Journal of Statistical Software

April 2014, Volume 57, Issue 9. http://www.jstatsoft.org/

A SAS Package for Logistic Two-Phase Studies

Walter Schill
Leibniz Institute for
Prevention Research
and Epidemiology – BIPS

Dirk Enders
Leibniz Institute for
Prevention Research
and Epidemiology – BIPS

Karsten Drescher
Statistical Office
of Bremen

Abstract

Two-phase designs, in which for a large study a dichotomous outcome and partial or proxy information on risk factors is available, whereas precise or complete measurements on covariates have been obtained only in a stratified sub-sample, extend the standard case-control design and have been proven useful in practice. The application of two-phase designs, however, seems to be hampered by the lack of appropriate, easy-to-use software. This paper introduces **sas-twophase-package**, a collection of SAS-macros, to fulfill this task. **sas-twophase-package** implements weighted likelihood, pseudo likelihood and semi-parametric maximum likelihood estimation via the EM algorithm and via profile likelihood in two-phase settings with dichotomous outcome and a given stratification.

Keywords: case-control study, two-phase, weighted likelihood, pseudo likelihood, maximum likelihood, profile likelihood, EM algorithm, SAS.

1. Introduction

Two-phase study is the name of a study design which was introduced into epidemiology by Breslow and Cain (1988), see Breslow (2005) for a historical account of the ideas. In a typical (logistic) two-phase study, data are collected in a double-sampling scheme: In a first phase, data are collected for a large number of participants on a (dichotomous) outcome and some covariates, leading to a stratification. In the second phase additional and/or more precise data on covariates are ascertained from these strata in a sub-sample. Thus, this setup encompasses two situations: (a) a measurement error scenario, in which a proxy variable exists in phase one which in phase two is measured with higher precision and (b) a scenario in which a covariate is missing in phase one but is available in phase two. This missing-value scenario makes sense only if the missing covariate is associated with the phase one variables. The statistical analysis makes use of both data sources.

This study design is most useful if in a large first phase study a dichotomous outcome (D) and cheap or easily collectable information (Z) are known for each participant, while it is expensive to ascertain precise and/or complete information on covariates. If complete data X were available for everybody, one would employ a logistic model $\mathsf{P}(D = 1|X = x) = 1/[1 + \exp(-\alpha - x^\top\beta)]$ to estimate the log odds ratio parameter vector β. The first phase study can be prospective or retrospective (case-control).

Regression methods based on weighted likelihood, pseudo likelihood and maximum likelihood to evaluate two-phase designs have been published in the past 25 years. Despite these developments, application of the two-phase design is scarce: In a review of five top-ranking epidemiology journals, Haneuse, Saegusa, and Lumley (2011) identified *one* article employing the two-phase design in the period 2002 – 2007 (4792 studies were reported, among these were 816 case-control studies). We extended this review to the years 2008 – 2012 and identified one additional article reporting a two-phase study (among 478 case-control studies). We attribute these findings in parts to the lack of appropriate, easy-to-use software.

Programs we are aware of include the following: Lumley (2004) describes the R package **survey** which covers a broad spectrum of survey designs including cohort studies. The weighted likelihood method, based on Horvitz and Thompson (1952), is available to analyze two-phase designs with binary outcome. The variance formula, however, differs from ours in that a finite population correction is implemented. Chris Wild (Scott and Wild 2006) provides an R package **missreg** (`http://www.stat.auckland.ac.nz/~wild/software.html`). The package offers several functions for regression analysis in response selective and missing data schemes. The function `bin2stg()` implements nonparametric maximum likelihood estimation for binary regression models using a profile likelihood. The logit-, probit- and complementary loglog link can be employed. Recently, Haneuse *et al.* (2011) published the R package **osDesign**, including capabilities for design- and power calculations. Their `tps()` function implements parameter estimation by weighted likelihood (without providing model based standard errors), the pseudo likelihood method of Breslow and Cain and also implements the maximum likelihood approach of Breslow and Holubkov (1997); `tps()` does not accept weights for the phase two data. Like the **survey** package, **osDesign** is available from the Comprehensive R Archive Network (`http://CRAN.R-project.org/`).

This paper introduces **sas-twophase-package**, a collection of SAS-macros to perform logistic regression with data in a two-phase setup. The methods presented in Schill and Drescher (1997) and Scott and Wild (1997) are implemented. To run the programs, SAS/STAT (SAS Institute Inc. 2008b) – and SAS/IML (SAS Institute Inc. 2008a) software must be available on your computing environment. The software and a comprehensive documentation can be downloaded from `http://www.tinyurl.com/schill-twophase`. The package replaces an older, undocumented version that was distributed with our two-phase planning tool (Schill, Wild, and Pigeot 2007).

The paper is organized as follows: After giving some notation, Section 3 summarizes the methodology section of Schill and Drescher (1997) and the findings of Scott and Wild (1997). We describe the structure of the macro package in Section 4 and present two examples. A discussion concludes the paper.

2. Notation

We assume that in a population the probability of a binary outcome D in a person with covariate vector $X = x$ is given by the logistic model

$$\mathsf{P}(D = 1|X = x) = F\left(\alpha + x^\top \beta\right) = \frac{\exp(\alpha + x^\top \beta)}{1 + \exp(\alpha + x^\top \beta)}, \tag{1}$$

where $F(\cdot)$ denotes the standard logistic distribution function and x denotes a $p \times 1$ vector including exposures, covariates and interactions. Focus lies on inference about the log odds ratio parameter $\beta = (\beta_1, \ldots, \beta_p)^\top$.

The first phase sample comprises data on outcome $D \in \{0, 1\}$ and measurements Z of partial or proxy information about X, leading to a stratification S with $J > 1$ strata. Let N_{ij} denote the number of first phase observations with $(D, S) = (i, j)$, $i = 0, 1$ and $j = 1, \ldots, J$. The first phase sample can be prospective or retrospective (case-control).

At the second phase of sampling, $0 < n_{ij} \leq N_{ij}$ individuals are randomly selected from within each cell of the first phase data for covariate ascertainment. For notational convenience, we assume X to be discrete taking values x_{jk}, $k = 1, \ldots, K_j$, say, within stratum j. Let n_{ijk} denote the number of second phase observations falling into cell (i, j, k).

3. Methods

Estimation based on weighted likelihood, pseudo likelihood and maximum likelihood are presented. We first describe the likelihood functions for prospective first phase samples, the extension to retrospective samples is given in Section 3.6. The parameter covariance formulae of the two-phase methods are complex and are not reproduced here. For a description of the (model-based) parameter covariance matrices that are provided with the package, refer to the underlying articles of Schill and Drescher (1997) and Scott and Wild (1997).

Assumption. With the exception of the weighted likelihood method, all other estimation methods require that stratum variable S and outcome D are conditionally independent, given $X = x$: $\mathsf{P}(D = 1|S = j, x) = \mathsf{P}(D = 1|x)$, that is, outcome probabilities depend on S only through X. An equivalent formulation is: $\mathsf{P}(S = j|x, D = 0) = \mathsf{P}(S = j|x, D = 1)$. This assumption is automatically fulfilled in a missing-value context, because Z or parts of Z are included in X and S is a function of Z. In a measurement error setting this assumption is termed 'assumption of non-differential error' and may be violated.

3.1. Weighted likelihood

The idea of this approach (Flanders and Greenland 1991; Reilly and Pepe 1995) stems from the survey research literature (Horvitz and Thompson 1952) and is to maximize the complete data likelihood, where the unknown cell counts are replaced by the observed counts n_{ijk}, weighted by the inverse selection probabilities N_{ij}/n_{ij} within each (D, S)-cell. Thus $\theta_{\mathrm{WL}} = (\alpha, \beta^\top)^\top$ is estimated by $\widehat{\theta}_{\mathrm{WL}}$, which is obtained by maximization of

$$L_{\mathrm{WL}} = \prod_i \prod_j \prod_k \mathsf{P}(D = i \mid x_{jk})^{\left(\frac{N_{ij}}{n_{ij}} n_{ijk}\right)}.$$

Implementation. The weighted likelihood method uses SAS/STAT procedure `proc logistic` to estimate the parameter. Then SAS/IML matrix language is used to compute the parameter covariance matrix.

3.2. Pseudo likelihood

The pseudo likelihood approach utilizes a marginal outcome model for the phase one data and derives stratum-specific outcome probabilities for the second phase data. Let γ_j denote stratum-specific log odds, defined as $\gamma_j = \log(\mathsf{P}(D = 1|S = j)/\mathsf{P}(D = 0|S = j))$, define $p_1(j) = \mathsf{P}(D = 1|S = j)$ and $p_0(j) = 1 - p_1(j)$. Furthermore, let $p_{1j}(x) = \mathsf{P}(D = 1|S = j, x, \text{Sample 2})$ denote the probability of sampling a "case" with covariate x from stratum j of the second phase sample with n_{1j} "cases" and n_{0j} "controls", $p_{0j}(x) = 1 - p_{1j}(x)$. Then

$$p_1(j) = F(\gamma_j)$$

and

$$p_{1j}(x) = F\left(\log\frac{n_{1j}}{n_{0j}} - \gamma_j + \alpha + x^\top\beta\right).$$

By the method of Schill *et al.* (1993), $\theta_{\mathrm{PL}} = (\gamma^\top, \alpha, \beta^\top)^\top$ is estimated by $\widehat{\theta}_{\mathrm{PL}}$, which is obtained by maximizing the pseudo likelihood L_{PL} of the two-phase setup:

$$L_{\mathrm{PL}} = \prod_i \prod_j \left\{ p_i(j)^{N_{ij}} \prod_k p_{ij}(x_{jk})^{n_{ijk}} \right\}.$$

We note a minor change in the parametrization of the marginal model compared to Schill and Drescher (1997).

The method of Breslow and Cain (1988) estimates θ_{PL} in two steps. First, the pseudo likelihood contributions of the first phase data are maximized, giving estimates

$$\widehat{\gamma}_j = \log\frac{N_{1j}}{N_{0j}},\ j = 1, \ldots, J.$$

In the second step these estimates are plugged into the pseudo likelihood contributions of the second phase sample , i. e., the remaining parameters of θ_{PL} are estimated by maximizing

$$L_{\mathrm{BC}} = \prod_i \prod_j \prod_k p^*_{ij}(x_{jk})^{n_{ijk}},$$

where $p^*_{1j}(x) = 1 - p^*_{0j}(x) = F\left(\log\frac{N_{0j}n_{1j}}{N_{1j}n_{0j}} + \alpha + x^\top\beta\right)$. The resulting estimate is $\widehat{\theta}_{\mathrm{BC}}$.

Implementation. The pseudo likelihood methods use SAS/STAT procedure `proc logistic` to estimate the respective parameters. SAS/IML matrix language is used to compute the parameter covariance matrices.

3.3. Maximum likelihood via the EM algorithm

To compute ML estimates, Schill and Drescher (1997) justified the use of the EM algorithm (Dempster, Laird, and Rubin 1977) applied to a Poisson likelihood. In this approach, the (possibly unobserved) counts, N_{ijk} say, are Poisson distributed with expectation

$$\mu_{ijk} = \begin{cases} \exp(\delta_{jk} + \alpha + x_{jk}^\top \beta) & \text{if } i = 1 \\ \exp(\delta_{jk}) & \text{if } i = 0 \end{cases} . \tag{2}$$

In the E-step the unobserved cell counts N_{ijk} are replaced by their expectations conditional on the observed data n_{ijk} and the current estimates of the parameters, giving

$$\widehat{N}_{ijk} = n_{ijk} \quad + \quad (N_{ij} - n_{ij}) \, (\widehat{\mu}_{ijk} / \widehat{\mu}_{ij+}) \, . \tag{3}$$

The M-step then maximizes the Poisson likelihood as if the $\widehat{N}_{ijk}$ were the complete data. The parameter to be estimated is $\theta_{\text{ML\_EM}} = (\delta^\top, \alpha, \beta^\top)^\top$, $\delta^\top = (\delta_{11}, \ldots, \delta_{JK_J})$. δ represents a discrete parametrization of the covariate distribution of X in $\{D = 0\}$ and can be of high dimension if the second phase data are extensive with a wide variety of covariate patterns. The cost of using this extensively parameterized model is purely computational (Scott and Wild 1991), especially no efficiency loss in estimating α and β is incurred.

Implementation The EM algorithm is written entirely in SAS/IML. To speed up convergence, the proposal of Louis (1982) is implemented. Choosing good starting values for the Poisson model is important. The algorithm works as follows:

- Set as starting value $\theta^0 = (\alpha^0, \beta^0, \delta^0)$, where $\alpha^0 = \log(N_{1+}/N_{0+})$, β^0 is set to zero and the components of δ^0 are set to $\log(N_{0+}/\sum_j K_j)$, the log mean prevalence of covariate patterns in "controls". Let X denote the design matrix of the Poisson model (Equation 2), compute $\widehat{\mu} = \exp(X\theta^0)$.
- E-step: Compute "complete counts" $\widehat{N}_{ijk}$ according to Equation 3.
- M-step: Maximize the Poisson likelihood to obtain a new estimate of $\theta_{\text{ML\_EM}}$.
- Iterate E- and M-step until convergence.

The asymptotic variance-covariance matrix of $\widehat{\theta}_{\text{ML\_EM}}$ is a by-product of the algorithm.

3.4. Maximum likelihood via profile likelihood

To avoid estimating a potentially high dimensional nuisance parameter, Scott and Wild (1997) calculated the profile likelihood to obtain ML estimates for α and β of the underlying model in Equation 1. They derived an iterative cycle based on a pseudo likelihood: The approach fits a logistic regression model (the pseudo model) to the *phase two data* where the pseudo model includes stratum specific offsets that are updated at each cycle. The probabilities $p^\star_{ijk}$ of the pseudo model are

$$p^\star_{1jk} = 1 - p^\star_{0jk} = F\left(\log \frac{\kappa_{1j}}{\kappa_{0j}} + \alpha + x_{jk}^\top \beta\right), \tag{4}$$
$$j = 1, \ldots, J, \; k = 1, \ldots, K_j.$$

The κ_{ij} are computed as

$$\begin{aligned} \kappa_{ij} &= \frac{n_{ij} - \gamma_{ij}}{N_{ij} - \gamma_{ij}}, \\ \gamma_{ij} &= n_{ij} - \sum_k n_{+jk} p^{\star}_{ijk}. \end{aligned} \tag{5}$$

The parameter to be estimated is $\theta_{\text{ML\_SW}} = (\alpha, \beta^\top)^\top$. Note that, if in stratum j say, phase one and phase two sample sizes agree, i. e., $n_{0j} = N_{0j}$ and $n_{1j} = N_{1j}$, the offset for this stratum is zero.

Implementation. Estimation of $\theta_{\text{ML\_SW}}$ is implemented as an iterated sequence of data- and proc steps of SAS/BASE- and SAS/STAT software:

- Start the algorithm with the Breslow-Cain approach, i. e., choose as offsets $\log\left(\frac{n_{1j}N_{0j}}{n_{0j}N_{1j}}\right)$, $j = 1, \ldots, J$, and apply the pseudo model (Equation 4) to the phase two data.
- Update offsets via Equations 5.
- Estimate α and β using the pseudo model.
- Iterate until convergence.

The asymptotic variance-covariance matrix of $\widehat{\theta}_{\text{ML\_SW}}$ is computed in SAS/IML.

3.5. Interrelations between methods

Depending on model specification, stratification and recruitment some relations between methods may be established.

- If the model includes the stratum variable S as a factor, i. e., S is parameterized via dummy variables, the two pseudo likelihood methods agree and give the ML estimates of $(\alpha, \beta^\top)^\top$.
- If the sampling fractions n_{ij}/N_{ij} are constant, the WL- and BC-estimates of α and β agree.
- In the complete data case we have $N_{ij} = n_{ij}$ for all i and j and weighted likelihood and the Breslow-Cain method yield ML estimates: $(\widehat{\alpha}_{\text{WL}}, \widehat{\beta}^\top_{\text{WL}})^\top = (\widehat{\alpha}_{\text{BC}}, \widehat{\beta}^\top_{\text{BC}})^\top = (\widehat{\alpha}_{\text{ML}}, \widehat{\beta}^\top_{\text{ML}})^\top$.

If the validity of the non-differential error assumption is in question as in Example 1 (see below), the weighted likelihood approach should be chosen. However, if this assumption is fulfilled, the other approaches should be preferred and maximum likelihood is the most efficient method. Furthermore, if the recruitment fractions for the second phase sample vary widely over strata, the weighted likelihood approach can be badly inefficient (see Schill and Drescher 1997; Breslow and Holubkov 1997; Breslow and Chatterjee 1999). With respect to efficiency, there is not much to choose between the two pseudo likelihood methods: Because

of the above stated relationships, we prefer the Breslow-Cain method if the marginal, stratum specific parameters are of interest.

3.6. Retrospective first phase sample

If the first phase sample is retrospective (unmatched or frequency-matched case-control), the meaning of the intercept parameter(s) changes: In the case of an unmatched study, all methods estimate as intercept a parameter $\alpha_0 = \alpha + \log(\mathsf{P}(D = 1)/\mathsf{P}(D = 0))$ instead of α. In this case, an offset $\log(N_{1+}/N_{0+})$ has been added to the linear predictor.

4. The sas-twophase-package

The folder **sas-twophase-package** contains four directories:

- `data` – provides datasets for the enclosed examples,
- `documentation` – contains `TPDocu.pdf` (documentation and usage) and `TPMethods.pdf` (methods and implementation),
- `examples` – contains three SAS programs that execute the examples given in `TPDocu.pdf`,
- `macros` – `twophase.sas` is the calling macro that must be included in the SAS program. The folder contains also a preparatory program, some utility programs to enhance screen output and the macros that perform the estimation methods. All of these macros are called by `twophase`.

Usage. We give a brief introduction to using the package, for a detailed description refer to `TPDocu.pdf` in the documentation folder. The package has to be stored somewhere on your computer. It expects separate first and second phase datasets, where the first phase dataset is a cross-tabulation of outcome and stratum variable. The second phase dataset contains outcome, stratum and regressor variables and an optional weight variable. To employ the various estimation methods, `%include` the program `twophase.sas` into a SAS program and run the macro `twophase` by providing appropriate arguments. The results are saved in the work directory and can optionally be viewed on the output screen.

5. Examples

The examples were executed with SAS 9.2 on a PC under Windows XP. The datasets of the examples are part of the data-folder of **sas-twophase-package**.

Example 1. Carroll, Gail, and Lubin (1993) consider the problem of estimating the odds ratio of a disease D in a case-control study where the binary exposure X is measured with error. Table 1 presents data where exposure to Herpes Simplex Virus Type 2 (HSV-2) is measured by a refined western blot procedure (X) and a less refined blot procedure (Z) in women with cervical cancer ($D = 1$) and in controls ($D = 0$).

Complete data				Incomplete data		
D	X	Z	Count	D	Z	Count
1	0	0	13	1	0	318
1	0	1	3	1	1	375
1	1	0	5	0	0	701
1	1	1	18	0	1	535
0	0	0	33			
0	0	1	11			
0	1	0	16			
0	1	1	16			

Table 1: Case-control data of Carroll *et al.* (1993).

We are interested in the association of X and D, i. e., in estimating β from the logistic model $\mathsf{P}(D = 1|X = x) = 1/[1 + \exp(-\alpha - x\beta)]$ using the complete and incomplete data. We want to employ weighted likelihood and maximum likelihood.

In the following it is assumed that **sas-twophase-package** is stored on drive `G:`. The program `twophase.sas` has to be included into the SAS program and the location of the data has to be specified. The first lines of a program could look like

```
%let path_tp=%str(g:\sas-twophase-package\macros);
%include "&path_tp.\twophase.sas";
libname in "g:\sas-twophase-package\data";
```

Then some data manipulations are necessary to define the stratum variable, S say. Z is not valid since stratum variables are required to attain values $1, \ldots, J$. Note that the first phase data are the combination of complete and incomplete data of Table 1. A call of macro `twophase`, where WL- and both ML-estimates are requested, could look like the following:

```
%twophase(folder = &path_tp,
  path_ph1 = carr1, path_ph2 = carr2,
  methods = wl ml_em ml_sw,
  compare = 1, outest = 1,
  caco = d, svar = s ,
  counts_ph1 = count, weights_ph2 = count,
  regr = x);
```

The results are as follows:

```
          Weighted Regression   ML (EM-Algorithm)     ML (Profile Lik.)
             estim     stderr      estim     stderr      estim     stderr

X          0.60808    0.35034    0.95792    0.23659    0.95792    0.23659
_ALPHA    -0.31383    0.18685   -0.51352    0.14142   -0.51352    0.14142
```

The considerable difference between the estimated log odds ratios under WL ($\widehat{\beta}_{\mathrm{WL}} = 0.608$) and ML ($\widehat{\beta}_{\mathrm{ML}} = 0.958$) is noteworthy: We attribute this difference to the fact that the non-differential error assumption (Section 3) is probably not met. In fact, as Carroll *et al.* (1993)

	Stratum							
	1	2	3	4	5	6	7	8
Cases	135	19	32	28	360	80	89	96
Controls	347	62	42	42	210	48	46	42

Table 2: HdA first phase data.

noted, there is a substantial amount of misclassification in these data and the sensitivity of Z as a measure of X seems to be higher among cases: $\mathsf{P}(Z = 1|X = 1, D = 0) = 0.500$, $\mathsf{P}(Z = 1|X = 1, D = 1) = 0.783$, $p = 0.049$ by Fisher's exact test. Since the WL method does not require the conditional independence assumption, one would rather rely on the WL estimate.

Example 2. This example presents an analysis with a continuous covariate. The data are derived from the "HdA study", a two-phase case-control study of Pohlabeln *et al.* (2002) with focus on lung cancer risk due to the intensity of occupational asbestos exposure. However, a precise intensity assessment in terms of asbestos fibreyears was affordable only for a 20% subsample ($n_{0+} = n_{1+} = 164$) of the original study ($N_{0+} = N_{1+} = 839$) . For the original study the duration of occupational asbestos exposure was known (see Pohlabeln *et al.* (2002) for details). We want to derive a smoking-adjusted effect of fibreyears on lung cancer incidence.

In this example, we take as stratum variable S the variable `STRATA`, constructed as cross-tabulation of 'duration of asbestos exposure' (4 levels) and 'heavy smoking' (2 levels). `'STRATA=1'` then represents 'non- or mild smokers' & 'duration of asbestos exposure=0', ..., `'STRATA=4'` represents 'non- or mild smokers' & 'long duration', ... `'STRATA=8'` is the group of 'medium- or heavy smokers' & 'long duration of exposure' (Table 2).

The second phase dataset has additional information on smoking history (variable `SMOKE` with levels 0 (non-smoker), 1 (mild smoker), 2 (medium smoker) and 3 (heavy smoker)) and the continuous variable `FY`, log(asbestos fibreyears+1). `FY` is deemed an appropriate cumulative exposure measure in occupational epidemiology. We want to fit the logistic model $\mathsf{P}(\texttt{CASE = 1}|\texttt{FY, SMOKE}) = F\left(\alpha + \beta_1\texttt{SMOKE1} + \cdots + \beta_3\texttt{SMOKE3} + \beta_{\texttt{FY}}\texttt{FY}\right)$.

We have to perform the same preparatory steps as in Example 1. To demonstrate the meaning of the marginal stratum parameters in the pseudo likelihood approaches, we include the Breslow-Cain method. Furthermore, to illustrate the efficiency gain of the two-phase methods over a so-called ''complete-case" analysis, where only the data of the second phase sample are analyzed, the appropriate methods are chosen. A call of macro `twophase` could look like:

```
%twophase(folder = &path_tp,
  path_ph1 = hdac1, path_ph2 = hdac2,
  methods = pl_bc ml_em s2,
  compare = 1, outest = 1,
  caco = case, svar = strata,
  counts_ph1 = count, weights_ph2 = count,
  regr = smoke1 smoke2 smoke3 fy);
```

The results are as follows:

	PL (Breslow-Cain)		ML (EM-Algorithm)		Sample2-Analysis	
	estim	stderr	estim	stderr	estim	stderr
FY	0.13211	0.07039	0.16389	0.05739	0.14560	0.08239
SMOKE1	0.94050	0.54140	0.84504	0.54383	0.89373	0.53473
SMOKE2	1.98080	0.47861	1.93990	0.47946	2.01755	0.52330
SMOKE3	2.41971	0.50837	2.40276	0.50382	2.44167	0.55246
_ALPHA	-1.62627	0.45578	-1.61585	0.45468	-1.71063	0.47797
_S1	-0.94405	0.08891	.	.	.	.
_S2	-1.18270	0.25764	.	.	.	.
_S3	-0.27193	0.22951	.	.	.	.
_S4	-0.40547	0.23904	.	.	.	.
_S5	0.53900	0.07180	.	.	.	.
_S6	0.51083	0.17592	.	.	.	.
_S7	0.65999	0.17490	.	.	.	.
_S8	0.82668	0.17844	.	.	.	.

As we have also selected the (default) Breslow-Cain method, estimates of the marginal, stratum log odds are displayed: For instance, `_S4 - _S1` $= 0.539$ is an estimate of $\gamma_4 - \gamma_1$, the marginal log odds ratio of 'long asbestos duration' vs. 'no asbestos duration' among 'non- or mild smokers'. As noted in Pohlabeln *et al.* (2002), the complete-case analysis yields also valid estimates in this example since recruitment into the phase two sample did not depend on phase one variables. By comparing the results of two-phase- and complete case analyses, one can see the information gain due to including the first phase data.

6. Discussion

The methods implemented in this package require a given stratification of the first and second phase data. Stratification is a separate, important topic in the design of two-phase studies. In Example 2, for instance, the stratification can be criticized because it not fully employs the smoking information in the first phase data, which in fact would be available from the original case-control study. The rationale for choosing the stratification based on `STRATA` is to include into the stratification as much first phase information as possible on asbestos exposure, which is the focus of the analysis. A cross-classification with a 4-level smoking variable would lead to empty strata, a situation that the methods cannot cope with. However, a more sophisticated stratification is possible: It assembles all non-smokers into one stratum and then cross-tabulates the remaining with respect to 'smoking' (3 levels) and 'duration of asbestos exposure' (4 levels). Not only are the smoking parameters estimated more precisely, but also a slightly more precise fibreyear estimate is obtained (data not shown).

Which variables should be used for stratification? The second phase data of Example 2 are certainly not typical since they represent a systematic sub-sample of a case-control study in that all enrollments of two years of follow-up were included. The two-phase design is especially effective when 'rare exposures' can be over-sampled. Furthermore, if the second phase sample arises by means of a survey, participation may depend on some other personal characteristics like, for example, sex or age. Even if the outcome model would not include such variables,

relevant exposures could be associated with these. In this case, such variables should also be used for stratification. An illustrative example that also highlights some other topics is provided by the pharmaco-epidemiological study of Behr, Schill, and Pigeot (2012).

We have implemented two approaches to obtain semi-parametric maximum likelihood estimates. The EM algorithm uses a non-parametric estimate of the covariate distribution, which can involve a large number of nuisance parameters, especially if continuous covariates are included in the outcome model. Although Louis' idea to speed up the algorithm is implemented and works fine, the algorithm can still be time- and resource-consuming: In Example 2, for instance, EM had to estimate approximately 150 parameters and used about 3 seconds on a fast PC. In a series of computations around the paper of Behr *et al.* (2012), we generated second phase datasets of size $n = 500$, 1000, 2000 and 10000 with continuous covariates and different stratifications. Let n_δ denote the number of nuisance parameters, i. e., the number of distinct covariate patterns in the second phase dataset. With $n = 500$, we had $n_\delta \approx 480$ and the algorithm took approximately 3 minutes, with $n = 1000$, $n_\delta \approx 890$ and approximately 25 minutes were necessary. With $n = 2000$, $n_\delta \approx 1560$ and EM took 2–3 hours. With $n = 10000$, the algorithm failed due to lack of memory. In conclusion, this form of the EM algorithm shows acceptable performance for small or medium numbers of covariate patterns in the second phase data. In other instances, the approach based on the profile likelihood should be chosen. Like the weighted-and pseudo likelihood methods, nuisance parameters need not to be estimated. Due to its implementation as a repeated cycle of data- and proc steps, however, the algorithm usually needs a few seconds.

Estimating regression parameters by maximum likelihood for binary data in a two-phase setup is obviously not an easy task. Our package implements ML estimation by the EM algorithm and via a pseudo model derived from the profile likelihood. The `bin2stg()` function of the **missreg** package also uses the profile likelihood to obtain semi-parametric ML estimates. The `tps()` function of the **osDesign** package solves a constrained maximization problem to obtain ML estimates, the approach has been shown to be equivalent to the profile likelihood solution (Scott, Lee, and Wild 2007). However, for certain data constellations `tps()` does not produce ML estimates. Moreover, we observed slight differences in the standard error estimates of `tps()` compared to `bin2stg()` or our package (the latter agree). In case of conflict we judge results of `bin2stg()` as gold standard.

Acknowledgments

This work was partly supported by Deutsche Forschungsgemeinschaft, grant PI 345/5-1. We thank Marcus Seiler for work on a previous version of the package. We also thank Sigrid Behr for testing and for carefully reading the manuscript.

References

Behr S, Schill W, Pigeot I (2012). "Does Additional Confounder Information Alter the Estimated Risk of Bleeding Associated with Phenprocoumon Use – Results of a Two-Phase Study." *Pharmacoepidemiology and Drug Safety*, **21**(5), 535–545.

Breslow NE (2005). "Case-Control Study, Two-Phase." In P Armitage (ed.), *Encyclopedia of Biostatistics*, pp. 734–741. John Wiley & Sons.

Breslow NE, Cain KC (1988). "Logistic Regression for Two-Stage Case-Control Data." *Biometrika*, **75**, 11–20.

Breslow NE, Chatterjee N (1999). "Design and Analysis of Two-Phase Studies with Binary Outcome Applied to Wilms Tumour Prognosis." *Journal of the Royal Statistical Society C*, **48**(4), 457–468.

Breslow NE, Holubkov R (1997). "Maximum Likelihood Estimation of Logistic Regression Parameters under Two-Phase, Outcome Dependent Sampling." *Journal of the Royal Statistical Society B*, **59**, 447–461.

Carroll RJ, Gail MH, Lubin JH (1993). "Case-Control Studies with Errors in Covariates." *Journal of the American Statistical Association*, **88**, 185–199.

Dempster AP, Laird NN, Rubin DB (1977). "Maximum Likelihood from Incomplete Data via the EM-Algorithm." *Journal of the Royal Statistical Society B*, **39**, 1–38.

Flanders D, Greenland S (1991). "Analytic Methods for Two-Stage Case-Control Studies and Other Stratified Designs." *Statistics in Medicine*, **10**, 739–747.

Haneuse S, Saegusa T, Lumley T (2011). "**osDesign**: An R Package for the Analysis, Evaluation, and Design of Two-Phase and Case-Control Studies." *Journal of Statistical Software*, **43**(11), 1–29. URL http://www.jstatsoft.org/v43/i11/.

Horvitz D, Thompson D (1952). "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of of the American Statistical Association*, **47**(11), 663–685.

Louis T (1982). "Finding the Observed Information Matrix When Using the EM Algorithm." *Journal of the Royal Statistical Society B*, **44**, 226–233.

Lumley T (2004). "Analysis of Complex Survey Samples." *Journal of Statistical Software*, **9**.

Pohlabeln H, Wild P, Schill W, Ahrens W, Jahn I, Bolm-Audorff U, Jöckel KH (2002). "Asbestos Fibreyears and Lung Cancer: A Two-Phase Case-Control Study with Expert Exposure Assessment." *Occupational and Environmental Medicine*, **59**, 410–414.

Reilly M, Pepe MS (1995). "A Mean Score Method for Missing and Auxilliary Covariate Data in Regression Models." *Biometrika*, **82**, 299–314.

SAS Institute Inc (2008a). *SAS/IML 9.2 User's Guide*. SAS Institute Inc., Cary, NC. URL http://www.sas.com/.

SAS Institute Inc (2008b). *SAS/STAT 9.22 User's Guide*. SAS Institute Inc., Cary, NC. URL http://www.sas.com/.

Schill W, Drescher K (1997). "Logistic Analysis of Studies with Two-Stage Sampling: A Comparison of Four Approaches." *Statistics in Medicine*, **16**, 117–132.

Schill W, Jöckel KH, Drescher K, Timm J (1993). "Logistic Analysis in Case-Control Studies under Validation Sampling." *Biometrika*, **80**, 339–352.

Schill W, Wild P, Pigeot I (2007). "A Planning Tool for Two-Phase Case-Control Studies." *Computer Programs and Methods in Biomedicine*, **88**, 175–181.

Scott AJ, Lee AJ, Wild CJ (2007). "On the Breslow-Holubkov Estimator." *Lifetime Data Analysis*, **13**, 545–563.

Scott AJ, Wild J (1991). "Fitting Logistic Regression Models in Stratified Case-Control Studies." *Biometrics*, **47**, 497–510.

Scott AJ, Wild CJ (1997). "Fitting Regression Models to Case-Control Data by Maximum Likelihood." *Biometrika*, **84**(1), 57–71.

Scott AJ, Wild CJ (2006). "Calculating Efficient Semiparametric Estimators for a Broad Class of Missing-Data Problems." In EP Liski, J Isotalo, J Niemelä, S Putanen, GPH Styan (eds.), *Festschrift for Tarmo Pukkila on his 60th birthday*, pp. 301–314. University of Tampere.

Affiliation:

Walter Schill, Dirk Enders
Leibniz Institute for Prevention Research and Epidemiology – BIPS
Division of Biometry
Achterstr. 30
28359 Bremen, Germany
E-mail: schill@bips.uni-bremen.de, enders@bips.uni-bremen.de

Karsten Drescher
Statistical Office of Bremen
An der Weide 14–16
28195 Bremen, Germany
E-mail: Karsten.Drescher@statistik.bremen.de

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 57, Issue 9 *Submitted:* 2012-11-06
April 2014 *Accepted:* 2013-11-08