Documentation
Computing Adjusted Attributable Fractions from Two-Phase Case-Control Data
SAS®-Macro `af2p`

Author: **W. Schill, K. Drescher**
Date: **16.03.2014**
Version: **2.1**

Contents

1 General

Confounder-adjusted attributable fractions can be derived from case-control data (Bruzzi et *al.*, 1985), maximum likelihood estimates have been developed by Drescher and Osius (1993) and Greenland and Drescher (1993). This document describes the adaptation to two-phase case-control data and the implementation via the SAS®-macro `af2p`.
The macro, which requires SAS/IML software, is contained in the folder `sas-af2phase`. It can be downloaded from `www.tinyurl.com\schill-af2p`. The computations require results of a two-phase analysis, conducted with our `sas- twophase-package` (Schill *et al.*, 2014).

2 Attributable fractions

The confounder-adjusted attributable fraction is defined as the fraction of diseased that would have not occurred if the exposure of interest was absent:

$$
\begin{aligned}
\mathrm{AF} &= \frac{\mathbf{Pr}(\text{disease}) - \mathbf{Pr}(\text{disease} \mid \text{exposure of interest absent})}{\mathbf{Pr}(\text{disease})} \\
&= 1 - \frac{\mathbf{Pr}(\text{disease} \mid \text{exposure of interest absent})}{\mathbf{Pr}(\text{disease})}.
\end{aligned}
$$

More formally, let Z denote 'exposure of interest', U other exposure variables, D disease indicator,then

$$\mathrm{AF} = 1 - \frac{\int \mathbf{Pr}(D = 1 \mid Z = 0,\, u)\mathbf{Pr}(u)du}{\mathbf{Pr}(D = 1)}$$

and can be reformulated as

$$= 1 - \int R^{-1}(z|u)\mathbf{Pr}(z, u|D = 1)d(z, u),$$

where

$$R(z|u) = \frac{\mathbf{Pr}(D = 1|z, u)}{\mathbf{Pr}(D = 1|Z = 0, u)}$$

denotes the relative risk function.

We further work under the following assumptions:

- Logistic disease model: let the vector of explanatory variables X be partitioned into $X^T = (Z^T, U^T)$, $\beta = (\beta_Z^T, \beta_U^T)^T$. Then

$$\mathbf{Pr}(D = 1|x) = \left[1 + e^{-(\alpha + x^T\beta)}\right]^{-1} = \left[1 + e^{-(\alpha + z^T\beta_Z + u^T\beta_U)}\right]^{-1}. \tag{1}$$

- Rare disease assumption, which implies that the relative risk equals the odds ratio, defined as

$$\mathrm{OR}(z|u) = \frac{\mathbf{Pr}(D = 1|z, u)\mathbf{Pr}(D = 0|Z = 0, u)}{\mathbf{Pr}(D = 0|z, u)\mathbf{Pr}(D = 1|Z = 0, u)},$$

  which, under the logistic model, is given as

$$e^{z^T\beta_Z}.$$

Combining yields

$$\mathrm{AF} = 1 - \int e^{-z^T\beta_Z}\mathbf{Pr}(z, u|D = 1)d(z, u). \tag{2}$$

3 Estimation in two-phase case-control studies

Two-phase setup Phase one case-control data are stratified into $J \geq 1$ strata, with N_{0j} and N_{1j} denoting the number of controls and cases in stratum j, $j = 1, \dots J$. Let N_1 and N_0 denote the number of cases and controls in phase one. From within each cell of the phase one data, $1 \leq n_{0j} \leq N_{0j}$ controls and $1 \leq n_{1j} \leq N_{1j}$ cases are recruited for covariate ascertainment in phase two. We assume discrete covariates. Within each stratum j, $j = 1, \ldots, J$, $K_j \geq 1$ covariate patterns x_{jk} are observed in phase two, with n_{0jk}, n_{1jk} denoting the counts among controls and cases, respectively.

Since we work with discrete covariates, the logistic model parameters can also be obtained by using a Poisson model where the (possibly unobserved) phase one counts N_{ijk} are Poisson distributed with expectation

$$\mu_{ijk} = \begin{cases} e^{\delta_{jk}+\gamma+x_{jk}^T\beta} & \text{if } i = 1 \text{ (cases),} \\ e^{\delta_{jk}} & \text{if } i = 0 \text{ (controls).} \end{cases} \tag{3}$$

The parameter to be estimated is $\theta = (\gamma, \beta^T, \delta^T)^T$, where $\delta^T = (\delta_1, \ldots \delta_{JK_J})$ explicitly parameterizes the covariate distribution among controls. As in ordinary case-control studies, the intercept parameter in the logistic model has been changed.

In the discrete model considered here, the attributable fraction (Equation 2) is given as

$$\begin{aligned} 1 - \mathbf{AF} &= \sum_{jk} e^{-z_{jk}^T\beta_Z} \Pr(x_{jk}|D = 1) \\ &= \sum_{jk} e^{-z_{jk}^T\beta_Z} \frac{\mu_{1jk}}{\mu_{1++}} = \frac{1}{\mu_{1++}} \sum_{jk} e^{\delta_{jk}+\gamma+u_{jk}^T\beta_U}, \end{aligned} \tag{4}$$

where '++' denote summation over j and k. Note that the sum in the last equation represents the expected number of cases that would remain if 'risk factor' Z was absent.

Evaluation To evaluate a logistic two-phase study, we use the EM algorithm applied to the Poisson model (Schill and Drescher, 1997). The algorithm proceeds with E-step and M-step:

- E-step: Compute pseudo-complete phase one counts via

$$\widetilde{N}_{ijk}^{(t+1)} = n_{ijk} + (N_{ij} - n_{ij}) \frac{\widehat{\mu}_{ijk}^{(t)}}{\widehat{\mu}_{ij+}^{(t)}}$$

- M-step: Perform a Poisson regression with the phase one pseudo counts.

After completion, the EM algorithm has produced ML estimates $\widehat{\theta}$, ML estimates of expected phase one counts $\widehat{\mu}_{ijk}$, along with a parameter covariance matrix $\widehat{\Sigma}_\theta$

$$\begin{aligned} \widehat{\mu}_{ijk} &= \left(e^{X\widehat{\theta}}\right)_{ijk} \\ \widehat{\Sigma}_\theta &= \left[X^T(\mathbf{Diag}(\widehat{\mu}) - \mathbf{BlockDiag}(D_{ij})X\right]^{-1}. \end{aligned}$$

where X denotes the design matrix of the Poisson model.

An ML estimate of $\mathbf{AF}$ is obtained from Equation 4 by plugging in estimated values:

$$\widehat{\mathbf{AF}} = 1 - \frac{1}{\widehat{\mu}_{1++}} \sum_{jk} e^{\widehat{\delta}_{jk}+\widehat{\gamma}+u_{jk}^T\widehat{\beta}_U}. \tag{5}$$

Note that $\widehat{\mu}_{1++} = N_1$, the number of cases in phase one.

Asymptotic variance of $\widehat{\mathbf{AF}}$ A benefit of using the fully parameterized Poisson model is that the asymptotic variance of $\widehat{\mathbf{AF}}$ is obtained nearly as by-product of the algorithm. However, the result is firstly valid only under Poisson sampling, not under case-control sampling, but see below.

We obtain the asymptotic variance of $\widehat{\mathbf{AF}}$ by applying the delta-method:

$$\mathbb{V}(\widehat{\mathbf{AF}}) = \frac{\partial \widehat{\mathbf{AF}}}{\partial \boldsymbol{\theta}^T} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} \frac{\partial \widehat{\mathbf{AF}}}{\partial \boldsymbol{\theta}}. \tag{6}$$

We write $\widehat{\mathbf{AF}} = 1 - \widehat{\mu}^*_{1++}/\widehat{\mu}_{1++}$, where $\widehat{\mu}_{1++} = \sum \widehat{\mu}_{1jk}$ and $\widehat{\mu}^*_{1++} = \sum \widehat{\mu}^*_{1jk}$, with $\widehat{\mu}^*_{1jk} = e^{\widehat{\delta}_{jk} + \widehat{\gamma} + u^T_{jk} \widehat{\beta}_U}$. We represent the vectors $\widehat{\mu}_1$ and $\widehat{\mu}^*_1$ as $\widehat{\mu}_1 = \exp(X_1 \widehat{\theta})$ and $\widehat{\mu}^*_1 = \exp(X^*_1 \widehat{\theta})$, where X_1 is the submatrix of the design matrix pertaining to the cases and X^*_1 is the modification of X_1 where all columns pertaining to Z are set to zero. One computes

$$\frac{\partial \widehat{\mathbf{AF}}}{\partial \theta^T} = -\frac{\partial}{\partial \theta^T} \left(\frac{\widehat{\mu}^*_{1++}}{\widehat{\mu}_{1++}} \right) = -\frac{\frac{\partial}{\partial \theta^T} \widehat{\mu}^*_{1++}}{\widehat{\mu}_{1++}} + (1 - \widehat{\mathbf{AF}}) \frac{\frac{\partial}{\partial \theta^T} \widehat{\mu}_{1++}}{\widehat{\mu}_{1++}}.$$

$\frac{\partial}{\partial \theta^T} \widehat{\mu}^*_{1++}$ and $\frac{\partial}{\partial \theta^T} \widehat{\mu}_{1++}$ are easily obtained when the algorithm has finished: They are given as $\sum_{jk} \widetilde{x}^{*T}_{jk} \widehat{\mu}^*_{1jk}$ and $\sum_{jk} \widetilde{x}^T_{jk} \widehat{\mu}_{1jk}$, respectively. Here, $\widetilde{x}^*_{jk}$ and $\widetilde{x}_{jk}$ are the 'complete' regressors, i. e., rows of the design matrix (including indicators for γ and δ_{jk}) such that $\widehat{\mu}^*_{1jk} = \exp(\widetilde{x}^{*T}_{jk} \widehat{\theta})$ and $\widehat{\mu}_{1jk} = \exp(\widetilde{x}^T_{jk} \widehat{\theta})$.

Variance under case-control sampling Equation 6 firstly is valid under Poisson sampling. To show that it is also valid under retrospective sampling, i. e., sampling conditional on N_0 and N_1, we use a reparametrization. Define the parameter $\xi = (\xi_1^T, \xi_2^T)$, where $\xi_1 = (\beta^T, \mathrm{Pr}(x_{jk}|D = 1)^T)^T$ and $\xi_2^T = (\mu_{0++}, \mu_{1++})$. Then (N_0, N_1) are ancillary for ξ_1. Since $\mathbf{AF}$ depends on ξ_1 alone, it follows that $\widehat{\mathbf{AF}}$ and (N_0, N_1) are asymptotically independent, hence $\mathbb{V}(\widehat{\mathbf{AF}}|N_0, N_1) = \mathbb{V}(\widehat{\mathbf{AF}})$.

Standard case-control studies The discrete covariate approach taken here can also be applied to standard case-control studies. To use our package in this case, an artificial phase one dataset has to be created with one stratum and stratum sizes n_0 and n_1. The E-step of the EM algorithm is empty and a Poisson model is fit to the data.

4 Implementation in macro `af2p`

The SAS-macro `af2p` implements the above calculations. For any term or combination of terms in the linear predictor (except the intercept of course) attributable fractions can be computed. The regressors pertaining to Z are required as a list of terms. To set up β_Z and X^*_1 properly, the exact sequence of terms in the linear predictor must be provided as another argument.
The program requires output of a two-phase analysis that was performed by running the EM algorithm, as called by the macro `twophase` of `sas-twophase-package` (Schill *et al.*, 2014).

4.1 Parameters of macro `af2p`

```
%macro af2p(
       pathdesign=em_design,
       pathparm  =em_theta,
       pathfinal =em_final,
       pathout   =af,
       regress   =,
       factor    =,
       nameF     =,
       study     =,
       s1pr      =0);
```

Arguments The following arguments can be used.

Parameter	**Description**
`pathdesign`	Path of the dataset that contains design matrix of underlying two-phase analysis via the EM algorithm. The value is of the form `pathdesign =` *libname.filename*. If the dataset is located in the workspace, *libname* can be omitted. **Value**: string **Default**: `em_design`
`pathparm`	Path of the dataset that contains parameter estimate and covariance matrix of the underlying analysis. See `pathdesign` for details. **Value**: string **Default**: `em_theta`
`pathfinal`	Path of the dataset that contains estimated expected counts of the underlying analysis. See `pathdesign` for details. **Value**: string **Default**: `em_final`
`pathout`	Path of the dataset that contains estimated attributable fraction and standard error. See `pathdesign` for details. **Value**: string **Default**: `af`
`regress`	Exact form of the linear predictor that was used in the underlying analysis by running macro `twophase`. **Value**: string **Default**: none
`factor`	All terms that pertain to Z, the factor whose attributable risk is to be estimated. Each single term in `factor` must be an element of the terms in `regress`. **Value**: string **Default**: none
`nameF`	Description of factor. **Value**: string **Default**: none
`study`	Description of underlying analysis or study. **Value**: string **Default**: none
`s1pr`	Indicator whether underlying phase one study is retrospective (`s1pr=0`) or prospective (`s1pr=1`). Must agree with the value of the according parameter when calling macro `twophase`. **Value**: numric **Default**: 0

4.2 Usage

The macro `af2p` is part of the folder `sas-af2phase` and can be downloaded from `www.tinyurl.com\schill-af2p`. The folder must be stored somewhere on your computer. The folder also contains this documentation and an example SAS program, see below. The macro requires certain outputs from macro `twophase`, which is part of the `sas-twophase-package` (url: `tinyurl.com/schill-twophase`).

5 Examples

5.1 HdA study

For a more detailed description of the ”HdA” study, refer to Schill *et al.* (2014) and the documentation of `sas-twophase-package`. In brief, the data are concerned with a two-phase case-control study on the lung cancer risk of occupational asbestos exposure. The phase one data constitute the cross-tabulation of case status and an eight-level stratum variable `STRATA`, which cross-classifies coarse smoking information (2 levels) and a four-level variable denoting duration of occupational asbestos exposure. The second phase dataset includes more precise exposure information: A four-level variable on smoking status (`SMOKE`) and a continuous variable `FY`, indicating log(fibreyears+1), an intensity measure of asbestos exposure.

Our first goal is to determine a smoking-adjusted relative risk estimate for log(fibreyears+1), where we choose as linear predictor for the two-phase analysis `'smoke1 smoke2 smoke3 fy'` (the first three terms denote dummy variables for mild-, mid- and heavy smoking). Based on this analysis, the second goal is to estimate adjusted attributable fractions for several risk factors: (1) smoking alone, (2) fibreyears alone and (3) both factors combined.

We assume that the folders `sas-twophase-package` and `attrib-frac-twophase` are stored on drive `e:` of the computer. A SAS program then `%incudes{.}` the necessary macros, generates the dummy variables and performs the two-phase analysis by choosing the EM algorithm.

```
*This is the path, where the macros of the twophase package are stored;
%let path_tp=%str(e:\sas-twophase-package\macros);
%include "&path_tp.\twophase.sas";

*This is the path, where macro af2p is stored;
%let path_af=%str(e:\sas-af2phase\macros);
%include "&path_af.\af2p.sas";

libname in  "e:\sas-twophase-package\data";

data hda1;
        set in.hdac_ph1;
run;

data hda2;
        set in.hdac_ph2;
        SMOKE1=(SMOKE=1);
        SMOKE2=(SMOKE=2);
        SMOKE3=(SMOKE=3);
run;

title'HdA Data, Stratif. A (8 Strata)';

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

The screen output of this part of the program is

```
         ML (EM-Algorithm)
            estim     stderr

FY        0.16389    0.05739
SMOKE1    0.84504    0.54383
SMOKE2    1.93990    0.47946
SMOKE3    2.40276    0.50382
_ALPHA   -1.61585    0.45730
```

The second part of the program computes the attributable fractions by running `af2p` three times:

```
%af2p(
       pathdesign=em_design,
       pathparm  =em_theta,
       pathfinal =em_final,
       pathout   =af1,
       regress   =smoke1 smoke2 smoke3 fy, /* original regressor */
       factor    =smoke1 smoke2 smoke3       ,
       nameF     =Smoking,
       study     =HdA,
       s1pr      =0);
%af2p(
       pathout   =af2,
       regress   =smoke1 smoke2 smoke3 fy,
       factor    =fy,
       nameF     =Asbestos exposure,
       study     =HdA);
%af2p(
       pathout   =af3,
       regress   =smoke1 smoke2 smoke3 fy,
       factor    =fy  smoke1 smoke2 smoke3,
       nameF=Asbestos + Smoking,
       study     =HdA);
```

The combined result, which also displays standard errors and 95% confidence intervals, would look like

```
                      Attributable      Std.      Lower      Upper
NameF                   Fraction       Error      95%-CL     95%-CL    Terms Removed

Smoking                  0.77974      0.10117    0.58143    0.97804    smoke1 smoke2 smoke3
Asbestos exposure        0.09784      0.03066    0.03773    0.15794    fy
Asbestos + Smoking       0.80128      0.09036    0.62418    0.97837    fy smoke1 smoke2 smoke3
```

5.2 Ille-et-Vilaine study

In this example we show how attributable fractions can be estimated from standard case-control data. The data are taken from the ”Ille-et-Vilaine” study, which is described in detail in the book of Breslow and Day (1980). This study is a population based case-control study on oesophageal cancer with focus on alcohol consumption and smoking habits as risk factors.

Records of 200 cases and 775 controls are included, covariates considered are the categorical variables age (variable `AGE` in 6 levels), alcohol consumption in gram per day (variable `ALC` in 4 levels) and tobacco consumption in gram per day (variable `TOB` in 4 levels). A categorical main effect model is used to derive the age-adjusted log odds ratios for the distinct levels of alcohol - and tobacco consumption. Finally, we estimate attributable fractions to alcohol consumption, tobacco use and both exposures combined.

First, the places where programs and data are stored have to be made available.

```
*This is the path, where the two-phase macros are stored;
%let path_tp=%str(d:\ESave\sas-twophase-package\macros);
*include twophase-Macro;
%include "&path_tp.\twophase.sas";
```

```
*This is the path, where the attributable risk macro is stored;
%let path_af=%str(d:\ESave\sas-af2phase\macros);
*include attributable fraction-Macro;
%include "&path_af.\af2p.sas";

*libname of input data;
libname in "d:\ESave\sas-af2phase\data";
```

Second, an artificial stratum variable `sdum` is created and an artificial phase one dataset with two observations is generated. The data are analyzed as two-phase data.

```
     /* Define artificial stratum variable sdum */
     data illev2;set in.ille_et_vilaine;
          sdum=1;
     proc sort;by sdum;
     run;
     /* Produce artificial phase one dataset */
     proc freq data=illev2 noprint;
          tables case/nocol norow nopercent out=illev1;
          weight n;
          by sdum;
     run;
          %twophase(folder      =&path_tp,
               path_ph1    =illev1,
               path_ph2    =illev2,
               methods     =ml_em wl s2,
               compare     =1,
               outest      =1,
               caco        =case,
               svar        =sdum,
               counts_ph1  =count,
               weights_ph2 =n,
               regr        =age2 age3 age4 age5 age6 alc2 alc3 alc4 tob2 tob3 tob4,
               s1pr        =0);
```

The results are as follows:

```
         ML (EM-Algorithm)     Weighted Regression   Sample2-Analysis
            estim     stderr      estim     stderr      estim     stderr

AGE2      1.98088    1.10407    1.98088    1.10407    1.98088    1.10407
AGE3      3.77629    1.06804    3.77629    1.06804    3.77629    1.06804
AGE4      4.33518    1.06505    4.33518    1.06505    4.33518    1.06505
AGE5      4.89641    1.07638    4.89641    1.07638    4.89641    1.07638
AGE6      4.82654    1.12130    4.82654    1.12130    4.82654    1.12130
ALC2      1.43463    0.25006    1.43463    0.25006    1.43463    0.25006
ALC3      1.98072    0.28476    1.98072    0.28476    1.98072    0.28476
ALC4      3.60287    0.38504    3.60287    0.38504    3.60287    0.38504
TOB2      0.43805    0.22832    0.43805    0.22832    0.43805    0.22832
TOB3      0.51262    0.27298    0.51262    0.27298    0.51262    0.27298
TOB4      1.64100    0.34411    1.64100    0.34411    1.64100    0.34411
_ALPHA   -5.54087    1.08304   -5.41304    1.08273   -5.54087    1.08304
```

Finally, we compute the attributable fractions and present the combined results.

```
   %af2p(pathout=af1,
           regress=age2 age3 age4 age5 age6 alc2 alc3 alc4 tob2 tob3 tob4,
            factor=alc2 alc3 alc4,
            NameF =Alcohol,
            Study =Ille et Vilaine,
           s1pr   =0);
     %af2p(pathout=af2,
           regress=age2 age3 age4 age5 age6 alc2 alc3 alc4 tob2 tob3 tob4,
            factor=tob2 tob3 tob4,
            NameF =Tobacco,
            Study =Ille et Vilaine,
           s1pr   =0);
     %af2p(pathout=af3,
           regress=age2 age3 age4 age5 age6 alc2 alc3 alc4 tob2 tob3 tob4,
            factor=tob2 tob3 tob4 alc2 alc3 alc4,
            NameF =Tobacco+Alcohol,
            Study =Ille et Vilaine,
           s1pr   =0);
     data af;set af1 af2 af3;
     run;
     title 'Attributable fractions in the Ille et Vilaine study';
     proc print data=af label noobs;
          var NameF af Cl_af cu_af;
     run;
```

```
Attributable fractions in the Ille et Vilaine study

                   Attributable     Lower      Upper
NameF                Fraction       95%-CL     95%-CL

Alcohol               0.72436      0.62714    0.82158
Tobacco               0.29401      0.14909    0.43892
Tobacco+Alcohol       0.80001      0.71913    0.88089
```

Acknowledgement

We thank Dirk Enders for review of this document and review of the software.

Technischer Anhang

Wenn die hase 1 Studie retrospektiv ist, wird *nach* Beendigung des EM Algorithmus der Intercept modifiziet, indem $\log(N_1/N_0)$ von$\theta(1)$ abgezogen wird. Für das so modifizierte $\widehat{\theta}$ gilt *nicht* $\exp(X_1\widehat{\theta}) = \widehat{\mu}_1$. Deswegen muss diese "Korrektur" rückgängig gemacht werden. Um das zu tun, müssen die originalen $\widehat{\mu}_i$ auch eingelesen werden. Die Varianzkorrektur bleibt dagegen bestehen (ist allerddings irrelevant).

References

Breslow, N.E. and Day, N.E. (1980): Statistical Methods in Cancer Research Volume 1 - The analysis of case-control studies. IARC Scientific Publications No 32.

Bruzzi, P., Green, S.B., Byar, D.P., Brinton, L.A. and Schairen, C. (1985): Estimating the Population Attributable Risk for Multiple Risk Factors from Using Case-Control Data. American Journal of Epidemiology 122, 904-914.

Drescher, K. and Osius, G. (1993): . Attributable Risk Estimation from Case-Control Data with Regression Adjustment for Confounders. Institute of Statistics, University Bremen. Unpublished Manuscript..

Greenland, S. and Drescher, K.(1993): Maximum Likelihood Estimation of the Attributable Fraction from Logistic Models. Biometrics 49, 865-872.

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

Schill, W., Enders, D. and Drescher, K. (2014): A SAS Package for Logistic Two-Phase Studies. Journal of Statistical Software (submitted).