Mathematical Statistics
Stockholm University

Efficient ascertainment schemes for maximum likelihood estimation

Maria Grünewald and Ola Hössjer

Research Report 2008:7

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 **2008:7**,
http://www.math.su.se

Efficient ascertainment schemes for maximum likelihood estimation

Maria Grünewald and Ola Hössjer*

June 2008

Abstract

While well chosen sampling schemes may substantially increase efficiency of observational studies, some sampling schemes may instead decrease efficiency. Rules of thumb how to choose sampling schemes are only available for some special cases. In this paper we provide tools to compare efficiencies, and cost adjusted efficiencies, of different sampling schemes, in order to facilitate this choice. The method can be used for both categorical and continuous outcome variables. Some examples are presented, focusing on data from ascertainment sampling schemes. A Monte Carlo method is used to overcome computational issues wherever needed. The results are illustrated in graphs.

KEY WORDS: Ascertainment, Cost adjusted efficiency, Fisher information, Efficient design, Outcome dependent sampling, Multistage design, Continuous outcome variables

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: mariag@math.su.se.

1 Introduction

A well chosen sampling scheme can substantially increase the efficiency of a study. However, it is not always obvious how to sample well. Neyman (1938) presents the possibility of two-stage sampling to increase efficiency in field sampling, and concludes that two-stage sampling sometimes, but not always, reduces the variance of estimates of means. Since then various authors have investigated the effects of two-stage and multistage sampling in different settings, most of which focus on binary outcome variables. In some special cases, such as case-control studies, there are rules of thumb to follow with regards to efficiency, see for example Maydrech & Kupper (1978), but in most other settings more elaborate calculations are necessary to discriminate between different options. Multistage sampling is described in the context of genetic epidemiology by, among others, Whittemore & Halpern (1997): Case-control status of prostate cancer is first ascertained and then more expensive measures such as family history of disease and DNA samples are collected. Asymptotic variances of Horvitz-Thompson estimates are derived. Reilly (1996) investigates optimal allocation of available resources for two-stage data with binary outcomes. Complete information is there available from variables sampled in Stage 1, while Stage 2 variables are sampled more sparsely with probabilities determined by Stage 1 data. Cost is allowed to differ between sampling in Stage 1 and sampling in Stage 2. The author emphasizes the usefulness of pilot studies to obtain information needed to find the optimal allocation. Zhou et al. (2007) investigate outcome dependent sampling where the outcome variable is continuous. Power of tests based on a semi-parametric estimator are compared with the power of an inverse probability weighted estimator and the power of a maximum likelihood estimator based on a simple random sample.

The aim of this paper is to provide tools for comparing sampling schemes with respect to efficiency of maximum likelihood estimates, in order to facilitate study design in observational studies. A general theory of the efficiency calculations for multistage designs is presented in Grünewald & Hössjer (2008), while in this paper we will focus on what is often referred to as "the ascertainment problem". Selection is made on one or more variables, both categorical and continuous distributions are accommodated. The case-control design, where sampling probabilities are determined by the outcome of a binary outcome variable, is a special case of this design. The data can be thought of as originating from a two stage design, where initially, all individuals in the study are sampled for some variables (Stage 1) and then a subset of individuals are sampled for remaining variables (Stage 2). In particular, in an observational study response variables are collected at Stage

1 and explanatory variables (or covariates) at Stage 2. Selection in reverse order, starting with explanatory variables, yields a setting which is similar to experimental design, an area that has been thoroughly investigated, see for example Melas (2006) for an overview.

The ascertained sample arises when there is failure to record data on individuals sampled only in Stage 1, so that only those with data from both Stage 1 and Stage 2 remain in the sample. The ascertainment probability

$$\pi(z) = P(J = 2|Z = z, \theta)$$

is the probability of being sampled at Stage 2, where J denotes sampling stage, Z data, and θ are the model parameters. Data resulting from this scenario would be similar to that which may emerge for example when the study is hospital-based, and the study base is ill defined. External sources may then have to be used to uncover the population distribution of the disease.

The failure to record individuals with incomplete data means that the likelihood will be different from that of the usual two-stage design. A possible way to handle ascertainment, pursued in this paper, is to condition on the ascertainment event. With experimental design, or when selecting on an explanatory variable, the ascertainment event is ancillary when the main goal is to estimate or test hypothesis for response variable parameters, such as the effect of the explanatory variable. Therefore, the ascertained data can be analyzed without correcting for the selection scheme. However, this is an exception in contrast to, for instance, response variable selection, where ascertainment has to be corrected for. Fisher (1934) provides an early example in the context of segregation analysis of ascertainment correction by means of conditioning. The resulting likelihood is expressed in terms of weighted distributions, using weights proportional to ascertainment probabilities. Patil (2002) gives an overview of weighted distributions. They are useful for meta analysis, truncation, missing data, damaged observations, analysis of family data, and when no proper sampling frame is available (Patil & Taillie 1989). In Patil et al. (1973), the efficiency of weighted distributions are compared with that of un-weighted distributions by studying the difference of information matrices.

Another possibility (not pursued here) is to consider the joint likelihood of Stage 2 data and ascertainment, using missing data methodology. Grünewald & Humphreys (2008) use the Stochastic EM-algorithm to evaluate the resulting likelihood. Missing data techniques may also be used for multistage designs without ascertainment. A major distinction from the classical missing data setting is that data is missing by design. Little & Rubin (2002) give a thorough description of how to handle missing data.

We investigate asymptotic efficiencies of parameter estimates of ascertained data sets compared to the corresponding full data sets, where all individuals sampled at Stage 1 are also sampled at Stage 2. For comparison, the efficiencies of prospective and retrospective versions of the ascertained likelihood are also considered, as well as the efficiency of two-stage sample designs. A smaller efficiency may be acceptable if the sampling cost is considerably reduced compared to the full sample. For this reason, a cost-adjusted efficiency is introduced, which gives the efficiency per unit cost of sampling. To quantify the cost-efficiency tradeoff, we plot the efficiency and cost-adjusted efficiency as functions of the ascertainment probability.

This paper is organized as follows: In Section 2 notation to describe data under ascertainment is introduced. Likelihoods under ascertainment, two-stage data and full data are presented in Section 3. Fisher information matrices resulting from these are presented in Section 4, and a Monte Carlo method to overcome computational issues in the calculations is described in Section 5. In Section 6 efficiency, and cost adjusted efficiency, is expressed in terms of the Fisher information. To illustrate the tools presented in the paper some examples are provided in Section 7. Usefulness of the methods, and possible extensions, are discussed in Section 8.

2 Model

Let $Z = (X, Y)$ be a set of random variables, where X are explanatory variables, and Y response variables. Let Z_1 be an incomplete version of Z, representing data collected in Stage 1, and $A = \{J = 2\}$ be the event that data is ascertained, i.e. that all of Z is collected. Then

$$\pi(z_1) = P(A|Z_1 = z_1; \theta) = P(A|Z_1 = z_1)$$

is the selection probability, assumed to depend only on the Stage 1 variable z_1 and not the model parameters θ. We will focus on retrospective designs, where $Z_1 = Y$, so that $\pi(y) = P(A|Y = y)$ is the selection probability. The model is described in Figure 1, where $\theta = (\theta_X, \theta_Y)$ consists of regression parameters θ_Y and remaining parameters θ_X, that affect the distribution of X. Typically θ_Y are the structural parameters of main interest, whereas θ_X are nuisance parameters. The figure corresponds to a two stage design with Y collected at Stage 1 and X at Stage 2. We also assume that no information is available about subjects that were not ascertained at Stage 2.

In the examples below X will be a binomially distributed variable while Y will be either a binomially distributed variable, a normally distributed variable or two interdependent normally distributed variables. The framework is however flexible to other choices of distributions and data structures. We will use $P(.)$ to denote both probability density functions and probability mass functions.

We will assume $\pi(y)$ to be known. To calculate the efficiency of the sampling scheme the parameter values, θ, must also be specified. Sometimes knowledge from previous studies can be used, but if no such data is available a pilot study is highly recommended.

3 Choice of likelihood

To correct for ascertainment we condition the likelihood on the fact that the data is ascertained. This likelihood can be constructed in different ways, depending on wether we also impose conditioning on response variables (retrospective likelihood) or explanatory variables (prospective likelihood). It is well known that conditioning the likelihood on non-ancillary statistics affects the efficiency of estimates, see for example Liang (1983), and it may also influence how the ascertainment scheme affects the efficiency. A comparison of the efficiency of different ascertainment corrected likelihoods in family based case-control studies is provided by Kraft & Thomas (2000). Of the four likelihoods investigated the joint likelihood was confirmed to be the most effective and the conditional likelihood for stratum-matched case-control data was the least efficient. The relative efficiency of the prospective and the retrospective likelihoods varied depending on data structure and genetic model.

We will here investigate likelihoods for three types of data: a likelihood for full data (3.1), a likelihood for data from a two stage design (3.2), and a likelihood for data under ascertainment (3.3). In likelihoods (3.1)-(3.3) variables X and Y will be modeled jointly. For data under ascertainment we will also investigate a prospective likelihood (3.5), and a retrospective likelihood (3.4). The retrospective likelihood has the attractive feature that the ascertainment cancels out of the formula. However, due to the loss of information in conditioning on the non-ancillary statistic Y, this likelihood turns out to be ill conditioned, so that the parameters describing Y as dependent on X usually are not identifiable unless some of the parameter values are assumed known. An exception to this is the effect parameter in logistic regression model, for which the prospective and retrospective likelihood give the same profile likelihood, see for example Kagan (2001) or Chen (2003). Calculations for the retrospective likelihood were performed for some, but not all, of the examples in Section 7. Calculations for the two-stage design will be presented in one of the examples.

The likelihoods are written as

$$\begin{aligned}
L_{\text{full}}(\theta) &= \prod_{i=1}^{n} P(z^i|\theta) && (3.1)\\
L_{\text{two}}(\theta,\pi) &= \prod_{i;J^i=1} P(y^i|\theta)(1-\pi(y^i)) \prod_{i;J^i=2} P(z^i|\theta)\pi(y^i) &&\\
&\propto \prod_{i;J^i=1} P(y^i|\theta) \prod_{i;J^i=2} P(z^i|\theta) = \prod_{i=1}^{n} P(y^i|\theta) \prod_{i;J^i=2} P(x^i|y^i,\theta) && (3.2)\\
L_{\text{asc}}(\theta,\pi) &= \prod_{i;J^i=2} P(z^i|A^i,\theta) && (3.3)\\
L_{\text{retr}}(\theta) &= \prod_{i;J^i=2} P(x^i|y^i,A^i,\theta) = \prod_{i;J^i=2} P(x^i|y^i,\theta) && (3.4)\\
L_{\text{pr}}(\theta,\pi) &= \prod_{i;J^i=2} P(y^i|x^i,A^i,\theta) && (3.5)
\end{aligned}$$

where n is the number of individuals and $z^i = (x^i, y^i)$, A^i and J^i represent full data, ascertainment, and number of stages of data collection for individual i.

4 Information matrices

Assuming $\theta = (\theta_1, \ldots, \theta_p)$, we define the score function as the $1 \times p$ vector

$$\psi(z;\theta) = \frac{\partial \log P(z|\theta)}{\partial \theta}$$

for fully observed data, z. Let us further introduce the five information matrices

$$\begin{aligned}
I_Z(\theta) &= Cov(\psi(Z;\theta)),\\
I_Y(\theta) &= Cov(E(\psi(Z;\theta)|Y)),\\
I_{X|Y,A}(\theta,\pi) &= E[Cov(\psi(Z;\theta)|Y)|A],\\
I_{Y|X,A}(\theta,\pi) &= E[Cov(\psi(Z;\theta)|X,A)|A],\\
I_{Z|A}(\theta,\pi) &= Cov(\psi(Z;\theta)|A).
\end{aligned}$$

With overall ascertainment probability $P(A|\theta,\pi) = P(J=2) = \int \pi(y)P(y;\theta)dy$, the Fisher information matrices resulting from likelihoods (3.1)-(3.5) can be expressed as

$$\begin{aligned}
I_{\text{full}}(\theta) &= nI_Z(\theta),\\
I_{\text{two}}(\theta,\pi) &= nI_Y(\theta) + I_{\text{retr}}(\theta,\pi),\\
I_{\text{asc}}(\theta,\pi) &= nP(A|\theta,\pi)I_{Z|A}(\theta,\pi),\\
I_{\text{retr}}(\theta,\pi) &= nP(A|\theta,\pi)I_{X|Y,A}(\theta),\\
I_{\text{pr}}(\theta,\pi) &= nP(A|\theta,\pi)I_{Y|X,A}(\theta,\pi).
\end{aligned}$$

Due to discrepancy in the amount of data contributing to the information, we may infer the inequalities

$$\begin{aligned}
I_{\text{full}}(\theta) &\geq I_{\text{two}}(\theta,\pi),\\
I_{\text{two}}(\theta,\pi) &\geq I_{\text{asc}}(\theta,\pi),\\
I_{\text{asc}}(\theta,\pi) &\geq I_{\text{pr}}(\theta,\pi),\\
I_{\text{asc}}(\theta,\pi) &\geq I_{\text{retr}}(\theta,\pi)
\end{aligned}$$

where $I_1 \geq I_2$ means that $I_1 - I_2$ is non-negative definite.

5 Monte Carlo Estimation

In some situations, such as when Y is normally distributed, analytical solutions to the expectations described above are not available. Monte Carlo simulations can then be used to overcome the computational difficulties. In the Monte Carlo calculations the ascertainment scheme only enters the calculations in the simulation of the data, so even complex ascertainment schemes are easily accommodated. The Monte Carlo samples are obtained by simulating data according to the model and applying the ascertainment scheme. Assume that $z_k^* = (x_k^*, y_k^*)$, $k = 1, \ldots, K$, is a random sample from $P(\cdot;\theta)$. Then

$$\begin{aligned}
\hat{I}_Z(\theta) &= K^{-1}\sum_{k=1}^{K} \psi(z_k^*;\theta)^T\psi(z_k^*;\theta),\\
\hat{I}_Y(\theta) &= K^{-1}\sum_{k=1}^{K} \hat{E}(\psi(Z;\theta)|y_k^*)^T\hat{E}(\psi(Z;\theta)|y_k^*),
\end{aligned}$$

where $\hat{E}(\psi(Z;\theta)|y)$ is the sample mean of all $\psi(z_k^*;\theta); y_k^* = y$ and T denotes vector transposition. We then get

$$\begin{aligned}
\hat{I}_{full}(\theta) &= n\hat{I}_Z(\theta), \\
\hat{I}_{asc}(\theta,\pi) &= nK^{-1}\sum_{k=1}^{K}\pi(y_k^*)(\psi(z_k^*;\theta)-\hat{\mu})^T(\psi(z_k^*;\theta)-\hat{\mu}), \\
\hat{I}_{pr}(\theta,\pi) &= nK^{-1}\sum_{k=1}^{K}\pi(y_k^*)\widehat{Cov}(\psi(Z;\theta)|x_k^*,A), \\
\hat{I}_{retr}(\theta,\pi) &= nK^{-1}\sum_{k=1}^{K}\pi(y_k^*)\widehat{Cov}(\psi(Z;\theta)|y_k^*), \\
\hat{I}_{two}(\theta,\pi) &= n\hat{I}_Y(\theta)+\hat{I}_{retr}(\theta,\pi),
\end{aligned}$$

where

$$\begin{aligned}
\hat{\mu} &= \sum_{k=1}^{K}\pi(y_k^*)\psi(z_k^*;\theta)/\sum_{k=1}^{K}\pi(y_k^*), \\
\widehat{Cov}(\psi(Z;\theta)|x,A) &= \sum_{k;x_k^*=x}\pi(y_k^*)(\psi(z_k^*;\theta)-\hat{\mu}(x))^T(\psi(z_k^*;\theta)-\hat{\mu}(x))/\sum_{k;x_k^*=x}\pi(y_k^*), \\
\hat{\mu}(x) &= \sum_{k;x_k^*=x}\pi(y_k^*)\psi(z_k^*;\theta)/\sum_{k;x_k^*=x}\pi(y_k^*),
\end{aligned}$$

and similarly, $\widehat{Cov}(\psi(Z;\theta)|y)$ is the sample covariance of all $\psi(z_k^*;\theta); y_k^* = y$.

6 Cost Efficiency Tradeoff

Efficiency, e, will be evaluated in terms of the Fisher information, I:

$$e(\theta,\pi) = I_{rr}^{-1}(\theta,\pi)/I_{full,rr}^{-1}(\theta) \tag{6.1}$$

where $I(\theta,\pi)$ is the information of the selected sample, and $I_{full}(\theta)$ is the information of a full size simple random sample (SRS). Hence the efficiency is based on the asymptotic variance I_{rr}^{-1} of the r^{th} component of θ, $r = 1,...,p$. Other scalar functions of the Fisher information than I_{rr}^{-1} are discussed in Grünewald & Hössjer (2008). Note that $I(\theta,\pi)$ in (6.1) can be any of the information matrices $I_{two}(\theta,\pi)$, $I_{asc}(\theta,\pi)$, $I_{retr}(\theta,\pi)$ or $I_{pr}(\theta,\pi)$, resulting in efficiencies $e_{two}(\theta,\pi)$, $e_{asc}(\theta,\pi)$, $e_{retr}(\theta,\pi)$ and $e_{pr}(\theta,\pi)$. In this paper the main focus will be on $e_{\text{asc}}(\theta,\pi)$.

If represented graphically with ascertainment probability, $P(A|\theta,\pi)$, on the x-axis and efficiency, $e(\theta,\pi)$, on the y-axis, the efficiency of a SRS, i.e. a sample with $\pi(y)\equiv a$, $0<a\leq 1$, will give a straight line with equation $P(A)=e$. A beneficial ascertainment scheme will give an efficiency above that line. This graphical representation is useful for example when there is a limited number of cases to select, and we want to ensure that we lose no more than a certain percentage of the total efficiency by sampling the controls more sparsely.

If there are no restrictions with regard to available subjects to sample, interest may instead be in the cost efficiency tradeoff. One way to formulate this is via the cost adjusted efficiency

$$CE(\theta,\pi)=e(\theta,\pi)/\text{RAC}(\theta,\pi),$$

where $\text{RAC}(\theta,\pi)$ is the relative average cost of the sample compared to a full sample. With $C_1\geq 0$ the cost of sampling individuals at Stage 1, and $C_2>C_1$ the total cost of sampling individuals at Stages 1 and 2, we write

$$RAC(\theta,\pi)=C_2^{-1}\int\left(C_1(1-\pi(y))+C_2\pi(y)\right)P(y;\theta)dy=\frac{C_1}{C_2}+\frac{C_2-C_1}{C_2}P(A|\theta,\pi).$$

The cost adjusted efficiency thus quantifies how cost efficient the present design π is compared to a SRS design. A beneficial sampling scheme will give $CE(\theta,\pi)>1$, and the most cost efficient sampling scheme is identified as the highest point on the curve. Depending on what efficiency is used we use notation $CE_{\text{two}}(\theta,\pi)$, $CE_{\text{asc}}(\theta,\pi)$ etc. For examples of how $CE_{\text{two}}(\theta,\pi)$ can be used to evaluate efficiency in a two-stage design see Thomas et al. (2004), and in regression analysis with incomplete covariate data see Reilly & Pepe (1995).

For ascertainment samples we put $C_1=0$, implying that the cost of sampling first stage data not used in the analysis is ignored, and hence $RAC(\theta,\pi)=P(A|\theta,\pi)$. For two-stage data it may be relevant to put $C_1>0$, and adjust C_1/C_2 to reflect the relative cost of first and second stage sampling, in order to accommodate different data collection scenarios.

7 Examples

In this section three models will be investigated with respect to how the ascertainment scheme affects the efficiency of the maximum likelihood estimates. The ascertainment schemes compared in the examples are chosen as illustrations, but other ascertainment schemes may be more relevant, or more efficient, in a specific study.

The results are presented in graphs. The values are standardized by the efficiency of a full sample.

All calculations are made in the software R (R Development Core Team 2005).

Model i: Logistic regression

Logistic regression is frequently used in case-controls studies in epidemiology. A property of the logit link function is that sampling probabilities cancel out of the effect estimates (Anderson 1972), which facilitates analysis. Often an equal number of cases and controls are sampled, and interest is in estimating the effect parameter. See Maydrech & Kupper (1978) for calculations of cost and sample size is case-control studies.

For simplicity we use a binomial distribution for X as well as Y:

$$\begin{aligned}\theta &= (\alpha_X, \alpha_Y, \beta_{XY}),\\ X &\sim \mathrm{Bin}(1, \frac{\exp(\alpha_X)}{1+\exp(\alpha_X)}),\\ Y|\{X=x\} &\sim \mathrm{Bin}(1, \frac{\exp(\alpha_Y+\beta_{XY}x)}{1+\exp(\alpha_Y+\beta_{XY}x)}).\end{aligned}$$

We here sample all observations with outcome $y = 1$, and a proportion a of observations where $y = 0$. That is

$$\pi(y) = \begin{cases} a; & y = 0,\\ 1; & y = 1.\end{cases}$$

Since Y is discrete, Monte Carlo simulations were not necessary in this example.

For the joint ascertainment likelihood the efficiency was calculated for all three parameters $\theta = (\alpha_X, \alpha_Y, \beta_{XY})$, whilst for the prospective likelihood the efficiency was only calculated for $\theta_Y = (\alpha_Y, \beta_{XY})$, since the likelihood was conditioned on α_X. When all parameters were included in the calculations for the retrospective likelihood the model turned out to be ill conditioned, the information matrix was positive semidefinite and had rank 2 instead of 3. Further investigation indicated that α_X and α_Y could not be estimated simultaneously. To overcome this problem we assumed α_X to be known in the calculations presented for the retrospective likelihood in model i.

Figure 2 illustrates the efficiency, $e(\theta, \pi)$, for the estimates of a set of parameters θ: $(\alpha_X = -1, \alpha_Y = -2, \beta_{XY} = 2)$. With this set of parameters

$P(Y=1) \approx 0.22$ in the SRS. In the other ascertainment schemes individuals with $y=1$ are over-sampled. The efficiency of a SRS of the same size is included for comparison and a vertical line is drawn at the ascertainment probability that gives equally many cases, $(y=1)$, and controls, $(y=0)$. The efficiencies from the three likelihoods are standardized by the full sample efficiency of the most efficient likelihood. The efficiencies using the ascertained likelihood and the prospective likelihood were so similar that they could not be distinguished. The retrospective likelihood gave a lower efficiency in the estimation of α_Y but can not be distinguished from the ascertained and prospective likelihoods in the estimation of β_{XY}. The prospective and retrospective likelihoods generate the same profile likelihood for the estimation of β_{XY} (Chen 2003), so it is not surprising that those estimates have the same efficiency. Focusing on the joint ascertainment likelihood, the gain in efficiency, compared to a SRS, is mainly present in the estimation of α_Y and β_{XY}, while α_X is relatively unaffected.

The benefit of a specific ascertainment scheme will depend on the parameter values in the model. Figure 3 exemplifies how changing one parameter affects $e_{asc}(\theta,\pi)$ in Model *i*. The parameter values are as above but with α_Y taking the values $0, -2$ and -4. In the full SRS this gives $P(Y=1) \approx 0.60, 0.22$ and 0.045 respectively. Figure 4 illustrates the same example as Figure 3, but with cost adjusted efficiencies. Our calculations confirm that an equal number of cases and controls is an efficient choice when interest is in estimating β_{XY}.

Model *ii*: Linear regression

Even though response variables in epidemiological studies often are dichotomized to fit into the logistic regression model, the true nature of many variables, such as body mass index (BMI), blood pressure and plasma glucose, are continuous. Categorizing continuous variables often lead to a loss in efficiency, see for example Vargha et al. (1996). While preserving the continuous nature of the response variable it is still possible to use non-random ascertainment to increase efficiency. In this example a normal distribution is assumed for the response variable, and a linear regression model is used,

$$\begin{aligned}\theta &= (\alpha_X, \alpha_Y, \beta_{XY}, \sigma_Y),\\ X &\sim \mathrm{Bin}(2, \frac{\exp(\alpha_X)}{1+\exp(\alpha_X)}),\\ Y|\{X=x\} &\sim \mathrm{N}(\alpha_Y + \beta_{XY} \times x, \sigma_Y^2).\end{aligned}$$

Ascertainment probabilities depend on the value of Y. We here choose to specify a cut-off value t, to let

$$\pi(y) = \begin{cases} a; & y < t, \\ 1; & y \geq t, \end{cases}$$

and vary the value of a. Monte Carlo simulations were used for the computation.

In this example the efficiency of a two stage likelihood was calculated as a comparison to the efficiency in the ascertainment sample. For the ascertained likelihood and the two-stage likelihood the efficiency was calculated for all parameters θ, while for the prospective likelihood the efficiency was calculated for all parameters but α_X. The efficiency in calculating α_Y, β_{XY} and σ_Y was the same for the prospective likelihood as for the joint ascertainment likelihood. The retrospective likelihood suffered from problems similar to those in model *i*, the information matrix here had rank 3 instead of 4, and was positive semidefinite. As in model *i* the parameters α_X and α_Y could not be estimated simultaneously. We will not present any results for the prospective and retrospective likelihood for model *ii*.

In Figure 5 the cost adjusted efficiency is presented for parameter values $(\alpha_X = -4, \alpha_Y = 0, \beta_{XY} = 2, \sigma_Y = 1)$ and cut-of $t = 2$. For this set of parameters $P(Y < t) = 0.96$ in the SRS. A vertical line in the graph indicates the ascertainment scheme where $P(Y < t|A) = 0.5$. In the figure three different costs are investigated for the two stage sample: $(C_1, C_2) = (0, 1)$ illustrates a situation where Y can be observed free of cost, $(C_1, C_2) = (1/3, 1)$ means that, per individual, sampling Y is associated with half of the cost of sampling X, and $(C_1, C_2) = (1/2, 1)$ implies that Y is as expensive to observe as X. It would also be possible to choose $C_1 > 1/2$ if Y is more expensive to sample than X, even though this is not the typical situation where outcome dependent sampling is considered. The two stage sample where $(C_1, C_2) = (0, 1)$ is more cost efficient than the ascertainment sample, meaning that the free of cost first stage data does contribute with information. The discrepancy is most pronounced in the estimation of α_Y and σ_Y. When $(C_1, C_2) = (1/3, 1)$ the two stage sample is more cost efficient than the ascertainment sample for the estimation of α_Y and σ_Y, but not for the other parameters. When $(C_1, C_2) = (1/2, 1)$ the two stage sample is more efficient estimating α_Y, and also more efficient estimating σ_Y for some ascertainment schemes.

Model *iii*: Selection on more than one variable

Ascertainment probabilities can be based on the outcome of more than one variable. In this example we have included two normally distributed variables Y_1 and Y_2, that both affect the ascertainment probability, and let the explanatory variable be binomially distributed. The ascertainment scheme is

$$\pi(y_1, y_2) = \begin{cases} a; \ y_1 < t_1 \cap y_2 < t_2, \\ 1; \ \text{otherwise,} \end{cases}$$

where $0 < a \leq 1$ is varied. A dependence between Y_1 and Y_2 is also included in the model.

$$\begin{aligned} \theta &= (\alpha_X, \alpha_{Y_1}, \beta_{XY_1}, \sigma_{Y_1}, \alpha_{Y_2}, \beta_{XY_2}, \beta_{Y_1Y_2}, \sigma_{Y_2}), \\ X &\sim \text{Bin}(1, \frac{\exp(\alpha_X)}{1+\exp(\alpha_X)}), \\ Y_1|\{X = x\} &\sim \text{N}(\alpha_{Y_1} + \beta_{XY_1} \times x, \sigma^2_{Y_1}), \\ Y_2|\{X = x, Y_1 = y_1\} &\sim \text{N}(\alpha_{Y_2} + \beta_{XY_2} \times x + \beta_{Y_1Y_2} \times y_1, \sigma^2_{Y_2}). \end{aligned}$$

The model can be illustrated by the graph

$$\begin{tikzcd} & Y_1 \arrow[dr] \arrow[dd] & \\ X \arrow[ur] \arrow[dr] & & A. \\ & Y_2 \arrow[ur] & \end{tikzcd}$$

For model *iii* only the results for the joint ascertained likelihood will be presented. The efficiencies for the prospective likelihood gave results that could not be distinguished from the joint ascertainment likelihood. The efficiencies for the retrospective likelihood were not calculated, the information matrix was positive semidefinite and had rank 3 instead of 8, and no more than three parameters could be estimated simultaneously. We could not see any obvious pattern of which variables could be estimated together for the retrospective likelihood.

The parameter values used were $(\alpha_X = -3, \alpha_{Y_1} = 0, \beta_{XY_1} = 2, \sigma_{Y_1} = 1, \alpha_{Y_2} = 0, \beta_{XY_2} = 1, \beta_{Y_1Y_2} = 1, \sigma_{Y_2} = 1)$. The model was run for three different sets of cut-offs, $(t_1, t_2) = (1, 1), (2, 2)$ and $(3, 3)$. For these values $P(Y_1 < t_1 \cap Y_2 < t_1)$ was 0.68, 0.88 and 0.96 respectively in the SRS. The results are presented in Figure 6. Vertical lines mark the ascertainment schemes where $P(Y_1 < t_1 \cap Y_2 < t_1|A) = 0.5$. Most of the parameter estimates benefit from ascertainment schemes with $a < 1$, while $\hat{\alpha}_{Y_1}$ and $\hat{\alpha}_{Y_2}$ do not. In this example the largest benefits from small values of a are for high values of (t_1, t_2), that is, when a large proportion of data in the SRS is in the range $Y_1 < t_1 \cap Y_2 < t_1$. The setting in this example was chosen to illustrate that the effect of ascertainment on efficiency can be large even in complex models. For other values the benefit may be less obvious.

Monte Carlo simulations were used for computation in model *iii*.

8 Discussion

In this paper we have presented tools for comparing ascertainment schemes with respect to efficiencies and cost adjusted efficiencies. The efficiencies are expressed in terms of Fisher information matrices. Three examples were examined, and the results were presented in graphs. Monte Carlo simulations were used for computation in two of the examples.

From the examples investigated it is apparent that non-random ascertainment schemes sometimes increase efficiency compared to a SRS, but also that they sometimes perform worse. Note that the slope of the efficiency curve is often steep, so that the value of $P(A|\theta, \pi)$ that results in the highest efficiency is close to values with low efficiency, see for example the efficiency of $\hat{\beta}_{XY_1}$ in Figure 6. It is therefore important to investigate the efficiency in the specific study setting before collecting data. Similarly to power analysis and local experimental designs, the calculations require specifying parameter values, which in reality are unknown. Pilot studies are therefore a valuable tool to acquire more knowledge about the data. It is also advisable to calculate the efficiency for some different sets of parameter values.

In this paper $e = h(I) = I_{rr}^{-1}$ was used and we have investigated efficiency of parameter estimates individually. Alternatively $h(I)$ could include all components of θ, such as $h(I) = tr(I^{-1})$, so that the choice of selection scheme can be based on a single criterion.

In the efficiency calculations in this paper it is assumed that the cost of sampling is the same for all subjects within the same sampling design. In reality the cost of sampling can differ depending on the outcome of one or more of the sampled variables. For example the cost of sampling cases can differ from the cost of sampling controls in a case-control study. To obtain a sampling scheme that is cost-efficient differential costs can be incorporated in the calculations. An example of this can be found in Maydrech & Kupper (1978) where cost functions are presented for cohort and case-control studies.

Here maximum likelihood estimation has been used, based on likelihoods conditioned on ascertainment. In reality, due to computational issues, these parameter estimates are not always straightforward to obtain. Other estimation procedures might then be preferable. Reilly & Pepe (1995) use a mean score method for regression analysis with incomplete or auxiliary covariate data. Other methods to correct for ascertainment are also available. Neuhaus (2000) describes how adjusting link functions can correct for ascertainment in binary regression models. Simulation based methods to estimate parameters in data with ascertainment have been described for example by Clayton (2003) and by Grünewald & Humphreys (2008). When the estimation procedure differs from what was used in this paper, comparison of the efficiency of the ascertainment schemes can still be carried out analogously, using asymptotic variances appropriate to the estimation procedure rather than the inverse Fisher information matrix.

While case-control designs are frequently used in for example epidemiology, selection on continuous outcomes, or on multiple outcomes, is not as common. This may be due to the added complexity in the analysis of data, but also because it is not transparent which designs are efficient. Plotting efficiencies, or cost adjusted efficiencies, as suggested in this paper, may aid in the choice of design.

Acknowledgments

Maria Grünewald's work was supported by the Swedish Foundation for Strategic Research (SSF), grant number A3 02:129, and Ola Hössjer's work by the Swedish Research Council, grant number 621-2005-2810. The authors thank Keith Humphreys at Karolinska Institutet, Stockholm, for valuable comments.

References

Anderson, J. A. (1972), 'Separate sample logistic discrimination', *Biometrika* **59**(1), 19–35.

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.

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

Fisher, R. (1934), 'The effects of methods of ascertainment upon the estimation of frequencies', *Annals of Eugenics* **6**, 13–25.

Grünewald, M. & Hössjer, O. (2008), A general statistical framework for multistage designs. Working paper.

Grünewald, M. & Humphreys, K. (2008), A Stochastic EM type algorithm for estimation in data with ascertainment on continuous outcomes, Technical Report 2008:5, Department of Mathematics, Stockhom University, 10691 Stockholm, Sweden.

Kagan, A. (2001), 'A note on the logistic link function', *Biometrika* **88**(2), 599–601.

Kraft, P. & Thomas, D. C. (2000), 'Bias and efficiency in family-based gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods', *Am. J. Hum. Genet* **66**, 1119–1131.

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. (2002), *Statistical analysis with missing data*, Wiley series in probability and statistics, John Wiley & Sons, Inc., New York; Chichester.

Maydrech, E. & Kupper, L. (1978), 'Cost considerations and sample size requirements in cohort and case-control studies', *American journal of epidemiology* **107**, 201–205.

Melas, V. (2006), *Functional Approach to Optimal Experimental Design*, number 184 *in* 'Lecture Notes in Statistics', Springer, United States of America.

Neuhaus, J. M. (2000), 'Closure of the class of binary generalized linear models in some non-standard settings', *J.R. Statist. Soc. B* **62**(Part 4), 839–846.

Neyman, J. (1938), 'Contribution to the theory of sampling human populations', *Journal of the American Statistical Association* **33**(201), 101–116.

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

Patil, G. & Taillie, C. (1989), Probing encountered data, meta analysis and weighted distribution methods, Technical Report 89-0101, Department of Statistics, Center for Statistical Ecology and Environmental Statistics, The Pennsylvania State University, University Park, PA 16802.

Patil, G., Taillie, C. & Talwalker, S. (1973), *Statistics for the environment*, John Wiley & Sons Ltd, chapter Encounter Sampling and Modelling in Ecological and Environmental Studies Using Weighted Distribution Methods, pp. 45–71.

R Development Core Team (2005), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Reilly, M. (1996), 'Optimal sampling strategies for two-stage studies', *American journal of epidemiology* **143**(1), 92–100.

Reilly, M. & Pepe, M. S. (1995), 'A mean score method for missing and auxiliary covariate data in regression models', *Biometrika* **82**(2), 299–314.

Thomas, D., Xie, R. & Gebregziabher, M. (2004), 'Two-stage sampling designs for gene association studies', *Genetic Epiemiology* **27**, 401–414.

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.

Whittemore, A. S. & Halpern, J. (1997), 'Multi-stage sampling in genetic epidemiology', *Statistics in medicine* **16**, 153–167.

Zhou, H., Chen, J., Rissanen, T. H., Korrick, S. A., Hu, H., Salonen, J. T. & Longnecker, M. P. (2007), 'Outcome-dependent sampling: An efficient sampling and inference procedure for studies with a continuous outcome', *Epidemiology* **18**(4), 461–468.

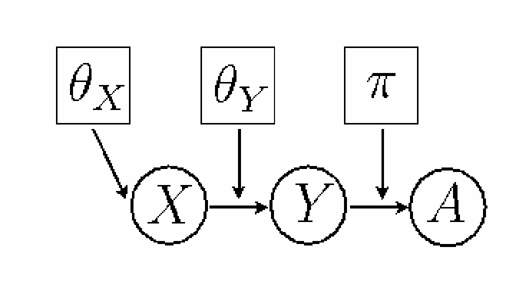

Figure 1: A retrospective sampling design. Sampling probabilities π depend on the outcome of explanatory variables Y but not on explanatory variables X, nor on model parameters $\theta = (\theta_X, \theta_Y)$.

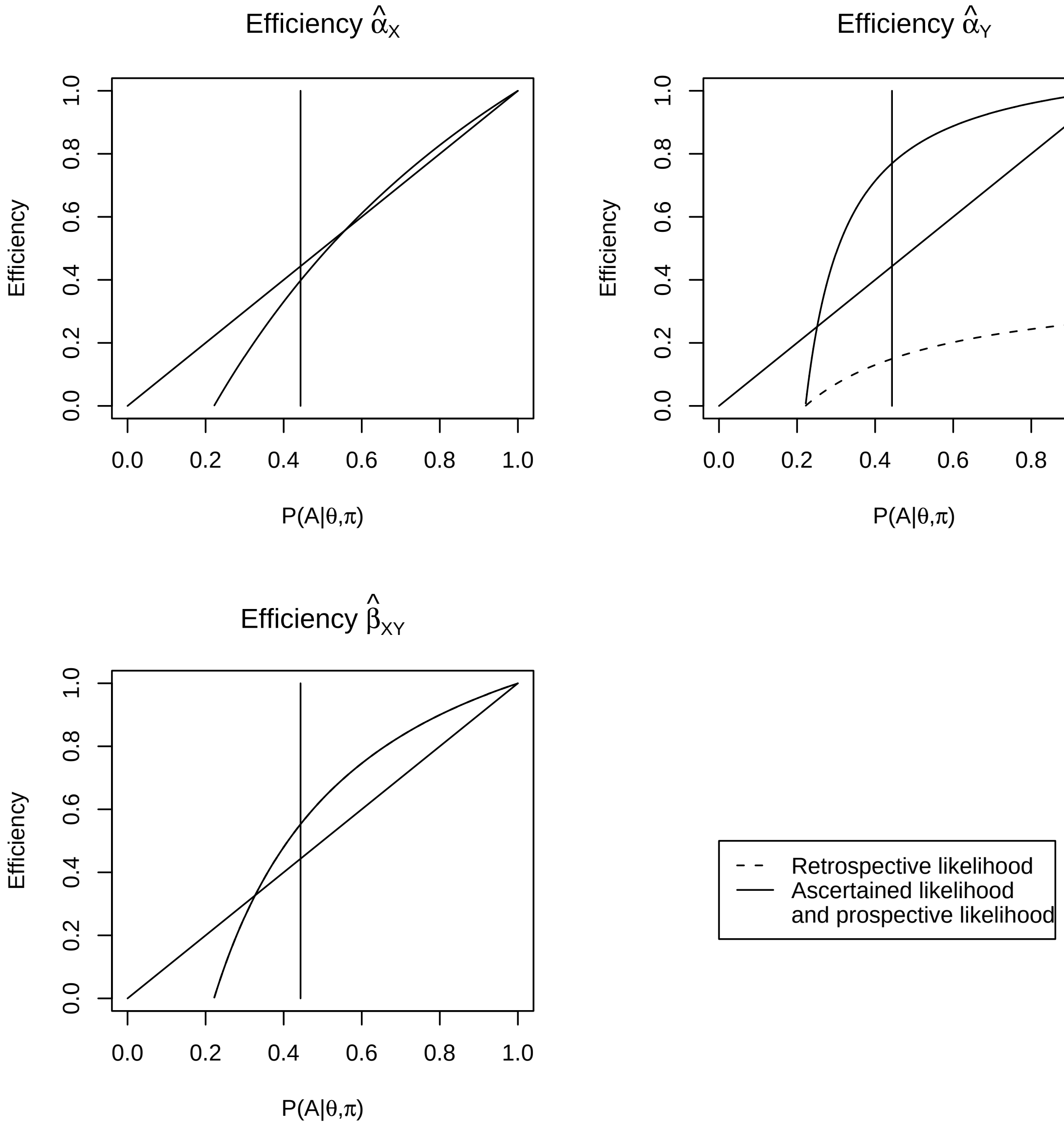

Figure 2: Model *i*. Efficiency, $e_{\text{asc}}(\theta, \pi)$, for estimation of parameters for different ascertainment schemes in logistic regression. $\alpha_X = -1, \alpha_Y = -2, \beta_{XY} = 2, 0 < a \leq 1$. α_X is not estimated by the retrospective likelihood.

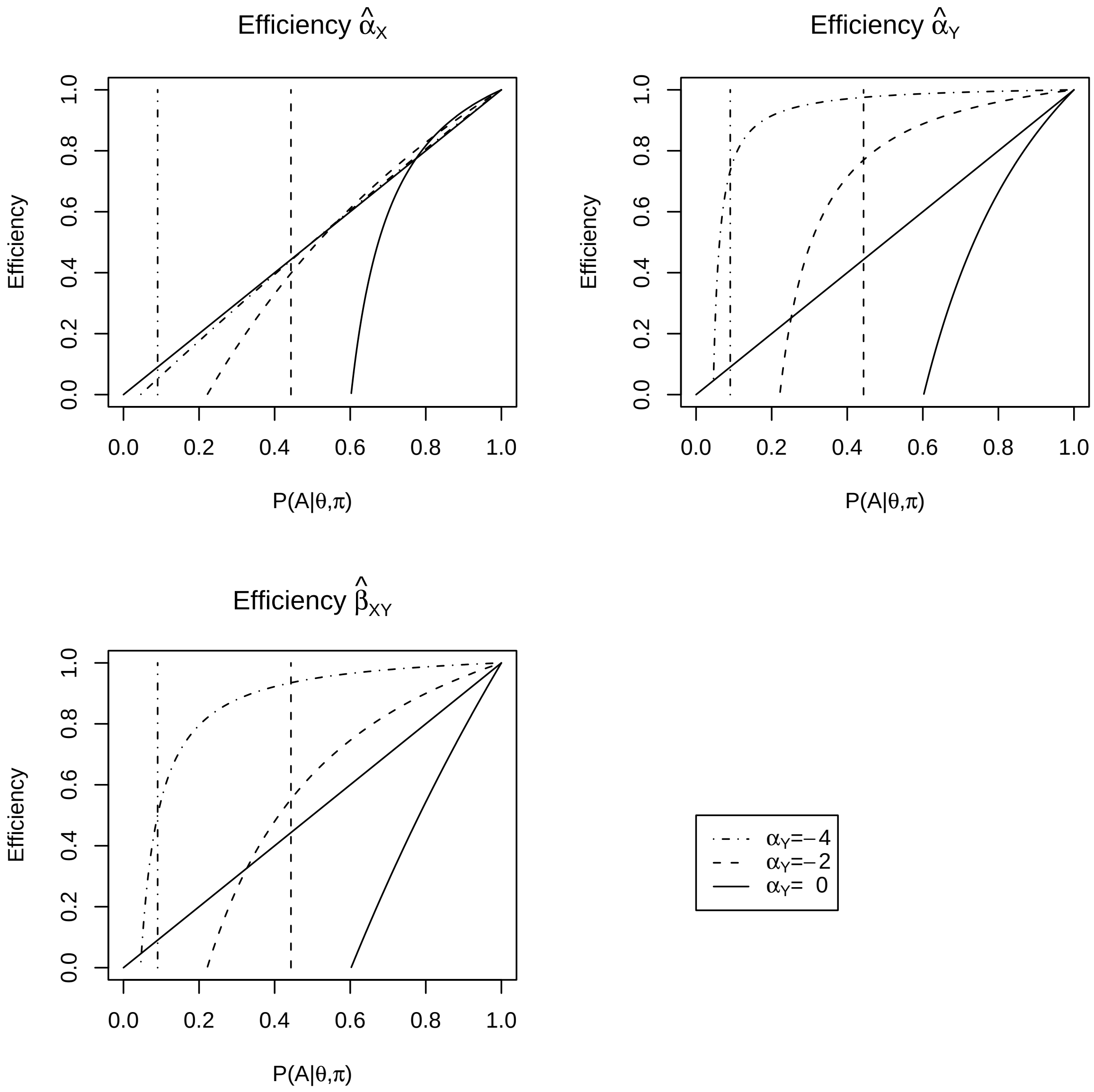

Figure 3: Model *i*. Efficiency, $e_{\text{asc}}(\theta, \pi)$, for estimation of parameters for different ascertainment schemes in logistic regression. $\alpha_X = -1, \alpha_Y = (0, -2, -4), \beta_{XY} = 2, 0 < a \leq 1$. Vertical lines at $P(Y = 1|A) = 0.5$, except for set of parameters with $\alpha_Y = 0$ where $P(Y = 1|A) > 0.5$ for the range of values plotted.

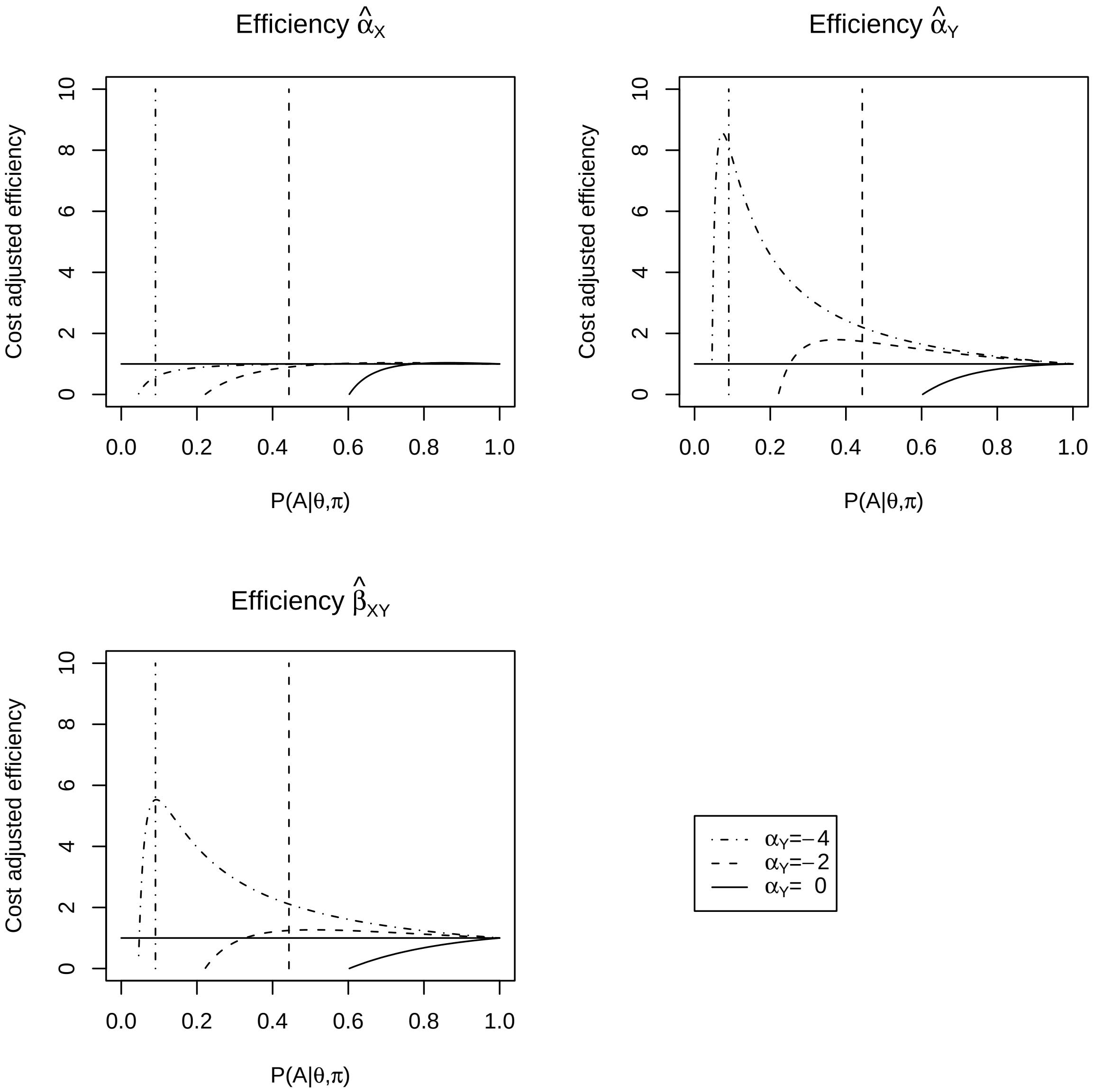

Figure 4: Model i. Cost adjusted efficiency, $CE_{\text{asc}}(\theta, \pi)$, for estimation of parameters for different ascertainment schemes in logistic regression. $\alpha_X = -1, \alpha_Y = (0, -2, -4), \beta_{XY} = 2, 0 < a \leq 1$. Vertical lines are drawn for designs such that $P(Y = 1|A) = 0.5$, except for set of parameters with $\alpha_Y = 0$ where $P(Y = 1|A) > 0.5$ for the range of values plotted.

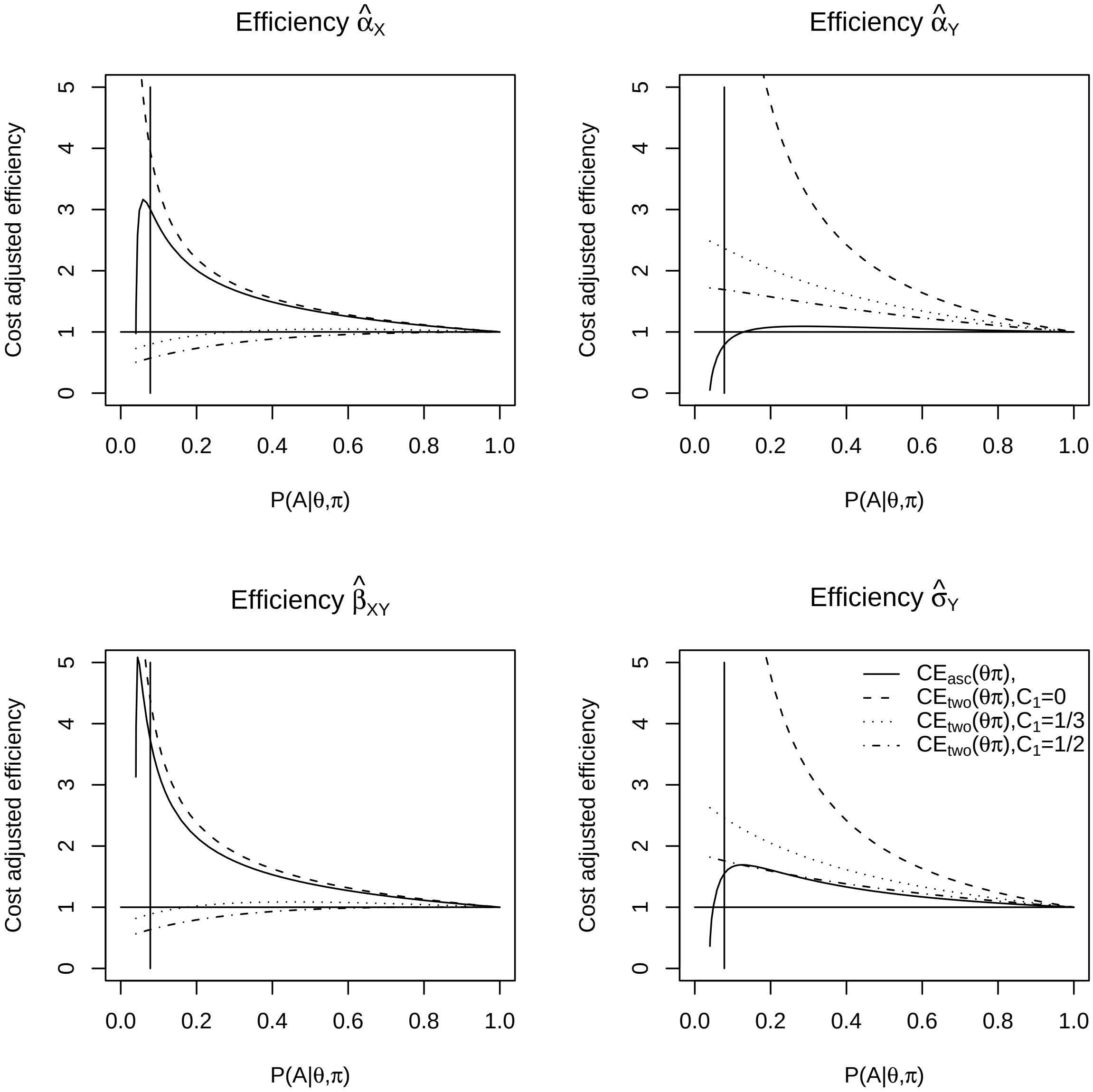

Figure 5: Model *ii*. Cost adjusted efficiency, for estimation of parameters for different ascertainment schemes in linear regression, using an ascertainment likelihood and a two stage likelihood. For the two stage likelihood different costs $C_1 = (0, 1/3, 1/2)$, $C_2 = 1$ are applied. A vertical line indicates the ascertainment scheme where $P(Y < t|A) = 0.5$. $\alpha_X = -4, \alpha_Y = 0, \beta_{XY} = 2, \sigma_Y = 1, t = 2, 0 < a \leq 1, K = 10000,$

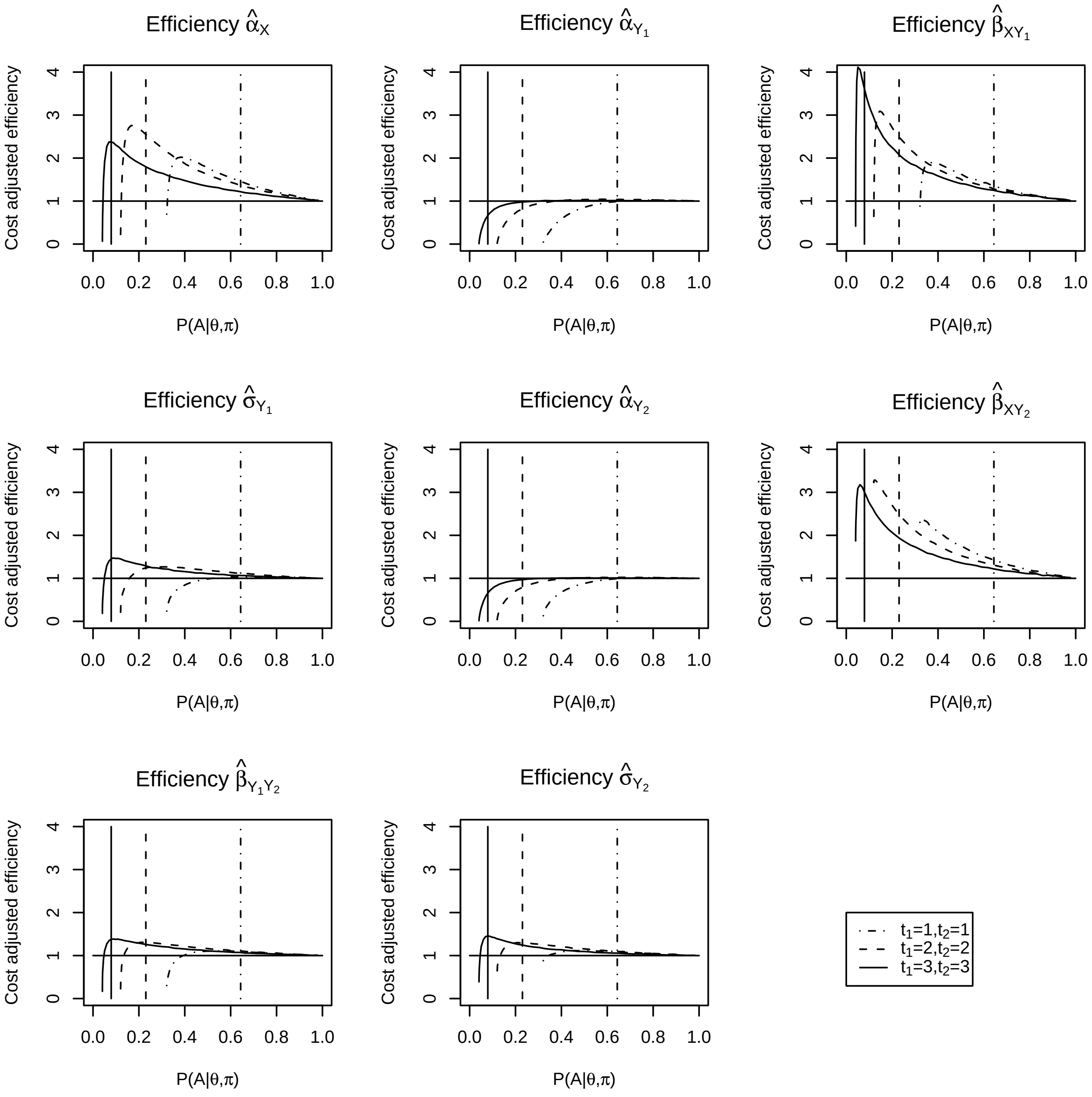

Figure 6: Model *iii*. Cost-adjusted efficiency, $CE_{\text{asc}}(\theta, \pi)$, where $\pi(y_1, y_2) = a$ is varied for $y \in (y_1 < t_1 \cap y_2 < t_2)$. Three different sets of (t_1, t_2) are used. Vertical lines indicate the ascertainment schemes where $P(Y_1 < t_1 \cap Y_2 < t_1|A) = 0.5$. $\alpha_X = -3, \alpha_{Y_1} = 0, \beta_{XY_1} = 2, \sigma_{Y_1} = 1, \alpha_{Y_2} = 0, \beta_{XY_2} = 1, \beta_{Y_1Y_2} = 1, \sigma_{Y_2} = 1,\ 0 < a \leq 1, K = 500000$.