Documentation
`sas-twophase-package:`
Macro `twophase`

Author: **D. Enders, W. Schill**
Date: **07.01.2014**
Version: **1**

Contents

Document: \sas-twophase-package\documentation\TPDocu.pdf

1 General

The purpose of the SAS-macro `twophase` is the conduct of logistic regression analyses with data in a two-phase setting. The estimation methods described in Schill and Drescher (1997) and Scott and Wild (1997) can be chosen. SAS/STAT- and SAS/IML software must be available on your computing environment.

2 Usage

The usage of the `twophase`-macro is as follows:

```
%twophase{ folder       = ,
           path_ph1     = ,
           path_ph2     = ,
           methods      = pl_bc,
           suffix       = ,
           compare      = 1,
           outproc      = 0,
           outcorr      = 0,
           outest       = 0,
           caco         = ,
           svar         = ,
           counts_ph1   = ,
           case         = 1,
           control      = 0,
           weights_ph2  = ,
           regr         = ,
           maxit        = 1000,
           epsil        = 1e-10,
           s1pr         = 0 }
```

3 Arguments

The following parameters can be set:

Parameter	Description
`folder`	Path of the folder containing all macros called by `twophase` . **Value**: string **Default**: none
`path_ph1`	Path of the phase one dataset. The value is of the form `path_ph1`=*libname.filename*. If the dataset is located in the workspace, libname can be omitted. **Value**: string

	Default: none
`path_ph2`	Path of the phase two dataset. See `path_ph1` for details. **Value**: string **Default**: none
`methods`	Specifies the estimation methods. Valid names are `ml_em`, `ml_sw`, `pl_bc`, `pl_sch`, `wl` and `s2`. Separate different names by blanks. The corresponding estimation methods are: `ml_em` - Maximum Likelihood via EM algorithm `ml_sw` - Maximum Likelihood (Profile Likelihood, Scott/Wild) `pl_sch` - Pseudo Likelihood (Schill) `pl_bc` - Pseudo Likelihood (Breslow-Cain) `wl` - Weighted Likelihood `s2` - Sample 2 (complete case) analysis **Value**: string **Default**: `pl_bc`
`suffix`	String to append on name of output dataset. **Value**: String **Default**: none
`compare`	If not set to zero, the estimated regression parameters and standard errors of all chosen methods are directed to the SAS-output. **Value**: numeric **Default**: 1
`outproc`	If not set to zero, parameter estimates, standard errors and confidence intervals of every method are directed to the SAS-output. **Value**: numeric **Default**: 0
`outcorr`	If not set to zero, the correlation matrix of every estimation method is directed to the SAS-output. Does only have an impact if `&outproc`=1. **Value**: numeric **Default**: 0
`outest`	If not set to zero, the estimates and standard errors of the phase one parameters are directed to the SAS-output, additional to the regression parameters. Does only have an impact if `&compare`=1. **Value**: numeric **Default**: 0
`caco`	Name of outcome variable. **Value**: string **Default**: None
`svar`	Name of stratum variable. `svar` must take on all the values $\mathbf{1, 2, \ldots, J}$ where $\boldsymbol{J}$ is the number of strata. **Value**: string **Default**: none
`counts_ph1`	Variable containing the observation counts in phase one.

	Value: string **Default**: none
`case`	Value of `&caco` (Case). **Value**: numeric **Default**: 1
`control`	Value of `&caco` (Control). **Value**: numeric **Default**: 0
`weights_ph2`	Optional variable in phase two containing weights for each observation. **Value**: string **Default**: none
`regr`	Names of the regression variables. The names must be separated by blanks. **Value**: string **Default**: none
`maxit`	Maximum number of iterations (whenever iterations occur in the estimation methods). **Value**: positive integer **Default**: 1000
`epsil`	Accuracy of calculations in the estimation methods. **Value**: numeric **Default**: $\mathbf{10^{-10}}$
`s1pr`	Indicates whether sampling in phase one is prospective or retrospective: `1`-prospective, `0`-retrospective. **Value**: 0/1 **Default**:0

4 Input data

Macro `twophase` expects separate phase one - and phase two datasets in a form described below. In a preparatory step, a rudimentary check on data consistency is performed: (a) the outcome variable should be in the specified range, (b) the stratum variable should be positive integer, (c) the count variable, if present, should be positive and (d) the covariates of the phase two dataset have to be non-missing. In case of violation, the respective records are removed for further processing. In any case the datasets are restructured and concatenated and in this form serve as input file for the estimation procedures. A list of variable names that should not be used in the input datasets is given in Section 5.

4.1 Phase one

The macro expects the phase one data to be cross-classified according to stratum and outcome status and reads three variables: `&caco` (outcome status), `&svar` (stratum) and `&counts_ph1` ("size" of group). `&caco` should be a dichotomous variable, `&svar` an enumeration of the strata and `&counts_ph1` a positive number, not necessarily an integer.

Example: Suppose the names are `&caco=D`, `&svar=STRAT` and `&counts_ph1 =COUNT`, `D` has values 0 (non-diseased) and 1 (diseased) and `STRAT` takes the values 1,2 and 3. Then a phase one dataset could be as follows:

D	STRAT	COUNT
0	1	20
0	2	12
0	3	12.5
1	1	25
1	2	6.56
1	3	77

Table 2: Exemplary phase one data.

Note that in a typical situation with observed data `COUNT` would be integer valued.

4.2 Phase two

The macro reads the variables `&caco`, `&svar`, the variables specified by `&regr` from the phase two dataset, and, if present, a positive, not necessarily integer-valued weight variable `&weight_ph2`. If `&weight_ph2` is not defined, a weight of 1 is assumed for observations in the phase two dataset. The variables specified in `&regr` enter in exactly this form into the linear predictor of the regression model. `twophase` has no capabilities to treat categorical variables as factors or to build interaction variables. This means that all data manipulation tasks like construction of dummies or building interaction terms have to be executed before the macro is called.

Example: Suppose `&regr=AGE SEX`, i. e. the regression will be performed with the covariates `AGE` and `SEX`. `AGE` has values 0 and 1 and `SEX` takes on the values 1 and 2. Then a phase two dataset could be as follows:

D	STRAT	AGE	SEX
0	1	0	1
0	1	0	1
0	2	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	3	1	2
1	2	1	1
0	1	1	1

Table 3: exemplary phase two data.

We could also have the same data in an aggregated form:

D	STRAT	AGE	SEX	WEIGHT
0	1	0	1	2
0	2	1	1	1
1	1	1	1	3
1	3	1	2	1
1	2	1	1	1
0	1	1	1	1

Table 4: exemplary phase two data, aggregated.

In the aggregated form we need to specify the weight variable `&weights_ph2= WEIGHT` in the call of `twophase`. Both versions of the phase two data are accepted as input data. A dataset is even accepted if it has duplicate observations with positive weights.

5 Processing and output data

The phase one data are restructured by keeping one observation per stratum with variables `_N0`, `_N1`, `_NSTRAT`, `_NZ0`, `_NZ1`, `_SAMPLE`, `_S`. For an explanation see Table 5 below.

The phase two dataset is sorted by `&caco &svar &regr`. Then `proc means` is invoked to calculate the sum of weights in the respective by-groups. The data are merged by `&svar &regr` and an observation now has `_MZ0` and `_MZ1`, the number of 'controls' and 'cases' per stratum and covariate pattern, where a covariate pattern is specified by the respective `&regr`-value. Stratum sums of `_MZ0` and `_MZ1` are calculated. This restructured phase two dataset has per stratum as many observations as there are covariate patterns in that stratum. Note that the outcome variable `&caco` has been eliminated.

Finally, the restructured phase one and phase two datasets are concatenated. Additional to the covariates specified in `&regr`, the following variables are stored:

Variable	Label
_NZ1	# Cases per Phase-1-Stratum
_NZ0	# Controls per Phase-1-Stratum
_N1	# Cases in Phase 1
_N0	# Controls in Phase 1
_NSTRAT	# strata
_MZ1	# Cases per Stratum and unique covar-pattern in Phase 2
_MZ0	# Controls per Stratum and unique covar-pattern in Phase 2
_M1	# Cases per Stratum in Phase 2
_M0	# Controls per Stratum in Phase 2
_S	Stratum-variable
_S1	Indicator Stratum 1
⋮	⋮
_SAMPLE	Phase (1/2: Phase 1/Phase 2)

Table 5: Design variables in the prepared dataset `prep`.

The variable names of Table 5 should not be used in the input datasets. The following variables are generated during the consistency check, their names should also not be present in the input datasets: `_za, _za_,_zb, _zb_,_zc, _zc_,_zd, _zd_`.

After preprocessing the macro performs the estimation methods specified in `&methods`. See Section 6 for details. For each method, one output dataset (exception: `em5macro`, see below) is generated and saved in the work directory. The name of the output data is a concatenation of the method name and the suffix "_est". The suffix can be extended by specifying the variable `&suffix` when calling `twophase`. For example, if `&methods=ml_em pl_bc`, the datasets `ml_em_est` and `pl_bc_est` will be saved. If additionally `&suffix=_a` is set, the datasets `ml_em_est_a` and `pl_bc_est_a` are generated.

The output data contains the parameter estimates, standard errors and covariance matrix for all parameters, including intercept (_ALPHA) and phase one parameters (provided the methods estimate these, see section 6 for details). Furthermore, the dataset contains the name of the macro that created the dataset (`_program`), the input data for the analysis (`_source`), the regression model (`_regress`), the time of the data creation (`_time`), the name of the estimation method (`_method`) and the time used for the processing of the macro (`_elaps`). An output dataset is printed in the example in section 7.1. Additional to this output dataset, intermediate results will be directed to the SAS output window if the relevant parameters, `&compare`, `&outproc`, `&outcorr` or `&outest` are not set to zero. This feature is demonstrated in the example in Section 7.3.

The `em5macro`, which provides ML estimates via the EM algorithm applied to a Poisson model, outputs three additional datasets to the workspace. These are `em_design&suffix`, containing the design matrix of the Poisson model, `em_theta&suffix`, containing the Poisson parameter estimate $\widehat{\boldsymbol{\theta}}$ (see below) and its covariance matrix and `em_final&suffix`, containing outcome - and stratum variable, regressor variables and expected counts under the Poisson model.

6 Parameter estimation

A detailed presentation of the estimation methods and their implementation may be found in the documentation `TPMethods.pdf` in this folder. We assume that in a population the probability of a binary outcome $D = 1$ in a person with covariate $X = x$ is given by the logistic model

$$\Pr(D = 1|X = x) = 1/[1 + \exp(-\alpha - x^T\beta)],$$

where x denotes a $p \times 1$ vector including exposures, covariates and interactions. Focus lies in the estimation of the log odds ratio parameter $\beta = (\beta_1, \ldots, \beta_p)^T$. Following Schill and Drescher (1997)and Scott and Wild (1997) , the Weighted Likelihood (WL), Maximum Likelihood (ML_EM, ML_SW) and Pseudo Likelihood (PL) methods as well as the analysis of phase 2 data alone ("complete-case analysis") estimate different sets of parameters:

Method	Estimated Parameter	Details
WL	$\theta_{\text{WL}} = (\alpha^*, \beta^T)^T$	
ML_SW	$\theta_{\text{ML\_SW}} = (\alpha^*, \beta^T)^T$	
ML_EM	$\theta_{\text{ML\_EM}} = (\alpha^*, \beta^T, \delta^T)^T$	δ parameterizes the (discrete) covariate distribution among 'non-diseased' and is potentially of high dimension. It is treated as a nuisance parameter and is not contained in the standard output.
PL	$\theta_{\text{PL}} = (\alpha^*, \beta^T, \gamma^T)^T$	γ describes the marginal phase one log odds with respect to stratum. The dimension of γ equals the number of strata.
S2	$\theta_{\text{S2}} = (\tilde{\alpha}, \tilde{\beta}^T)^T$	The complete case analysis usually only provides biased estimates.

Table 6: Estimation method and related parameter.

The interpretation of the intercept-term α^* depends on whether the phase one data are prospective or retrospective. If they are prospective, $\alpha^* = \alpha$, if they are retrospective, $\alpha^* = \alpha + \log(\Pr(D = 0)/\Pr(D = 1))$. In this case, an offset $\log(N_1/N_0)$ has been added to the linear predictor, where N_1 respective N_0 denote the number of cases and controls in phase one.

Compared to the presentation in Schill and Drescher (1997), we note two minor modifications in the parametrization. (1) In the pseudo likelihood methods, we use $\gamma = (\gamma_1, \ldots \gamma_J)^T$ instead of their $(\alpha_1, \xi_2, \ldots \xi_J)$. The ξ_j are obtained via $\xi_j = \gamma_j - \gamma_1$, $j = 2, \ldots, J$, and $\alpha_1 = \gamma_1$. (2) If the first phase sample is retrospective, their intercept estimates the term $\boldsymbol{\alpha + \log(N_1 \Pr(D=0)/N_0 \Pr(D=1))}$, different from ours.

7 Examples

7.1 Carroll data

In this example we use the `twophase`-macro in a simple setting. We show how a stratum variable is created and how an output data set looks like.
Carroll et al. (1993) published data of a measurement error scenario where for a large number of patients disease status $\boldsymbol{D}$ and a dichotomous, proxy measure of exposure $\boldsymbol{Z}$ is known, while the precise exposure $\boldsymbol{X}$ is measured only in a small validation sub-sample. Interest lies on the association between $\boldsymbol{D}$ and $\boldsymbol{X}$. The phase one data then consist of the disease variable `D`, the variable `Z` and observation counts `COUNT` for each combination of `D` and `Z`.

D	Z	COUNT
0	0	750
0	1	562
1	0	336
1	1	396

Table 7: Carroll phase one data.

The phase two data has an additional variable $\boldsymbol{X}$ and observation counts for all combinations of `D`×`Z`×`X`.

D	Z	X	COUNT
0	0	0	33
0	0	1	16
0	1	0	11
0	1	1	16
1	0	0	13
1	0	1	5
1	1	0	3
1	1	1	18

Table 8: Carroll phase two data.

In this form the data are not ready for the two-phase analysis, because a stratum variable `S` that numbers the strata is missing. In this example, `S` can be constructed from `Z` via `S=Z+1` to take the values **1** and **2**. The following SAS program does the preparation and then invokes the `twophase`-macro, where the `sas-twophase-package` is stored on drive

`G:`. We include only $\boldsymbol{X}$ in the regression model, perform the (default) PL_BC-method and print the result dataset.

```
*graphical options;
options nodate nonumber ls=64;

*This is the path, where the macros are stored;
%let path_tp=%str(G:\sas-twophase-package\macros);

*libname of input data;
libname in "G:\sas-twophase-package\data";

*include twophase-Macro;
%include "&path_tp.\twophase.sas";

data carr_ph1;
        set in.carrx1;
        S=Z+1;
run;

data carr_ph2;
        set in.carrx2;
        S=Z+1;
run;

%twophase(folder      =&path_tp,
          path_ph1    =carr_ph1,
          path_ph2    =carr_ph2,
          suffix      =_carr,
          compare     =0,
          caco        =D,
          svar        =S,
          counts_ph1  =COUNT,
          weights_ph2 =COUNT,
          regr        =X);

proc print data=pl_bc_est_carr;
run;
```

This produces the following SAS-output:

```
Obs     _parms      _estim      _stderr       _ALPHA          X

 1      _ALPHA    -0.30256     0.19631     0.038538    -0.07336
 2      X          0.56875     0.37337    -0.073355     0.13941
 3      _S1       -0.21943     0.04670     0.000000     0.00000
 4      _S2        0.23344     0.04665     0.000000     0.00000

Obs             _S1         _S2          _program

 1                0     .000000000    BC_MACRO.SAS
 2                0     .000000000
 3      .004309524     .000000000
 4                0     .004304612

Obs                      _source

 1      carr_ph1 (Phase 1), carr_ph2 (Phase 2)
 2
 3
 4

Obs     _regress                  _time

 1         X        2013.05.10 12:08:19.058000088
 2
 3
 4

Obs     _method               _elaps

 1       pl_bc     00:00:00.11 (hh:mm:ss.ff)
 2
 3
 4
```

7.2 Ohio data

In this example we use several estimation techniques and print the comparative table (setting `compare`=1). Prior to the two-phase analysis the datasets have to be transformed into the required shape.
The Ohio data are derived from Haneuse et al. (2011), who presented aggregated lung cancer mortality data for the 55-84 years old in the US state of Ohio in 1988. The phase one dataset has information on mortality status and the three ten-year age groups 55-64 (`AGE=0`), 65-74(`AGE=1`) and 75-84(`AGE=2`). The phase two dataset has, for sub-samples of size 100, additional information on race and sex. A race- and sex-adjusted analysis for the impact of age on lung cancer mortality is called for.

AGE	DEATH	NONDEATH
0	1785	1007046
1	2454	793901
2	1294	413697

Table 9: Ohio phase one data.

AGE	SEX	RACE	DEATH	NONDEATH
0	0	0	65	44
0	0	1	7	3
0	1	0	22	45
0	1	1	6	8
1	0	0	64	32
1	0	1	2	2
1	1	0	33	60
1	1	1	1	6
2	0	0	53	39
2	0	1	7	1
2	1	0	37	57
2	1	1	3	3

Table 10: Ohio phase two data.

In a first step the phase one and phase two datasets need to be transformed such that a separate variable indicating the disease status and variables with appropriate observation counts are present. Similar to the previous example, a stratum variable, in this case `S=AGE+1`, has to be constructed. Furthermore, dummy variables coding the factor `AGE` have to be defined for the phase two data.

```
*graphical options;
options nodate nonumber ls=78;
```

```
*This is the path, where the macros are stored;
%let path_tp=%str(G:\sas-twophase-package\macros);

*libname of input data;
libname in "G:\sas-twophase-package\data";

*include twophase-Macro;
%include "&path_tp.\twophase.sas";

data ohio_ph1;
        set in.ohio1;
        s=age+1;
run;

data ohio_ph2;
        set in.ohio2;
        s=age+1;
run;

*Cases and controls in different observations;
data ohio_ph1;
        length count 7.;
        set ohio_ph1(drop=nondeath rename=death=count in=a)
        ohio_ph1(drop=death rename=nondeath=count);
        if a then caco=1; else caco=0;
run;

data ohio_ph2;
        set ohio_ph2 (drop=nondeath rename=death=weights in=a)
           ohio_ph2 (drop=death rename=nondeath=weights);
        if a then caco=1; else caco=0;
        age1=(age=1);
        age2=(age=2);
run;

%twophase(folder      =&path_tp,
          path_ph1    =ohio_ph1,
          path_ph2    =ohio_ph2,
          methods     =wl pl_bc pl_sch ml_em ml_sw,
          compare     =1,
          outest      =1,
          caco        =caco,
          svar        =s,
```

```
          counts_ph1  =count,
          weights_ph2 =weights,
          regr        =age1 age2 race sex);
```

This code produces the following SAS-output. As described in Section 3.4 of `TPMethods.pdf`, the parameter estimates for the PL- and ML- methods coincide since the dummy variables for each stratum (`AGE1` and `AGE2`) are included in the regression model.

```
         Weighted Regression   PL (Breslow-Cain)     PL (Schill)
            estim      stderr     estim      stderr     estim     stderr

AGE1      0.70674     0.09095    0.68003    0.06801    0.68003    0.06801
AGE2      0.65480     0.08806    0.68810    0.06980    0.68810    0.06980
RACE      0.16738     0.33725    0.26567    0.31204    0.26567    0.31204
SEX      -1.13245     0.18025   -1.07442    0.17099   -1.07442    0.17099
_ALPHA    0.08609     0.08822    0.05538    0.08520    0.05538    0.08520
_S1          .           .      -0.34324    0.01949   -0.34324    0.01949
_S2          .           .       0.21288    0.01509    0.21288    0.01509
_S3          .           .       0.22472    0.02437    0.22472    0.02437

         ML (EM-Algorithm)     ML (Profile Lik.)
            estim      stderr     estim      stderr

AGE1      0.68003     0.06801    0.68003    0.06801
AGE2      0.68810     0.06980    0.68810    0.06980
RACE      0.26567     0.31204    0.26567    0.31204
SEX      -1.07442     0.17099   -1.07442    0.17099
_ALPHA    0.05538     0.08520    0.05538    0.08520
_S1          .           .          .          .
_S2          .           .          .          .
_S3          .           .          .          .
```

7.3 HdA data

This example presents an analysis with a continuous covariate in phase two.
The data are derived from the two-phase, age-matched case-control study of Pohlabeln et al. (2002) with focus on lung cancer risk due to intensity of occupational asbestos exposure. A precise intensity assessment of asbestos exposure in terms of asbestos fibreyears, however, was affordable only for a systematic, 20% sub-sample ($\boldsymbol{n_0 = n_1 = 164}$) of the original study. For the complete study ($\boldsymbol{N_0 = N_1 = 839}$) the duration of occupational asbestos exposure was known (see Pohlabeln et al. for details). We want to conduct a smoking-adjusted analysis for the association of fibreyears and lung cancer.
The phase one dataset then comprises information on disease status (variable `CASE`) and stratum counts. The stratum variable `STRATA` has 8 levels and was constructed as the

cross-classification of the variables ”duration of asbestos exposure” (4 levels) and ”heavy smoking” (2 levels). The phase two dataset has additional information on smoking history (variable `SMOKE` with 4 levels) and the continuous variable `FY`, log(asbestos fibreyears+1). `FY` is deemed an appropriate cumulative exposure measure in occupational epidemiology.

CASE	**STRATA**	**SMOKE**	**FY**	**COUNT**
0	1	0	0	17
0	1	1	0	36
0	1	1	0.7	1
0	1	1	1.5	1
0	2	0	0.1	2
0	2	0	2.1	1
0	2	1	0	4
0	2	1	0.2	1
0	2	1	0.3	1
⋮	⋮	⋮	⋮	⋮
1	8	3	7.6	2

Table 11: Excerpt of HdA phase two data.

We want to conduct a smoking-adjusted analysis for the association of fibreyears and lung cancer, so we included `FY` as well as dummy variables of `SMOKE` into the regression model. The results of all available estimation methods are concatenated in one comparatative table

```
options nodate nonumber ls=78;

*This is the path, where the macros are stored;
%let path_tp=%str(G:\sas-twophase-package\macros);

*libname of input data;
libname in "G:\sas-twophase-package\data";

*include twophase-Macro;
%include "&path_tp.\twophase.sas";

data hda1;
        set in.hdac_ph1;
run;

data hda2;
        set in.hdac_ph2;
        SMOKE1=(SMOKE=1);
```

```
        SMOKE2=(SMOKE=2);
        SMOKE3=(SMOKE=3);
run;

%twophase(folder      =&path_tp,
          path_ph1    =hda1,
          path_ph2    =hda2,
          methods     =ml_em ml_sw pl_bc pl_sch wl s2,
          suffix      =,
          compare     =1,
          outest      =1,
          caco        =case,
          svar        =strata,
          counts_ph1  =count,
          case        =1,
          control     =0,
          weights_ph2 =count,
          regr        =smoke1 smoke2 smoke3 fy,
          s1pr        =0);
```

The SAS-output:

```
         ML (EM-Algorithm)     ML (Profile Lik.)     PL (Breslow-Cain)
            estim      stderr     estim      stderr     estim      stderr

FY        0.16389    0.05739    0.16389    0.05739    0.13211    0.07039
SMOKE1    0.84504    0.54383    0.84504    0.54383    0.94050    0.54140
SMOKE2    1.93990    0.47946    1.93990    0.47946    1.98080    0.47861
SMOKE3    2.40276    0.50382    2.40276    0.50382    2.41971    0.50837
_ALPHA   -1.61585    0.45468   -1.61585    0.45468   -1.62627    0.45578
_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

         PL (Schill)           Weighted Regression   Sample2-Analysis
            estim      stderr     estim      stderr     estim      stderr

FY        0.12783    0.07378    0.11708    0.06081    0.14560    0.08239
SMOKE1    0.94200    0.54176    0.90879    0.53278    0.89373    0.53473
SMOKE2    1.98392    0.47849    1.97894    0.47023    2.01755    0.52330
SMOKE3    2.42252    0.50865    2.44983    0.49940    2.44167    0.55246
_ALPHA   -1.62716    0.45624   -1.62962    0.44913   -1.71063    0.47797
_S1      -0.93738    0.08456      .          .          .          .
_S2      -1.14631    0.22919      .          .          .          .
_S3      -0.33047    0.20098      .          .          .          .
_S4      -0.41292    0.20688      .          .          .          .
_S5       0.54064    0.06628      .          .          .          .
_S6       0.51708    0.15275      .          .          .          .
_S7       0.65840    0.14713      .          .          .          .
_S8       0.81449    0.14987      .          .          .          .
```

References

R.J. Carroll, M.H. Gail, and J.H. Lubin. Case-control studies with errors in covariates. *Journal of the American Statistical Association*, 88:185–199, 1993.

S. Haneuse, T. Saegusa, and T. Lumley. 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, 2011. URL `http://www.jstatsoft.org/v43/i11/`.

H. Pohlabeln, P. Wild, W. Schill, W. Ahrens, I. Jahn, U. Bolm-Audorff, and K.-H. Jöckel. Asbestos fibreyears and lung cancer: A two-phase case-control study with expert exposure assessment. *Occupational and Environmental Medicine*, 59:410–414, 2002.

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.

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