Chain Monte Carlo in practice*, first edn, Chapman & Hall, London.

Grünewald, M. (2004), Genetic association studies with complex ascertainment, Licenciate thesis 2004:5, Stockholm University, Department of Mathematics, Stockhom University, 10691 Stockholm, Sweden.

Gu, H., Abulaiti, A., Ostenson, C., Humphreys, K., Wahlestedt, C., Brookes, A. & Efendic, S. (2004), 'Single nucleotide polymorphisms in the proximal promoter region of the adiponectin (APM1) gene are associated with type 2 diabetes in Swedish caucasians.', *Diabetes* **53**, Suppl 1:S31–5.

Hammersly, J. M. & Handscomb, D. C. (1964), *Monte Carlo methods*, Methuen, London.

Hesterberg, T. (1995), 'Weighted average importance sampling and defensive mixture distributions', *Technometrics* **37**, 185–194.

Ibrahim, J. G., Lipsitz, S. R. & Chen, M.-H. (1999), 'Missing covariates in generalized linear models when the missing data mechanism is non-ignorable', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(1), 173–190.

Ip, E. H. S. (1994), 'A stochastic EM estimator in the presense of missing data -theory and applications.', *Technical report, Department of Statistics, Stanford University* .

Klos, K. L. & Kullo, I. J. (2007), 'Genetic determinants of hdl: monogenic disorders and contributions to variation', *Current Opinion in Cardiology* **22**(4), 344–351.

Liang, K.-Y. (1983), 'On information and ancillarity in the presence of a nuisance parameter', *Biometrika* **70**(3), 607–612.

Little, R. J. A. & Rubin, D. (1987), *Statistical analysis with missing data*, John Wiley & Sons, Hoboken, N.J.

Louis, T. A. (1982), 'Finding the observed information matrix when using the EM algorithm', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **44**(2), 226–233.

McLachlan, G. J. & Krishnan, T. (1997), *The EM algorithm and extensions*, John Wiley & sons Inc, chapter 6.

Patil, G. (2002), 'Weighed distributions', *Encyclopedia of Environmetrics* **4**, 2369–2377.

Ripatti, S., Larsen, K. & Palmgren, J. (2002), 'Maximum likelihood inference inference for multivariate frailty models using an automated monte carlo EM algorithm', *Lifetime Data Analysis* **8**, 349–360.

The R Development Core Team (2001), 'R', Version 1.4.0.

Tolonen, H., Kari, K. & Ruokokoski, E. (2000), 'Monica population survey data book', WWW-publications from the WHO MONICA Project "`http://www.ktl.fi/publications/monica/surveydb/title.htm`".

Tregouet, D., Escolano, S., Tiret, L., Mallet, A. & Golmard, J. L. (2004), 'A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm', *Annals of Human Genetics* **68**(2), 165–177.

Vargha, A., Rudas, T., Delaney, H. D. & Maxwell, S. E. (1996), 'Dichotomization, partial correlation, and conditional independence', *Journal of educational and behavioral statistics* **21**(3), 264–282.

Wacholder, S. & Weinberg, C. R. (1994), 'Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling', *Biometrics* **50**(2), 350–357.

Wei, G. C. G. & Tanner, M. A. (1990), ‘A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms’, *Journal of the American Statistical Association* **85**, 699–704.

Wu, C. (1983), ‘On the convergence properties of the EM algorithm’, *Annals of Statistics* **11**, 95–103.

Zhou, Haibo, Weaver, M. A., Qin, J., Longnecker, M. P. & Wang, M. C. (2002), ‘A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome’, *Biometrics* **58**(2), 413–421.

Figures and tables

	$Y_1 < 30$	$Y_1 \geq 30$
$Y_2 < 7.8$	0.1	0.3
$Y_2 \geq 7.8$	0.3	1

Table 5.1: Ascertainment probabilities in Model *ii*

	True values	Naive estimates	Importance sampling	Data augmentation	SEM type algorithm
$\hat{\beta}_{0X}$	-1.4	-0.107 (0.008)	-1.394 (0.008)	-1.403 (0.006)	-1.400 (0.007)
$\hat{\beta}_{0Y}$	24	24.409 (0.014)	23.996 (0.013)	24.007 (0.011)	24.027 (0.012)
$\hat{\beta}_{XY}$	4	4.152 (0.011)	4.008 (0.012)	3.991 (0.010)	3.993 (0.010)
$\hat{\sigma}_Y$	$\sqrt{(2)} \approx 1.414$	1.589 (0.006)	1.416 (0.005)	1.416 (0.005)	1.409 (0.005)

Table 5.2: Model *i*. Comparison of estimates when $\theta^* = \theta$. Results based on 100 simulations with $n^{obs} = 300$, $\dot{M} = 30000$ and $\ddot{M} = 50$. Standard errors for mean estimates are reported in parentheses.

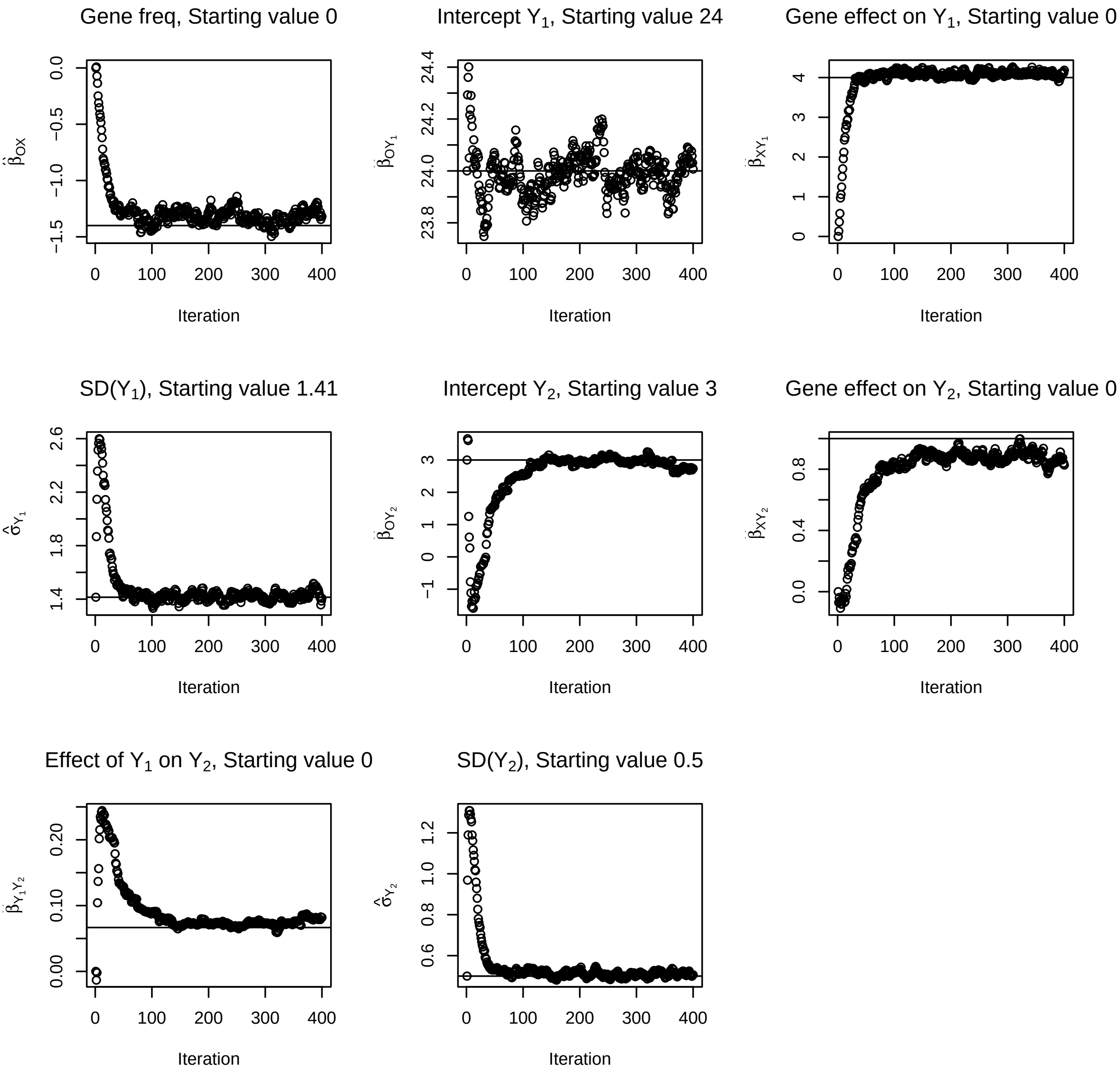

Figure 5.1: Model *ii*. The first 400 iterations in the SEM type algorithm for misspecified θ^*. True parameter values as solid lines. $n^{obs} = 300$.

	Standard errors based on observed variability in simulations	Standard errors calculated using method in Gilks et al. (1996)
$\hat{\beta}_{0X}$	0.072	0.073
$\hat{\beta}_{0Y}$	0.122	0.122
$\hat{\beta}_{XY}$	0.100	0.109
$\hat{\sigma}_Y$	0.047	0.054

Table 5.3: Model i. Comparison of standard errors calculated using method in Gilks et al. (1996) and standard errors reflecting observed variation between simulations. $\theta^* = \theta$. Results based on 100 simulations with $n^{obs} = 300$.

	True	$\beta^*_{XY} = \beta_{XY} = 4$	$\beta^*_{XY} = 2$	$\beta^*_{XY} = 0$
$\hat{\beta}_{0X}$	-1.4	-1.394 (0.008)	-1.189 (0.031)	3.213 (0.208)
$\hat{\beta}_{0Y}$	24	23.996 (0.013)	23.850 (0.021)	26.342 (0.183)
$\hat{\beta}_{XY}$	4	4.008 (0.012)	4.292 (0.035)	4.687 (0.030)
$\hat{\sigma}_Y$	$\sqrt{(2)} \approx 1.414$	1.416 (0.005)	1.401 (0.016)	2.001 (0.037)

Table 5.4: Model i. Importance sampling with β^*_{XY} misspecified, and the remaining parameters at ideal starting values. Results based on 100 simulations with $n^{obs} = 300$ and $\dot{M} = 30000$. Standard errors for mean estimates are reported in parentheses.

	True	$\beta^*_{XY} = \beta_{XY} = 4$	$\beta^*_{XY} = 2$	$\beta^*_{XY} = 0$
$\hat{\beta}_{0X}$	-1.4	-1.403	-1.409	-1.316
		(0.006)	(0.010)	(0.034)
$\hat{\beta}_{0Y}$	24	24.007	24.000	23.991
		(0.011)	(0.013)	(0.013)
$\hat{\beta}_{XY}$	4	3.991	3.984	4.094
		(0.010)	(0.014)	(0.041)
$\hat{\sigma}_Y$	$\sqrt{(2)} \approx 1.414$	1.416	1.392	1.434
		(0.005)	(0.007)	(0.009)

Table 5.5: Model i. Data augmentation method with β^*_{XY} misspecified, and the remaining parameters at ideal starting values. Results based on 100 simulations with $n^{obs} = 300$ and $\ddot{M} = 50$. Standard errors for mean estimates are reported in parentheses.

	True θ	θ^*	Data augmentation	Importance sampling
$\hat{\beta}_{0X}$	-1.4	0	-0.376 (0.105)	-1.001 (0.094)
$\hat{\beta}_{0Y_1}$	24	24	23.654 (0.109)	24.371 (0.090)
$\hat{\beta}_{XY_1}$	4	0	0.304 (0.043)	0.584 (0.102)
$\hat{\sigma}_{Y_1}$	$\sqrt(2) \approx 1.414$	$\sqrt(2)$	1.719 (0.020)	1.704 (0.038)
$\hat{\beta}_{0Y_2}$	3	3	4.825 (0.033)	4.977 (0.052)
$\hat{\beta}_{XY_2}$	1	0	0.557 (0.019)	0.645 (0.039)
$\hat{\beta}_{Y_1Y_2}$	$1/15 \approx 0.067$	0	0.061 (0.003)	0.015 (0.002)
$\hat{\sigma}_{Y_2}$	0.5	0.5	0.064 (0.008)	0.002 (0.004)

Table 5.6: Model *ii*. The data augmentation method and importance sampling under misspecified θ^*. Results based on 100 simulations with $n^{obs} = 300$, $\dot{M} = 30000$ and $\ddot{M} = 50$. Standard errors for mean estimates are reported in parentheses.