Documentation
sas-twophase-package
Methods and Implementation

Author: **Walter Schill, Karsten Drescher, Dirk Enders**
Date: **17.05.2013**
Version: **1**

Contents

1 General **2**

2 Notation **2**

3 Methods **2**
3.1 Weighted likelihood 3
3.2 Pseudo likelihood 3
3.3 Maximum likelihood via the EM algorithm 4
3.4 Maximum likelihood via profile likelihood 5
3.5 Interrelations between methods 6
3.6 Retrospective first phase sample 6

4 Implementation **6**
4.1 Weighted-, pseudo- and profile likelihood 7
4.2 EM algorithm 7

1 General

This document gives notation, summarizes methodology and describes implementation of two-phase methods as presented in Schill and Drescher (1997) and Scott and Wild (1997), as programmed in separate macros of the `sas-twophase-package`. Model-based covariance matrices are provided. `SAS/STAT`- and `SAS/IML` software (SAS Institute Inc.) must be available in the computing environment.

2 Notation

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

$$\Pr(D = 1|X = x) = \frac{\exp(\alpha + x^T\beta)}{1 + \exp(\alpha + x^T\beta)}, \tag{2.1}$$

where x denotes a $p \times 1$ vector including exposures, covariates and interactions. Focus lies on inference about the log odds ratio parameter $\beta = (\beta_1, \ldots, \beta_p)^T$.
The phase one sample comprises data on outcome $D \in \{0, 1\}$ and measurements of partial or proxy information Z 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 phase one 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 phase one data for covariate ascertainment. For 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

Regression methods based on weighted likelihood, pseudo likelihood and maximum likelihood are presented. We first describe the likelihood functions for prospective phase one samples, the extension to retrospective samples is given in section 3.6. Different sets of parameters are estimated with the diverse methods. For a description of the (model-based) parameter covariance matrices, refer to the underlying articles.

Assumptions **With the exception of the weighted likelihood method, all other estimation methods require that stratum variable S and outcome D are conditionally independent, given $X = x$: $\Pr(D = 1|S = j, x) = \Pr(D = 1|x)$,**

that is, outcome probabilities depend on S only through X. 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 may be violated.

3.1 Weighted likelihood

The idea of this approach (Flanders and Greenland, 1991; Reilly and Pepe, 1995) stems from survey research (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 $\theta_{\mathrm{WL}} = (\alpha, \beta^T)^T$ is estimated by $\widehat{\theta}_{\mathrm{WL}}$, which is obtained by maximizing

$$L_{\mathrm{WL}} = \prod_i \prod_j \prod_k \Pr(D = i \mid x_{jk})^{\left(\frac{N_{ij}}{n_{ij}} n_{ijk}\right)}. \tag{3.2}$$

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 $\gamma_j = \log(\Pr(D = 1|S = j)/\Pr(D = 0|S = j))$, define $p_1(j) = \Pr(D = 1|S = j)$. Furthermore, let $p_{1j}(x) = \Pr(D = 1|S = j, x, \text{Sample 2})$ denote the probability of sampling a "case" from stratum j of the second phase sample with n_{1j} "cases" and n_{0j} "controls" with covariate x. Let $F(u) = 1/[1 + \exp(-u)]$ denote the standard logistic distribution function.
Then

$$\begin{aligned} p_1(j) &= 1 - p_0(j) = F(\gamma_j), \\ p_{1j}(x) &= 1 - p_{0j}(x) = F\left(\log \frac{n_{1j}}{n_{0j}} - \gamma_j + \alpha + x^T\beta\right), \\ j &= 1, \ldots, J. \end{aligned}$$

$\theta_{\mathrm{PL}} = (\gamma^T, \alpha, \beta^T)^T$ is estimated by maximizing the pseudo likeliood L_{PL} of the two-phase setup (Schill et al., 1993):

$$L_{\mathrm{PL}} = \prod_i \prod_j \left\{ p_i(j)^{N_{ij}} \prod_k p_{ij}(x_{jk})^{n_{ijk}} \right\}. \tag{3.3}$$

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 $\theta_{\mathrm{BC}} = (\gamma^T, \alpha, \beta^T)^T$ 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 Sample 2, i. e., the remaining parameters of θ_{BC} are obtained by maximizing

$$L_{\mathrm{BC}} = \prod_i \prod_j \prod_k \widetilde{p}_{ij}(x_{jk})^{n_{ijk}}, \tag{3.4}$$

where $\widetilde{p}_{1j}(x) = 1 - \widetilde{p}_{0j}(x) = F\left(\log \frac{n_{1j} N_{0j}}{n_{0j} N_{1j}} + \alpha + x^T \beta\right)$.

3.3 Maximum likelihood via the EM algorithm

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

$$\mu_{ijk} = \begin{cases} \exp(\delta_{jk} + \alpha + x_{jk}^T \beta) & \text{if } i = 1 \\ \exp(\delta_{jk}) & \text{if } i = 0 \end{cases}. \tag{3.5}$$

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

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

The M-step then maximizes the Poisson likelihood, as if the $\widehat{N}_{ijk}$ were the complete data. The parameter to be estimated is $\theta_{\mathrm{ML\_EM}} = (\delta^T, \alpha, \beta^T)^T$, $\delta^T = (\delta_{11}, \ldots, \delta_{JK_J})$. δ represents a discrete parametrization of the covaariate distribution 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), meaning that no efficiency loss in estimating α and β is incurred. However, in case of a high-dimensional δ, the algorithm can be painfully slow or even fail due to lack of memory, because high-dimensional matrices need to be inverted (see Secton 4.2).

3.4 Maximum likelihood via profile likelihood

From the profile likelihood, Scott and Wild (1997) derive 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

$$p^{\star}_{1jk} = 1 - p^{\star}_{0jk} = F\left(\log\frac{\kappa_{1j}}{\kappa_{0j}} + \alpha + x^{T}_{jk}\beta\right), \quad j = 1,\ldots,J,\ k = 1,\ldots,K_j. \tag{3.6}$$

The κ_{ij} are computed as

$$\kappa_{ij} = \frac{n_{ij} - \gamma_{ij}}{N_{ij} - \gamma_{ij}}, \tag{3.7}$$

$$\gamma_{ij} = n_{ij} - \sum_{k} n_{+jk} p^{\star}_{ijk}. \tag{3.8}$$

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.

Estimation algorithm. **The parameter to be estimated is $\theta_{\text{ML\_SW}} = (\alpha, \beta^{T})^{T}$.**

- **P1** **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 to the phase two data.**
- **P2** **Update offsets via Equations (3.8) and (3.7).**
- **P3** **Estimate α and β using the pseudo model.**
- **P4** **Go to P2 until convergence.**

Asymptotic variance covariance matrix **The computation of the asymptotic variance-covariance matrix is based on Equation (18) of Scott and Wild. Since we are concerned with the binary logistic outcome model, their formulae reduce to simpler expressions. The information matrix I is given as**

$$I = I^{\star} + \sum_{j:\, a_j \neq 0} B_j B_j^{T} / k_j, \tag{3.9}$$

where $I^\star$ is the information matrix of the pseudo model. The other quantities are computed as:

$$
\begin{aligned}
B_j &= \sum_{x\in \text{ stratum } j} \frac{\partial P^\star(x)}{\partial \theta} = \sum_{k=1}^{K_j} n_{+jk} x_{jk} p^\star_{1jk} p^\star_{0jk}, \\
k_j &= \begin{cases} 1/a_j - w_j, \text{ if } a_j \neq 0, \\ \text{not defined otherwise,} \end{cases} \\
a_j &= \frac{1}{n_{0j} - \gamma_{0j}} - \frac{1}{N_{0j} - \gamma_{0j}} + \frac{1}{n_{1j} - \gamma_{1j}} - \frac{1}{N_{1j} - \gamma_{1j}}, \\
w_j &= \sum_{x\in \text{ stratum } j} (P^\star(x) - P^{\star 2}(x)) = \sum_{k=1}^{K_j} n_{+jk} p^\star_{1jk} p^\star_{0jk}.
\end{aligned}
$$

Note that strata j with $a_j = 0$ do not provide an increment in information. This is due to the fact that for these strata phase one and phase two sample sizes agree: $n_{0j} = N_{0j}$ and $n_{1j} = N_{1j}$ (see above).

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, the two pseudo likelihood methods agree and give the ML estimates of $(\alpha, \beta^T)^T$.**
- **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, $N_{ij} = n_{ij}$ and weighted likelihood and the Breslow-Cain method yield ML estimates: $(\widehat{\alpha}_{\text{WL}}, \widehat{\beta}^T_{\text{WL}})^T = (\widehat{\alpha}_{\text{BC}}, \widehat{\beta}^T_{\text{BC}})^T = (\widehat{\alpha}_{\text{ML}}, \widehat{\beta}^T_{\text{ML}})^T$.**

3.6 Retrospective first phase sample

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

4 Implementation

The preparatory program `prep.sas` reads the separate first and second phase data and outputs a combined, restructured dataset `prep`, sorted by stratum

and covariate pattern. This dataset serves as input to all estimation macros (refer to the document `TPDocu.pdf`, Section 5, in this folder).

4.1 Weighted-, pseudo- and profile likelihood

The weighted likelihood and pseudo likelihood implementations are `wlrmacro.sas`, `bc_macro.sas`, `zlrmacro.sas` and `sw_macro.sas`. These programs use `SAS/STAT` software `proc logistic` to obtain parameter estimates. To compute the (model-based) covariance matrices `proc iml` is invoked: data and parameter estimates are read into `iml` to obtain the appropriate design matrices and linear predictors. The two-phase covariance matrices are calculated; finally, an adjustment is made if the first phase sample is retrospective.

4.2 EM algorithm

Define $\theta = \theta_{\text{ML\_EM}}$ and let X denote the design matrix of the Poisson model (3.5), so that $\mu = \exp(X\theta)$. X has dimension $2n_x \times (1 + p + n_x)$, where $n_x = \sum_j K_j$ denotes the number of distinct covariate patterns.

Parameter starting values. **The algorithm is started with $\alpha = \log(N_1/N_0)$, where N_i denote the numbers of the outcome groups in phase one, β is set to 0 and the components of δ are set to $\log(N_0/n_x)$, i. e. log mean prevalence of covariate pattern in $\{D = 0\}$.**

Iterations. **The algorithm iterates E- and M-step until $\sum_l |\widehat{\theta}_l^{(t+1)} - \widehat{\theta}_l^{(t)}| < \epsilon$ with a specified ϵ.**

- **E-Step**
  (a) **Current expected counts: $\widehat{\mu}^{(t)} = \exp(X\widehat{\theta}^{(t)})$**
  (b) **Current weights (per stratum and outcome group): $\widehat{\mu}_{ijk}^{(t)}/\widehat{\mu}_{ij+}^{(t)}$**
  (c) **Current completed observations (per stratum and outcome group): $\widehat{N}_{ijk}^{(t)} = n_{ijk} + (N_{ij} - n_{ij}) \times \widehat{\mu}_{ijk}^{(t)}/\widehat{\mu}_{ij+}^{(t)}$.**
- **M-Step**
  The M-Step implements Louis (1982) idea for speeding up convergence of the EM algorithm: the first cycles use one step of the standard Newton-Raphson algorithm, the remaining steps implement Louis' modification. In any case, if the deviance increases after a step, the step length is halved. The deviance is $2\sum(\widehat{N}\log\widehat{N} - \widehat{\mu}\log\widehat{\mu} - \widehat{N} + \widehat{\mu})$, where summation is over i, j, k.
  Specifically

– **For the first 10 iterations, one Newton-Raphson step is performed. This gives**

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \widehat{\boldsymbol{\theta}}^{(t)} + [\boldsymbol{X}^T \mathbf{Diag}(\widehat{\boldsymbol{\mu}}^{(t)})\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\widehat{\boldsymbol{N}}^{(t)} - \widehat{\boldsymbol{\mu}}^{(t)})$$

– **After 9 iterations, Louis' idea is implemented via**

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \widehat{\boldsymbol{\theta}}^{(t)} + \mathbf{Cov}(\widehat{\boldsymbol{\theta}}^{(t)})\boldsymbol{X}^T(\widehat{\boldsymbol{N}}^{(t)} - \widehat{\boldsymbol{\mu}}^{(t)}).$$

Here, $\mathbf{Cov}(\widehat{\theta})$ is the parameter covariance matrix under EM, given as

$$\mathbf{Cov}(\widehat{\boldsymbol{\theta}}) = (\boldsymbol{X}^T[\mathbf{Diag}(\widehat{\boldsymbol{\mu}}) - \boldsymbol{D}]\boldsymbol{X})^{-1},$$

where $\boldsymbol{D}$ is a block-diagonal matrix built from the $2 \times J$ elements $\boldsymbol{D_{ij}} = (\boldsymbol{N_{ij}} - \boldsymbol{n_{ij}})(\mathbf{Diag}(\widehat{\pi}_{ij}) - \widehat{\pi}_{ij}\widehat{\pi}_{ij}^T)$ with $\widehat{\pi}_{ij}^T = (\widehat{\mu}_{ij1}, \ldots, \widehat{\mu}_{ijK_j})/\widehat{\mu}_{ij+}$.

Realization in SAS (em5macro.sas). **The data come prepared from the preparation program `prep.sas`. The first J observations represent the phase one data, the remaining n_x observations represent the phase two data, sorted by stratum and covariate pattern. Each of these n_x observations has**

- **the stratum $S \in \{1, \ldots, J\}$,**
- **the value of the regression variables x_{jk} (p columns), forming a covariate pattern,**
- **the number of cases (n_{1jk}) and controls (n_{0jk}) with this covariate pattern,**
- **the number of cases (n_{1j}) and controls (n_{0j}) in the respective stratum and the number of phase one counts per stratum (N_{1j} and N_{0j}).**

In a first preparation step, $(N_{ij} - n_{ij})$ are calculated and a variable Nr, giving a complete enumeration of the covariate patterns, is created. The next step "doubles" the phase 2 data in that separate observations for cases and controls are built. Furthermore, a second stratification variable S_2 is created which counts the number of cells in the phase one $2 \times J-$table.
The EM algorithm is performed using `SAS/IML` matrix language.
The following table gives a translation of terms from the above text to IML-variables:

Symbol	IML name	Description
Y	Y	disease status (0: controls)
N_0, N_1	N0, N1	number of controls and cases in phase 1
S	z	Stratum variable (values: 1 - J)
S_2	z2	Stratum variable, $S_2 = 2 * S - 1 + Y$, (values: 1 - $2S$)
x_{jk}	vars	covariate values (row vector per observation of dimension p)
	Nr	enumeration of pattern (values: 1 - n_x)
n_x	nxz	number of distinct covariate patterns
	m_com	phase 2 frequency of covariate pattern
n_{0jk}	m_com*(D=0)	among controls
n_{1jk}	m_com*(D=1)	among cases
$(N_{ij} - n_{ij})$	m_inc	number of observations with missing covariates (per S_2)

Weights. The weights are implemented via `B=Design(z2)`, which has the dummies of S_2 ($(i,j)-$cell in the phase 1 table.). Applying `BB=B*T(B)` to the vector of expected counts μ gives the desired (i,j)-cell sums (μ_{ij+}).

References

N.E. Breslow and K.C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75:11–20, 1988.

A.P. Dempster, N.N. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em-algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

D. Flanders and S. Greenland. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10:739–747, 1991.

D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of of the American Statistical Association*, 47(11):663–685, 1952.

M. Reilly and M.S. Pepe. A mean score method for missing and auxilliary covariate data in regression models. *Biometrika*, 82:299–314, 1995.

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

W. Schill and K. Drescher. Logistic analysis of studies with two-stage sampling: A comparison of four approaches. *Statistics in Medicine*, 16:117–132, 1997.

W. Schill, K.-H. Jöckel, K. Drescher, and J. Timm. Logistic analysis in case-control studies under validation sampling. *Biometrika*, 80:339–352, 1993.

A. J. Scott and C. J. Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57–71, 1997.

A.J. Scott and .J. Wild. Fitting logistic regression models in stratified case-control studies. *Biometrics*, 47:497–510, 1991.