The SAS PAR Macro

Ellen Hertzmark, Handan Wand, and Donna Spiegelman

March 8, 2012

Abstract

"What % of the cases would be prevented if it were possible to eliminate one or more risk factors from a target population?"

The %PAR SAS macro is designed to answer questions such as this by estimating the population attributable risk (PAR) and its 95% confidence interval. We calculate the full PAR and partial PAR, as defined below. The variance formulas implemented here apply only to cohort studies. Currently, the confidence intervals are not valid for case-control studies. Please write to us if you have a case-control study. Models with interaction terms are acceptable.
Population prevalences can be considered fixed (e.g. for sensitivity analysis), estimated from the same cohort from which the relative risks were estimated, or estimated from a population survey such as NHANES.

• **FULL PAR**: all measured risk factors are considered eliminated. All members of the target population who are exposed switch to the lowest risk category of all measured risk factors.

• **PARTIAL PAR**: One or more risk factors are considered eliminated, while others are allowed to remain unchanged. References: (Bruzzi *et al.*(1985), Spiegelman, Hertzmark, and Wand (2006)).

Keywords: SAS, macro, population attributable risk, attributable risk, pooled logistic regression, Cox models, proportional hazards regression

Contents

1 Description **2**

2 Invocation and Details **3**

3 Examples **6**

3.1 Example 1. Correct runs of the macro 7

3.1.1 Using the results of a logistic model 7

3.1.2 Using the results of a proportional hazards (Cox) model 8

3.2 Example 2. Errors in macro call 9

3.2.1 A variable left out of the lists 9

3.2.2 A variable appears in both *MODVAR* and *FIXEDVAR* lists 10

3.3 No *SAMPSZ* given for *VPREVOPT*=COHORT 11

3.3.1 No *MODVAR* . . . 12
3.3.2 RR is less than 1 for some combination of the predictors . . . 12
3.4 Example 3. Including interaction terms AND Removing only some of the BMI-related risk . . . 13
3.5 Example 4. Using prevalences from a complex survey design . . . 16

4 Frequently Asked Questions **21**
4.1 Q: Reference level is not lowest RR or OR . . . 21
4.2 Q: How can I use %PAR with interaction variables? . . . 21
4.3 Q: What if I have negative interactions? . . . 22
4.4 Q: SAS says I ran out of memory . . . 22
4.5 Q: The program has been running for a long time and is still not done. . . . 22
4.6 Q: The coefficient for my exposure is significant, but the confidence interval for PAR includes 0 . . . 23
4.7 Q: I ran my model with output from a PROC PHREG (a proportional hazards or Cox model) and got WARNINGS . . . 23

5 Useful Knowledge **23**
5.1 PAR is distributive in univariate models . . . 23
5.2 Choice of *VPREVOPT*: . . . 23

6 Computational Methods **24**
6.1 Variance-covariance matrix of the prevalences in a cohort study . . . 24
6.2 Comparison of *VPREVOPT*=COHORT and *VPREVOPT*=FIXED for various values of *SAMPSZ* in an example . . . 24
6.3 Computation of confidence intervals . . . 24

7 Credits **24**

8 References **25**

1 Description

Given the regression coefficients and their variance-covariance matrix, as well as the prevalence of each unique combination of the risk factors in the target population, the %PAR macro will compute the full or partial population attributable risk for a set of exposures. The macro can accommodate prevalences that come from cohort studies or from complex study designs, as well as prevalences

that are considered fixed.

2 Invocation and Details

As discussed in Spiegelman, Hertzmark, and Wand (2007), with references given to the original papers, in a multifactorial disease setting, the full or partial PAR is strictly valid only when all confounders are explicitly modeled. Since age is typically one of the strongest confounders, it is recommended that pooled logistic regression be used, rather than the Cox model, when PARs are to be calculated.

Before running %PAR, you need to make the necessary datasets. namely the dataset containing the log relative risk coefficients and their variance-covariance matrix and the dataset containing the prevalences. If you are using *VPREVOPT*=SURVEY, you also need to have the variance-covariance matrix of the prevalences.
The coefficients and their variance-covariance matrix can be obtained using `covout outest=`*data_set_name* in the procedure you are using to compute the relative risks, for example

```
proc logistic descending data=DATA covout outest=betadat;
  model CASE = VAR1 VAR2 .... VARn;
run;
```

If you are computing the PAR with frequencies from a cohort study, the frequency/prevalence dataset for use with the results of a pooled logistic model should be obtained as follows:

```
proc sort data=DATA;  by VAR1 VAR2 .... VARn;  run;
proc means noprint data=DATA;  var ID;
output  out=freqs  n=fq;
by VAR1 VAR2 .... VARn;
run;
```

DATA is the data set on which the analysis was done.
VAR1 VAR2 .... VARn is a complete list of the model variables in the analysis you are using to compute the PAR.
ID is any numeric variable that is non-missing for all observations used in the analysis.

The resulting data set will have one observation for each unique combination of the model variables, giving the values of the variables and the number of observations in that stratum (i.e. the stratum frequency). The name of the output dataset (`freqs` above) and that of the 'n' variable (`fq` above) can be any valid SAS names the user wants.
Unlike previous versions of %PAR, the prevalence data set does not need to be in any specific order (i.e. you don't have to worry about how you sort the data). Furthermore, unlike previous versions of %PAR, you do not have to 'fill in' the prevalence data set with the prevalences (0) of the combinations that do not occur in the data set.

NOTE: We used PROC MEANS rather than PROC FREQ because PROC FREQ 'freaks out' (i.e. complains about being out of memory) if the number of variables or categories is large).

Making the prevalence data set and the variance-covariance matrix of the prevalences for complex survey data will be described in detail in Example 4.

In order to run this macro, your program must know where find it. You can tell SAS where to find macros by using the options

```
mautosource sasautos= <directories where macros are located>.
```

For example, at the Channing Lab, an options statement might be

```
options nocenter ps=78 ls=125 replace formdlim='['
mautosource
sasautos=('/usr/local/channing/sasautos',
   '/proj/nhsass/nhsas00/nhstools/sasautos);
```

This will allow you to use all the SAS read macros for the data sets (`/proj/nhsass/nhsas00/nhstools/sasautos`), as well as other public SAS macros, such as %PM, %INDIC3, %EXCLUDE, %LOGITR, and %MPHREG9.

Note: This macro can use a lot of memory if you have a typical Channing model with a large number of variables (and thus a large number of unique combinations in the prevalence dataset). Even a small prevalence dataset, like that used in Example 4 below required over 11 megabytes of memory. To get 2 gigabytes of memory, use the shell command

```
qbs -q l -o memsize=2048M <program name>
```

This shell command waits till there is a server with the required amount of memory, then 'hogs' the machine till it's finished running. It is to everyone's advantage that you request the smallest amount of memory the job will run with, even though you can go all the way to 6144M.

The macro call is

```
%par(
  bdata= The name of the dataset containing the coefficients and their
         variance-covariance matrix.
         This may be a 2-level SAS name, e.g. LIBNAME.DATANAME,
         or a 1-level name (i.e. the dataset has already been read in during
         the program run, so WORK is understood as the libname).
  pdata= The name of the dataset containing the variable combinations
         and their frequencies or prevalences.
         Again, this may be a 2-level or 1-level name.
  vpdata= The name of the dataset containing the variance-
          covariance matrix of the prevalences.
          Use this only if you are using prevalences from a
          complex survey and VPREVOPT=SURVEY (See below).
          The rows (and columns) in VPDATA should be in the same order as
          the rows in PDATA.
  vpvarname= The 'prefix' of the variables in the dataset containing
             variance-covariance matrix of the prevalences, when
             VPREVOPT=SURVEY.
             The variables in the dataset should be named
```

```
          <vpvarname>1 <vpvarname>2 ..... <vpvarname><N obs in PDATA>.
n_or_p= Whether the PDATA dataset contains counts (N) or
        prevalences (P).
        In the code above, we produced counts, so
        n_or_p=N.
        Default=P.
n_or_pname= The name of the variable in PDATA that is the count
            or the prevalence for the stratum.
            The macro will use this information to make a dataset with
            the prevalences it needs.
            In the example above, n_or_pname=FQ,
            because the code in the PROC MEANS that made
            the dataset was 'n=fq'.
            Default=prev.
modvar= The list of modifiable variables.
        This list must be written out completely.
        Ordinary SAS notation like 'bmiq2-bmiq5' is not acceptable, but
        the results of %INDIC3 (e.g. '&bmiq_') are, if you are still in
        the program that ran %indic3.
        NOTE to those who like neatly formatted programs:  The macro
        counts the variables in this list by looking for spaces.
        If you want to indent a line, use spaces, not tabs.
fixedvar= The list of variables to be held fixed in computing a partial
          PAR.  This list should usually include age, gender, family
          history, and other similarly non-modifiable variables.
          If this list is empty, the full PAR will be computed.
          NOTE:  The lists MODVAR and FIXEDVAR should be disjoint.
          If there is any overlap, the macro will stop
          (See Example 3.2.2).
          NOTE to those who like neatly formatted programs:  The macro
          counts the variables in this list by looking for spaces.
          If you want to indent a line, use spaces, not tabs.
fullpar= T or F.
         Whether you want to compute the full PAR (i.e. the fraction of
         cases accounted for by your model).
         Default is F.
partialpar= T or F.
            Whether you want to compute the partial PAR (i.e. the fraction
            of cases accounted for by the modifiable variables).
            NOTE:  if MODVAR is empty, the macro will automatically set
            PARTIALPAR to F and FULLPAR to T.
            Default is T.
vprevopt= COHORT or FIXED or SURVEY.
          Defines the way you want to handle the variance-covariance
          of the prevalences.
          For COHORT and FIXED, the macro will do the
          necessary calculations.
          If the sample size is very large or if you only want to
          generalize to the exact group on which you did the analysis,
          you can use FIXED.
```

```
            See Section 5.2 for a discussion of when you
            can use vprevopt=FIXED in cohort studies.
            Default=FIXED.
  sampsz= The size of the sample on which the analysis is based.
          This is necessary only for VPREVOPT=COHORT.
          If VPREVOPT is COHORT and this parameter is left blank,
          the macro gives an 'ERROR in macro call' message and stops.
          See the example in section 3.2.5.
);
```

Because of the nature of the formulas used, the macro requires that all strata (defined as unique combinations of the fixed and modifiable variables) have relative risk at least 1 (equivalent to the sum of the estimated coefficients being at least 0), for all variables and for the fixed variables (i.e., the relative risk for each stratum must be at least 1, even if some of the individual coefficients are negative). If this is not so, the macro will tell you which strata are problematic. The easiest way to avoid problems is to use the lowest risk category as the reference level for each variable.

3 Examples

In examples 1-3, we use data from a study of the relation of overweight and obesity to total death in NHS. The analysis begins with 1980 data and ends in June, 2000.

The variable names are

```
prevcanc   reported cancer before baseline (1980)
prevchd    reported CHD before baseline (1980)
db30       diabetes onset before age 30
prevhrt    reported heart disease before baseline (1980)
age6069    age 60-69 (updated)
agege70    age 70 plus (updated)
bmilt185   BMI under 18.5 at baseline (1980)
ovwt       BMI 25-29.99 at baseline (1980)
obese      BMI 30 plus at baseline (1980)
smkmod     moderate level of smoking at baseline (1980)
smkhi      high level of smoking at baseline (1980)
pacumneg   no physical activity at baseline (1980)
nineties   time period is 1990s or later
```

The reference categories are

```
no report of cancer, CHD, or heart disease before the time period and no
   diabetes onset before age 30
age under 60
BMI 18.5 to 24.99
no smoking
any physical activity
time period earlier than 1990s
```

Note that we are using 'no physical activity' as our model variable. This is because the %PAR macro needs the relative risks in all strata to be at least 1, and physical activity is protective.

In a previous program we made the datasets with the coefficients and their variance-covariance matrix (*BDATA*) and the frequencies (*PDATA*) and stored them as permanent datasets using the code shown at the beginning of the 'Invocation and Details' section. The macro can accept a 2-level SAS dataset name (to read directly from the permanent dataset) or a 1-level name (if the dataset has already been read into the WORK (temporary) library).
NOTE: Since we ran %PAR in a different program from the one that made the data, we cannot use &bmiq_ notation.

Our interest is in the fraction of total deaths that could be attributed to overweight (`OVWT`) and obesity (`OBESE`). Since exercise is highly related to these variables, we will examine what would happen if everyone were normal weight or thinner and everyone did some exercise (i.e. `PACUMNEG=0`). Note that we are not examining the effect of extreme thinness, but are considering it unmodifiable. In other words, we are considering the effect of eliminating only some of the BMI-related exposures.

3.1 Example 1. Correct runs of the macro

3.1.1 Using the results of a logistic model

The macro call is

```
title3 'results from logistic regression';
title4 'correct run without interactions';
%par(bdata=bplain1, pdata=pplain1, modvar=ovwt obese pacumneg,
n_or_p=n, n_or_pname=fq,
fullpar=t,
fixedvar=bmilt185 smkmod smkhi nineties age6069 agege70
 prevcanc prevchd db30 prevhrt);
```

The macro does not require the variables to be in the same order as the model. Every variable in the model must be listed in either *MODVAR* or *FIXEDVAR* (and not both) with the same name as was used in the model. See Examples 3.2.1 and 3.2.2. We did not have to specify *VPREVOPT*=FIXED, since that is the default.
The results are

```
------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program parrun   24MAY2011   15:46    stleh            1

results from logistic regression
correct run without interactions

option for the variance-covariance matrix of the prevalences is FIXED .

Full PAR (95% CI) for
  OVWT OBESE PACUMNEG
  BMILT185 SMKMOD SMKHI NINETIES AGE6069 AGEGE70 PREVCANC PREVCHD DB30 PREVHRT
```

```
           0.787 (0.747  , 0.821 )

Partial PAR (95% CI) for
  modifiable vbls :  OVWT OBESE PACUMNEG
  fixed vbls :  BMILT185 SMKMOD SMKHI NINETIES AGE6069 AGEGE70 PREVCANC PREVCH
D DB30 PREVHRT
           0.236 (0.203  , 0.268 )
------------------------------------------------------------------------------
```

The macro first gives the full PAR and its 95% confidence interval. Note that it lists the model variables before giving the PAR. Then the macro gives the partial PAR and its 95% confidence interval. It lists which variables were considered modifiable and which were considered fixed.

Since the model included age, as well as the presence of life-threatening diseases, it could account for 78.7% of the deaths. Overweight, obesity, and lack of exercise accounted for 23.6% of the deaths.

3.1.2 Using the results of a proportional hazards (Cox) model

Even though we recommend using pooled logistic regression to get the inputs to %PAR, we illustrate the use of %PAR with the results of a Cox model. The proportional hazards model contains the same predictors as the logistic model above. We *CANNOT* stratify by age and time period, because all variables must be explicit in the model. The model is

```
title2 'Cox model using all data';
proc phreg data=one covout outest=bplainc;
model talld*alldead(0)=
prevcanc prevchd db30 prevhrt
bmilt185 ovwt obese smkmod smkhi pacumneg nineties age6069 agege70;
run;
data here.bplainc;  set bplainc;  run;
```

Since we have actual person-time in the model, we used this to make *PDATA*.

```
proc sort data=one;  by
prevcanc prevchd db30 prevhrt
bmilt185 ovwt obese smkmod smkhi pacumneg nineties age6069 agege70;
run;
proc means noprint data=one;  var talld;
output  out=pplsum1  sum=fq;
by prevcanc prevchd db30 prevhrt
 bmilt185 ovwt obese smkmod smkhi pacumneg nineties age6069 agege70;
run;
proc means noprint data=pplsum1;  var fq;
output  out=pplsumtot  sum=fqtot;
run;

data pplsum1;
if _n_ eq 1 then set pplsumtot;
set pplsum1;
```

```
fq=fq/fqtot;
run;
data here.pplsum1;  set pplsum1;  run;
```

Here `fq` is the proportion of total person-time with the given set of covariate values.
The call to %PAR is

```
title3 'results from cox regression';
%par(bdata=here.bplainc, pdata=here.pplsum1, modvar=ovwt obese pacumneg,
 n_or_p=p, n_or_pname=fq, vprevopt=fixed, fullpar=t,
fixedvar=bmilt185 smkmod smkhi nineties age6069 agege70
 prevcanc prevchd db30 prevhrt);
```

The macro produced a harmless WARNING in the output, because Cox models have no intercept.
Finally, the output is

```
-------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program parrun   24MAY2011   16:02    stleh            1

results from cox regression

option for the variance-covariance matrix of the prevalences is FIXED .

Full PAR (95% CI) for
  OVWT OBESE PACUMNEG
  BMILT185 SMKMOD SMKHI NINETIES AGE6069 AGEGE70  PREVCANC PREVCHD DB30 PREVHR
T
           0.769 (0.728  , 0.804 )

Partial PAR (95% CI) for
  modifiable vbls :  OVWT OBESE PACUMNEG
  fixed vbls :  BMILT185 SMKMOD SMKHI NINETIES AGE6069 AGEGE70  PREVCANC PREVC
HD DB30 PREVHRT
           0.231 (0.2  , 0.262 )
-------------------------------------------------------------------------------
```

The computed full and partial PARs differ slightly from those we got from the pooled logistic model above because the coefficients in the Cox model differ slightly from those in the logistic model and because the time-based 'frequencies' used for the Cox model are slightly different from the counts used in for the logistic model.

3.2 Example 2. Errors in macro call

3.2.1 A variable left out of the lists

Here is a macro call in which one variable (`pacumneg`) is in neither the *MODVAR* list nor the *FIXEDVAR* list.

```
title3 'missing variable';
%par(bdata=bplain1, pdata=pplain1, modvar=ovwt obese ,
n_or_p=n, n_or_pname=fq,
fixedvar=bmilt185 smkmod smkhi nineties age6069 agege70 prevcanc prevchd db30 prevhrt, notes=n
```

The results are

```
------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program parrun   24MAY2011   16:17    stleh            1
vprevopt=fixed
missing variable
ERROR in macro call: The variable names in the macro call do not
   match those in the V-C matrix of the coefficients.
The macro will stop.
   See the output.

------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program parrun   24MAY2011   16:17    stleh            2
vprevopt=fixed
missing variable
Variable names that are not the same in the macro call
  and the BDATA dataset

                  in V-C    in macro
Obs      name     matrix      call

  8    PACUMNEG       1          0

------------------------------------------------------------------------------
```

The macro tells you that the variable pacumneg is in the original model (and therefore its variance-covariance matrix), but not in the macro call. This is a fatal error.

3.2.2 A variable appears in both *MODVAR* and *FIXEDVAR* lists

In this example, the variable pacumneg is in both the *MODVAR* and *FIXEDVAR* lists. The results are

```
------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program parrun   24MAY2011   16:17    stleh            7
vprevopt=fixed
overlapping vbls
ERROR in macro call:  The FIXEDVAR list overlaps with the MODVAR list.
The macro will stop.

------------------------------------------------------------------------------
```

```
/udd/stleh/doctn/par  Program parrun   24MAY2011   16:17    stleh           8
vprevopt=fixed
overlapping vbls
Variable names that are in both FIXEDVAR and MODVAR

Obs      name      _overlap_

 1     PACUMNEG        1
-------------------------------------------------------------------------------
```

The macro helpfully told you which variable overlapped.

3.3 No *SAMPSZ* given for *VPREVOPT*=COHORT

The macro call is

```
title3 'no sample size for a cohort';
%par(bdata=bplain1, pdata=pplain1, modvar=ovwt obese pacumneg,
n_or_p=n, n_or_pname=fq, vprevopt=cohort,
fixedvar= bmilt185 smkmod smkhi nineties age6069 agege70
 prevcanc prevchd db30 prevhrt);
```

The macro notes that you did not give a value for *SAMPSZ*, then proceeds to compute the PAR using *VPREVOPT*=FIXED.

```
-------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program parrun   24MAY2011   16:17    stleh           5
vprevopt=fixed
no sample size for a cohort
WARNING in macro run:  You specified COHORT variance-covariance
     matrix, but did not give a sample size.
  The macro will use VPREVOPT=FIXED.

-------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program parrun   24MAY2011   16:17    stleh           6
vprevopt=fixed
no sample size for a cohort

option for the variance-covariance matrix of the prevalences is FIXED .

Partial PAR (95% CI) for
  modifiable vbls :  OVWT OBESE PACUMNEG
  fixed vbls :  BMILT185 SMKMOD SMKHI NINETIES AGE6069 AGEGE70 PREVCANC PREVCH
D DB30 PREVHRT
```

```
          0.236 (0.203  , 0.268 )

------------------------------------------------------------------------------
```

3.3.1 No *MODVAR*

The macro call is

```
title3 'no modifiable variables';
%par(bdata=bplain1, pdata=pplain1,
n_or_p=n, n_or_pname=fq,
fixedvar=ovwt obese pacumneg bmilt185 smkmod smkhi nineties
 age6069 agege70 prevcanc prevchd db30 prevhrt);
```

The results are

```
------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program parrun   24MAY2011   16:17    stleh            3
vprevopt=fixed
no fixed vbls

option for the variance-covariance matrix of the prevalences is FIXED .

------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program parrun   24MAY2011   16:17    stleh            4
vprevopt=fixed
no modifiable variables

option for the variance-covariance matrix of the prevalences is FIXED .

Full PAR (95% CI) for

  OVWT OBESE PACUMNEG BMILT185 SMKMOD SMKHI NINETIES AGE6069 AGEGE70 PREVCANC
PREVCHD DB30 PREVHRT
          0.787 (0.747  , 0.821 )

------------------------------------------------------------------------------
```

Even though the macro call has the default settings (including *FULLPAR*=F), the macro produced the full PAR, since there were no modifiable variables given.

3.3.2 RR is less than 1 for some combination of the predictors

Because of the formulas used, the macro requires RR to be at least 1 for each combination of the model covariates *jointly* and each combination of the fixed covariates. If this is not true, the macro will stop. As an example, we show what happened when, instead of `pacumneg` (no physical activity), we use `pacumpos` (any physical activity) as the predictor. The macro call is

```
%par(bdata=here.bplain0, pdata=here.pplain0, modvar=ovwt obese pacumpos,
n_or_p=n, n_or_pname=fq,
fixedvar=bmilt185 smkmod smkhi nineties age6069 agege70 prevcanc prevchd
db30 prevhrt);
```

The macro output is

```
------------------------------------------------------------------------------

/udd/stleh/helpme/frankhu/parex  Program cucu   29JAN2007   17:24    stleh   1
ERROR in macro run:  Not all of the RRs for the combinations of
   all or fixed variables are at least 1.
   PAR does not compute correctly in these cases.
   Please change your reference levels.
The macro will stop.
```

The macro helpfully prints out the combinations in which the offending RRs are found.

```
------------------------------------------------------------------------------

/udd/stleh/helpme/frankhu/parex  Program cucu   29JAN2007   17:24    stleh   2


variable combinations with _rr_ or _fixrr_ < 1

                P   B           N           P                               
                A   M           I   A   A   R   P       P                   _
                C   I   S       N   G   G   E   R       R                   f
          O     U   L   M   S   E   E   E   V   E       E                   i
      O   B     M   T   K   M   T   6   G   C   V   D   V       _           x
 O    V   E     P   1   M   K   I   0   E   A   C   B   H       r           r
 b    W   S     O   8   O   H   E   6   7   N   H   3   R       r           r
 s    T   E     S   5   D   I   S   9   0   C   D   0   T       _           _

  7   0   0     1   0   0   0   0   0   0   0   0   0   0   0.75730    1.00000
  9   0   0     1   0   0   0   1   0   0   0   0   0   0   0.85746    1.13226
 27   1   0     1   0   0   0   0   0   0   0   0   0   0   0.90978    1.00000
125   0   0     1   0   0   0   0   0   0   0   1   0   0   0.82100    1.08412
127   0   0     1   0   0   0   1   0   0   0   1   0   0   0.92959    1.22751
145   1   0     1   0   0   0   0   0   0   0   1   0   0   0.98631    1.08412
```

Since all the observations printed out have `pacumpos`=1 (i.e. some physical activity), it is reasonable to start with changing the reference level for this variable, i.e. making `pacumneg` (i.e. the model variable is 'no physical activity'), as we did in the other examples.

3.4 Example 3. Including interaction terms AND Removing only some of the BMI-related risk

In this example, we have included interaction terms (created in the data set) for the BMI variables with age and with calendar time (`nineties`).

The interaction variables used are

```
effect             variable name
------             -------------
bmilt185*nineties  skinny90s
ovwt*nineties      ovwt90s
obese*nineties     obese90s
bmilt185*sixties   skinny60s
ovwt*age6069       ovwt60s
obese*age6069      obese60s
bmilt185*agege70   skinnyold
ovwt*agege70       ovwtold
obese*agege70      obeseold
```

The variance-covariance matrix of the coefficients was made using the following code.

```
proc logistic descending data=one covout outest=bintx1;  model alldead=
prevcanc prevchd db30 prevhrt
bmilt185 ovwt obese smkmod smkhi pacumneg nineties age6069 agege70
skinny90s ovwt90s obese90s skinny60s ovwt60s obese60s skinnyold ovwtold obeseold;
run;
```

The prevalence file was made by the following code.

```
proc sort data=one;  by
prevcanc prevchd db30 prevhrt
bmilt185 ovwt obese smkmod smkhi pacumneg nineties age6069 agege70
skinny90s ovwt90s obese90s skinny60s ovwt60s obese60s skinnyold ovwtold obeseold;
run;
proc means noprint data=one;  var alldead;
output  out=pintx1  n=fq;
by prevcanc prevchd db30 prevhrt
bmilt185 ovwt obese smkmod smkhi pacumneg nineties age6069 agege70
skinny90s ovwt90s obese90s skinny60s ovwt60s obese60s skinnyold ovwtold obeseold;
run;
```

Although this dataset has the same set of frequencies as `pplain1` used in Example 1, we have rerun the proc means with the interaction terms so that the interaction variables would be in the prevalence dataset. In fact, we could have run Example 1 using this prevalence dataset, or we could have added the interaction variables to the prevalence dataset used in Example 1.

Here is the summary of the odds ratios from this analysis.

```
            Odds Ratio Estimates

                Point          95% Wald
Effect       Estimate      Confidence Limits

PREVCANC        3.282       3.064       3.516
```

```
PREVCHD         1.084       0.938       1.253
DB30            4.127       3.174       5.368
PREVHRT         3.161       2.677       3.731
BMILT185        2.057       1.653       2.559
OVWT            1.283       1.164       1.414
OBESE           1.671       1.489       1.875
PACUMNEG        1.321       1.266       1.378
NINETIES        1.116       1.040       1.198
AGE6069         3.017       2.813       3.236
AGEGE70         6.524       5.962       7.141
SKINNY90S       0.755       0.563       1.013
OVWT90S         1.010       0.891       1.145
OBESE90S        1.145       0.989       1.326
SKINNY60S       1.290       0.962       1.730
OVWT60S         0.906       0.799       1.027
OBESE60S        0.979       0.846       1.133
SKINNYOLD       1.689       1.136       2.510
OVWTOLD         0.883       0.753       1.036
OBESEOLD        0.853       0.705       1.031
```

In the pooled logistic regression, the interaction terms `skinny90s`, `ovwt60s`, `obese60s`, `ovwtold`, and `obeseold` had negative coefficients. Nonetheless, since the coefficients of the interaction terms were small compared to the coefficients of the main effects, no combinations of covariates had RR less than 1, and the macro was able to run.
Here is the call to %PAR.

```
title4 'correct run, with interactions';
%par(bdata=bintx1, pdata=pintx1, modvar=ovwt obese pacumneg
 ovwt90s obese90s ovwt60s obese60s ovwtold obeseold,
n_or_p=n, n_or_pname=fq,
fullpar=t,
fixedvar=bmilt185 smkmod smkhi nineties age6069 agege70
 prevcanc prevchd db30 prevhrt skinny90s skinny60s skinnyold);
```

Note that in this call to %PAR, we have included only the effects associated with high BMI in the *MODVAR*. The effects associated with low BMI are listed in *FIXEDVAR*. This corresponds to the removal of risk associated with high, but not low, BMI. Here are the results

```
------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program parrun   08MAR2012   13:04    stleh           15


correct run, with interactions

option for the variance-covariance matrix of the prevalences is FIXED .

Full PAR (95% CI) for
  OVWT OBESE PACUMNEG  OVWT90S OBESE90S OVWT60S OBESE60S OVWTOLD OBESEOLD
  BMILT185 SMKMOD SMKHI NINETIES AGE6069 AGEGE70  PREVCANC PREVCHD DB30 PREVHR
```

```
T SKINNY90S SKINNY60S SKINNYOLD
           0.767 (0.658  , 0.844 )

Partial PAR (95% CI) for
  modifiable vbls :  OVWT OBESE PACUMNEG  OVWT90S OBESE90S OVWT60S OBESE60S OV
WTOLD OBESEOLD
  fixed vbls :  BMILT185 SMKMOD SMKHI NINETIES AGE6069 AGEGE70  PREVCANC PREVC
HD DB30 PREVHRT SKINNY90S SKINNY60S SKINNYOLD
           0.223 (0.083  , 0.355 )
```

The model with the interactions did not account for any more of the total deaths than the model without interactions, but the confidence intervals widened somewhat. This is not surprising, since the likelihood ratio test p-values for the `ovwt60s ovwtold ovwt90s` and `obese60s obeseold obese90s` interactions (each of which has 3 degrees of freedom) were both greater than 0.15. For a discussion of when the PAR has the distributive property, see section 5.1.

Although it does not appear in the output, the number of observations in `PINTX1` was the same as the number of observations in `BPLAIN1`, because the interactions do not produce any new unique combinations of the variables.

3.5 Example 4. Using prevalences from a complex survey design

In this example, we retrieved data on age, sex, weight, height, and current smoking from NHANES 2003-2004. SAS PROC SURVEYFREQ will give the variance-covariance matrix of the estimated weighted numbers of people, but only by using an undocumented option. Furthermore, SAS does not put this out as a dataset, but only shows it in the output. For ease of reading from a text file that is a SAS output, we made a very simple model so we could have a relatively small number of unique combinations of the variables. The whole variance-covariance matrix of the frequencies is only printed out for bivariate tables. To get our data into this form, we made some combined variables and ran PROC SURVEYFREQ using the following program.

```
libname rvd '/udd/stleh/helpme/rvandam';
libname bfh '/udd/stleh/helpme/frankhu/parex';

/* read in data downloaded from nhanes 2003-2004 web site */
data nh0304;  infile '/udd/stleh/helpme/rvandam/nh0304.dat' missover delimiter='|';
input seqno age gender bmi wt ht triskf eversmoke samwt mv1 mv2;

/* change order of smoking variable (1, 2)-->(1, 0) */
eversmoke=2-eversmoke;  if eversmoke lt 0 then eversmoke=.;

/* change gender to indicator with 1=male */
msex=2-gender;  if msex lt 0 then msex=.;

/* make up overweight and obese variables */
if bmi ne . then do;
  obese=0;  if bmi ge 30 then obese=1;
  ovwt=0;  if bmi ge 25 and bmi lt 30 then ovwt=1;
end;
```

```
agegp=int(age/5);  /* 5-year age groups.  don't worry that
                      the lowest group is not labeled 0 or 1 */

/* make smoking and obese into one formal variable so can get the
   smoking*obese*agegp table as a 2-way table in PROC SURVEYFREQ  */
smokob=10*eversmoke+obese;
run;

/* cluster, strata, and weight below are related to the sampling design */
proc surveyfreq data=nh0304;
cluster mv1;  strata mv2;
tables smokob*agegp / wchisq debugwcov;  /* WCHISQ DEBUGWCOV is the
                                            undocumented option you have to use
                                            to get the v-c matrix of the
                                            weighted frequencies */
weight samwt;
ods output crosstabs=tables;
/* limiting to women with known obesity and smoking status and
   agegroups that overlap with our data (ages 50-74) */
where msex eq 0 and 10 le agegp le 14 and obese in (0, 1) and eversmoke ne .;
run;

/* make permanent data set */
data bfh.prevso;  set tables;  run;
```

Part of the output of PROC SURVEYFREQ is shown below.

```
/udd/stleh/helpme/rvandam  Program nh0304   14FEB2007   17:37    stleh

The SURVEYFREQ Procedure

            Data Summary

Number of Strata                 15
Number of Clusters               30
Number of Observations          800
Sum of Weights           31560652.5


                          Table of smokob by agegp

                                    Weighted    Std Dev of              Std Err of
smokob    agegp     Frequency     Frequency      Wgt Freq    Percent       Percent
-------------------------------------------------------------------------------------
     0       10            58       3610140        780300    11.4387        1.7472
             11            41       2335692        293883     7.4006        1.0186
             12            51       1783053        362444     5.6496        0.8872
             13            56       1733339        325185     5.4921        0.7552
             14            50       1703429        351254     5.3973        0.8128
```

	Total	256	11165652	1579836	35.3784	2.4748
1	10	34	1793307	465846	5.6821	1.4379
	11	21	1005715	284149	3.1866	0.8300
	12	55	1212853	186174	3.8429	0.5516
	13	41	1004979	231840	3.1843	0.6050
	14	32	894738	296437	2.8350	0.9317
	Total	183	5911592	851651	18.7309	2.2074
10	10	44	2342235	372297	7.4214	1.1273
	11	29	1805287	371237	5.7201	0.9675
	12	64	2096361	338701	6.6423	0.6151
	13	39	1214364	181557	3.8477	0.5099
	14	39	1408715	288612	4.4635	0.7126
	Total	215	8866962	983458	28.0950	1.0944
11	10	37	1663698	266541	5.2714	0.7435
	11	24	1626711	278892	5.1542	0.7434
	12	33	824558	261438	2.6126	0.7715
	13	30	912657	239562	2.8918	0.7047
	14	22	588821	181665	1.8657	0.6441
	Total	146	5616446	653909	17.7957	1.6281
Total	10	173	9409381	1241233	29.8136	2.3486
	11	115	6773405	777594	21.4615	1.7649
	12	203	5916825	841781	18.7475	1.5166
	13	166	4865339	593610	15.4158	0.7810
	14	143	4595703	581848	14.5615	1.1609
	Total	800	31560652	3247876	100.000	

Covariance Matrix of the Estimated Cell Totals

6.08868E11 5.25925E10 1.32529E11 1.1413E11 1.6593E11 1.11055E10 -1.804E10 3.5691E10 1.98392E10 -1.7787E9 9.36186E10 1.18609E11 1.67302E11
6.42732E10 1.28066E11 3.74025E10 9.36096E10 3.70861E10 1.00314E11 -2.5545E10 5.25925E10 8.63671E10 7657389948 9094294792 4.48119E10
-2.9254E10 5124764008 -1.6103E9 9220707776 1.33099E10 1.64459E10 4.30051E10 4455856042 -5.17849E9 8395226897 1.2641E10 1.59781E10 -
3.88198E9 859093388 -1.4437E10 1.32529E11 7657389948 1.31366E11 5.59797E10 7.12572E10 -1.8567E10 2.22573E10 4.00909E10 3.19171E10
8727974639 -2.5789E10 7.96129E10 7.318E10 1.30122E10 5.82201E10 1.75586E10 3.31893E10 1.39511E10 4.23484E10 -2.1763E10 1.1413E11 9094294792
5.59797E10 1.05745E11 6.60966E10 -1.7513E10 1.13022E10 3.44216E10 2.70397E10 4.21835E10 2.45929E10 6.40034E10 6.08314E10 2.85787E10
6.73746E10 4.74794E10 6386004012 2.49769E10 -9.20621E9 -2.8294E10 1.6593E11 4.48119E10 7.12572E10 6.60966E10 1.23379E11 -1.5791E10
1.44627E10 1.21643E10 1.18072E10 -3.91087E9 1.39748E10 5.27894E10 7.76205E10 2.52301E10 5.75841E10 4.13223E10 1.88849E10 2975140297
4.68427E10 -9.37965E9 1.11055E10 -2.9254E10 -1.8567E10 -1.7513E10 -1.5791E10 2.17012E11 3.42123E10 -2.219E10 4.66426E10 -4.0489E10 7.40847E10
-3.5277E10 3.68308E10 1.76745E10 -4.4602E10 -3.23271E9 7.69878E10 5.41495E10 9602416713 4.58097E10 -1.804E10 5124764008 2.22573E10
1.13022E10 1.44627E10 3.42123E10 8.07409E10 1571759351 3.93244E10 2.34721E10 -6.9655E9 2.75156E10 4.63773E10 1719973781 2.56375E10
2.36938E10 1.13008E10 2314190467 1.04881E10 1793767846 3.5691E10 -1.6103E9 4.00909E10 3.44216E10 1.21643E10 -2.219E10 1571759351 3.46609E10
1.67731E10 1.29279E10 1227729084 3.78967E10 2.15362E10 -658540247 2.98656E10 7340834411 4693530026 1.41151E10 -8.69663E9 -1.2532E10
1.98392E10 9220707776 3.19171E10 2.70397E10 1.18072E10 4.66426E10 3.93244E10 1.67731E10 5.37499E10 1.33905E10 2.64717E10 2.93408E10

4.12381E10 -855832812 7896024465 1.60447E10 3.74435E10 2.03317E10 -6.94004E9 -1.19061E9 -1.7787E9 1.33099E10 8727974639 4.21835E10 -3.91087E9 -4.0489E10 2.34721E10 1.29279E10 1.33905E10 8.78749E10 5555126210 4.84325E10 4100830730 2.00709E10 2.6625E10 1.4698E10 -1.181E10 -2.6782E10 -2.1956E10 -4.0402E10 9.36186E10 1.64459E10 -2.5789E10 2.45929E10 1.39748E10 7.40847E10 -6.9655E9 1227729084 2.64717E10 5555126210 1.38605E11 -1.8714E10 2.78952E10 2.00484E10 -1.6792E10 3.95859E10 4.58373E10 1.98097E10 -1.0304E10 3249671436 1.18609E11 4.30051E10 7.96129E10 6.40034E10 5.27894E10 -3.5277E10 2.75156E10 3.78967E10 2.93408E10 4.84325E10 -1.8714E10 1.37817E11 5.77083E10 -335304323 6.90154E10 1.30563E10 8048869393 -2.9373E9 2636125315 -4.08E10 1.67302E11 4455856042 7.318E10 6.08314E10 7.76205E10 3.68308E10 4.63773E10 2.15362E10 4.12381E10 4100830730 2.78952E10 5.77083E10 1.14718E11 2.16463E10 5.33597E10 3.50794E10 3.06241E10 1.68766E10 3.572E10 3102561326 6.42732E10 -5.17849E9 1.30122E10 2.85787E10 2.52301E10 1.76745E10 1719973781 -658540247 -855832812 2.00709E10 2.00484E10 -335304323 2.16463E10 3.29628E10 1.60628E10 1.91829E10 6716505578 7184536490 1.26462E10 -5.18528E9 1.28066E11 8395226897 5.82201E10 6.73746E10 5.75841E10 -4.4602E10 2.56375E10 2.98656E10 7896024465 2.6625E10 -1.6792E10 6.90154E10 5.33597E10 1.60628E10 8.32969E10 3.10147E10 -3.36132E9 1.30957E10 1.61908E10 -2.502E10 3.74025E10 1.2641E10 1.75586E10 4.74794E10 4.13223E10 -3.23271E9 2.36938E10 7340834411 1.60447E10 1.4698E10 3.95859E10 1.30563E10 3.50794E10 1.91829E10 3.10147E10 7.10441E10 -5.10694E9 2.45518E10 -3.12719E9 -1.0612E10 9.36096E10 1.59781E10 3.31893E10 6386004012 1.88849E10 7.69878E10 1.13008E10 4693530026 3.74435E10 -1.181E10 4.58373E10 8048869393 3.06241E10 6716505578 -3.36132E9 -5.10694E9 7.77809E10 2.27243E10 2.03422E10 2180428076 3.70861E10 -3.88198E9 1.39511E10 2.49769E10 2975140297 5.41495E10 2314190467 1.41151E10 2.03317E10 -2.6782E10 1.98097E10 -2.9373E9 1.68766E10 7184536490 1.30957E10 2.45518E10 2.27243E10 6.835E10 -1.2319E10 1.47808E10 1.00314E11 859093388 4.23484E10 -9.20621E9 4.68427E10 9602416713 1.04881E10 -8.69663E9 -6.94004E9 -2.1956E10 -1.0304E10 2636125315 3.572E10 1.26462E10 1.61908E10 -3.12719E9 2.03422E10 -1.2319E10 5.73899E10 6600631854 -2.5545E10 -1.4437E10 -2.1763E10 -2.8294E10 -9.37965E9 4.58097E10 1793767846 -1.2532E10 -1.19061E9 -4.0402E10 3249671436 -4.08E10 3102561326 -5.18528E9 -2.502E10 -1.0612E10 2180428076 1.47808E10 6600631854 3.30023E10 In the matrix above, the lines that begin with spaces are continuations of the previous lines, split to fit on the page.

Finally, here is code using %PAR on these data. Note that since the variance-covariance matrix we get from PROC SURVEYFREQ is for the estimated frequencies, we have to divide by the square of the sum of the weighted frequencies to get the variance-covariance matrix of the prevalences.

```
data vc;  infile '/udd/stleh/helpme/frankhu/parex/surveyvcfreq.txt' missover;
input w1-w20;
array vcs vc1-vc20;  array ws w1-w20;
do over vcs;  vcs=ws/31560652**2;  end;
                    /* the denominator here is from the weighted total
                       population used in this analysis, which comes from
                       the last line of the table
                       Total Total in the 'Weighted Frequency' column  */
keep vc1-vc20;
run;

data prev;  set rvd.prevso end=_end_;  prev=percent/100;
                                          /* percent is a variable
                                             in the ODS dataset */
/* make age indicators as in NHS model */
%indic3(vbl=agegp, prefix=agegp, min=11, max=14, reflev=10, usemiss=0);
/* recover eversmoke and obese from the combined variable smokob */
eversmoke=int(smokob/10);
obese=mod(smokob, 10);
if obese in (0, 1) and 10 le agegp le 14 and eversmoke ne . ;
            /* this gets rid of the 'total' and blank lines */
keep agegp11 agegp12 agegp13 agegp14 eversmoke obese prev;
run;


title2 'using vprevopt=survey';
```

```
%par(bdata=bfh.bsurvey, pdata=prev, modvar=obese,
fullpar=t,
notes=notes,
n_or_p=p, n_or_pname=prev,
vpvarname=vc,
fixedvar=agegp11 agegp12 agegp13 agegp14 eversmoke, vprevopt=survey, vpdata=vc);
```

The output of %PAR is

```
------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program trypopn   24MAY2011   17:14    stleh    7
using vprevopt=survey

option for the variance-covariance matrix of the prevalences is SURVEY .

Full PAR (95% CI) for
  OBESE
  AGEGP11 AGEGP12 AGEGP13 AGEGP14 EVERSMOKE
           0.805 (0.723  , 0.864 )

Partial PAR (95% CI) for
  modifiable vbls :  OBESE
  fixed vbls :  AGEGP11 AGEGP12 AGEGP13 AGEGP14 EVERSMOKE
           0.214 (-0.099  , 0.489 )

------------------------------------------------------------------------------
```

Just for comparison, we also ran %PAR with vprevopt=fixed.

```
------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program trypopn   24MAY2011   17:31    stleh          9
using vprevopt=fixed

option for the variance-covariance matrix of the prevalences is FIXED .

Full PAR (95% CI) for
  OBESE
  AGEGP11 AGEGP12 AGEGP13 AGEGP14 EVERSMOKE
           0.805 (0.779  , 0.827 )

Partial PAR (95% CI) for
  modifiable vbls :  OBESE
  fixed vbls :  AGEGP11 AGEGP12 AGEGP13 AGEGP14 EVERSMOKE
           0.214 (0.191  , 0.237 )

------------------------------------------------------------------------------
```

As another comparison, we ran %PAR using the frequencies from the NHS data, with the following results:

```
------------------------------------------------------------------------------

/udd/stleh/doctn/par  Program trypopn   24MAY2011   17:31    stleh          10
using vprevopt=fixed

option for the variance-covariance matrix of the prevalences is FIXED .

Full PAR (95% CI) for
  OBESE
  AGEGP11 AGEGP12 AGEGP13 AGEGP14 EVERSMOKE
           0.749 (0.722  , 0.775 )

Partial PAR (95% CI) for
  modifiable vbls :  OBESE
  fixed vbls :  AGEGP11 AGEGP12 AGEGP13 AGEGP14 EVERSMOKE
           0.076 (0.067  , 0.085 )
------------------------------------------------------------------------------
```

4 Frequently Asked Questions

4.1 Q: Reference level is not lowest RR or OR

A: Your PAR question is then something like 'What fraction of cases would not occur if everyone had the reference level of the covariate?'
Make sure that you really care about this question, rather than, '...if everyone had the minimum risk level?'
If you have a univariate model, %PAR will not work, but it is easy to compute the point estimate, and not so hard to compute the confidence interval, by hand.
If you have a multivariate model, the usability of %PAR will depend on whether the distribution of covariates is such that all actual cells of the contingency table have RR/OR ge 1. For example, if age is stronger than the exposure for which you are not using the minimum risk level as the reference, and there are no observations with the lower levels of risk in the age groups with the lowest levels of risk, all cells may have RR/OR ge 1. Otherwise, %PAR will not work.
One possible strategy is to combine groups so that the group including the reference level is indeed the minimum risk group.

4.2 Q: How can I use %PAR with interaction variables?

A: Interaction variables are just like any other variables in this macro. In this version, since you don't have to make up the prevalence file in a special way, all coefficients will be taken care of appropriately.
See Example 3.

4.3 Q: What if I have negative interactions?

A: In many cases, these negative interactions will not be enough to make the RR/OR for the cell less than 1, as the sum of the two coefficients for the main effects will be greater in absolute value than the interaction coefficient. See Example 3. If this is not the case, you may be able to reparameterize your model. If nothing works, you cannot use %PAR. How can a situation arise where reparameterization is not possible? Consider the following set of estimated coefficients from a hypothetical study similar to Example 3.

```
variable     coefficient
--------     -----------
old age      0.1
obese        0.15
old*obese   -0.3
```

Now consider the net coefficients (i.e. the log(RR) or log(OR) for the unique combinations of the predictors) for a dataset with all possible logical combinations of old age and obesity.

```
old age   obesity  net coefficient
-------   -------  ---------------
no        no       0
no        yes      0.15
yes       no       0.10
yes       yes     -0.05
```

There is no way to reparameterize the data so that all net coefficients will be positive.

4.4 Q: SAS says I ran out of memory

```
ERROR: Out of memory.
```

A: You probably need to ask for more memory. The shell command on the Channing Laboratory system is

```
qbs -q l -o memsize=2048M <program.name>
```

This shell command waits till a server has the required amount of memory available, then runs on that server. NOTE: The program that ran the models to make Table 2 of Spiegelman, Hertzmark, and Wand used 1.8 gigabytes of memory. On the Channing Laboratory system, you can increase `memsize` in increments of 1024M up to 6144M.

4.5 Q: The program has been running for a long time and is still not done.

A: If you are using VPREVOPT=SURVEY, the macro takes a very long time, which depends on the size of the prevalence dataset. See 'Choice of VPREVOPT' below.
NOTE: All the models reported in Table 2 of Spiegelman, Hertzmark, and Wand used a little over 5 cpu-hours, with the second column taking 4 hours and 41 minutes.

4.6 Q: The coefficient for my exposure is significant, but the confidence interval for PAR includes 0

A: Because we use nonlinear transformations to compute the confidence interval, this can happen. Even so, 0 should be close to the confidence bound.

4.7 Q: I ran my model with output from a PROC PHREG (a proportional hazards or Cox model) and got WARNINGS

A: We do not recommend using the output from Cox models as input to %PAR.
Nonetheless, the warnings shown below are completely harmless and result from the fact that the Cox model has no intercept and from the fact that the COVOUT for proportional hazards models does not include the variable `_link_`.

```
WARNING: The variable intercept in the DROP, KEEP, or RENAME list has never
         been referenced.
WARNING: The variable _link_ in the DROP, KEEP, or RENAME list has never been
         referenced.
```

5 Useful Knowledge

5.1 PAR is distributive in univariate models

Note that the PAR% is distributive in a univariate model(Wacholder et al. AJE 1994; 140:303-9). That is, in a univariate model, the PAR from an exposure with multiple levels (e.g. bmi1(ref), bmi2, bmi3, bmi4, bmi5) equals the sum of the PARs of the non-referent individual levels grouped into a single level (e.g. bmi1(ref), bmi2+bmi3+bmi4+bmi5). Therefore, you can simplify your problem for PAR calculations by turning all the multiple level variables into binary ones (reference level vs. not). For multivariate models, Wacholder's result does not hold, although in examples we have looked at, it seems to be a reasonable approximation (Spiegelman, et al, 2006). Approximate distributivity may be the reason that the partial PAR for the model with interactions (Example 3) is the same as the partial PAR for the model without interactions (Example 1).

5.2 Choice of *VPREVOPT*:

The speed at which %PAR runs depends heavily on the size of the *VPREVNAME* dataset. This depends on the number of unique combinations of the covariates. If it is reasonable to use *VPREVOPT*=FIXED, do so. Otherwise, try to minimize the number of unique combinations by grouping variables. See **Computational Methods** below for a comparison of *VPREVOPT*=FIXED with *VPREVOPT*=COHORT.

6 Computational Methods

6.1 Variance-covariance matrix of the prevalences in a cohort study

If *VPREVOPT*=COHORT, the variance-covariance matrix of the prevalences is computed based on the multinomial distribution.
Suppose that we have n estimated prevalences, $\hat{p}_1, \ldots, \hat{p}_n$.
According to multinomial distribution theory, the covariance for the estimates $\hat{p}_1, \ldots, \hat{p}_n$ is

$$\mathrm{Cov}(p_i, p_j) = p_i(1 - p_i)/n \text{ if } i = j$$

$$\mathrm{Cov}(p_i, p_j) = -p_i p_j/n \text{ if } i \neq j$$

Since for large n all the elements of the matrix will approach 0, the component of the variance of PAR that derives from the variance-covariance matrix of the prevalences also goes to 0, so that an adequate estimate of the confidence interval can be gotten by using *VPREVOPT*=FIXED.

6.2 Comparison of *VPREVOPT*=COHORT and *VPREVOPT*=FIXED for various values of *SAMPSZ* in an example

VPREVOPT=COHORT was compared to *VPREVOPT*=FIXED for the 'crude' models in column 1 of Table 2 in Spiegelman, Hertzmark, and Wand (see refs.). SAMPSZ=221000, 22100, 2210, and 221 were tried. No difference to 3 decimal paces (in the point estimate or the 95% confidence bounds) was seen with N=22100. No difference greater than .001 was found for n=2210. The largest absolute difference seen for n=221 was

```
VPREVOPT   95% CI

FIXED      (.323, .639)
COHORT     (.310, .647)
```

It therefore seems 'safe' to use *VPREVOPT*=FIXED for any cohort with sample size above 10000.

6.3 Computation of confidence intervals

The variance of the PAR is approximated using the multivariate delta method (see Spiegelman, Hertzmark, and Wand). To improve the asymptotic behavior of the confidence bounds, the point estimate is transformed using Fisher's Z transformation. The variance of this is then computed using the delta method with the original variance estimate. After this, upper and lower confidence limits for the transformed variable are computed. Then all values are taken back to the original scale by the inverse of Fisher's Z transformation.

7 Credits

Written by Ellen Hertzmark, Handan Wand, and Donna Spiegelman for the Channing Laboratory.
Questions can be directed to Ellen Hertzmark
`stleh@channing.harvard.edu`, (617) 432-4597.

8 References

1. Spiegelman, D., Hertzmark, E., Wand, H.C.: Point and interval estimates of partial population attributable risk in cohort studies: Examples and Software. In press, Cancer Causes and Control 2007. (available at `http://www.hsph.harvard.edu/faculty/spiegelman/par.html`).

2. Bruzzi P, Green SB, Byar DP, Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol 122:904-914, 1985

3. Smith-Warner, S.A., Spiegelman, D., Yaun, S.-S., Adami, H.-O., Beeson, L. van den Brandt, P.A., Colditz, G.A., Folsom, A.R., Fraser, G.E., Goldbohm, R.A., Miller, A.B., Potter, J.D., Rohan, T.E., Wand, H., Willett, W.C., Wolk, A., Hunter, D.J., Population attributable risk of postmenopausal breast cancer due to established breast cancer risk factors (in progress), 1999

4. Wacholder S., Benichou J., Heineman F.E., Hartge P., and Hoover R.N., Attributable Risk: Advantages of a Broad Definition of Exposure, Am J Epidemiol 140: 303-309, 1994

5. Walter, S.D. The Estimation and Interpretation of Attributable Risk in Health Research, Biometrics 32, 829-849, 1976

6. Shah, B.V., Barnwell, B.G., Bieler, G.S. SUDAAN User's Manual, Release 7.0. Research Triangle Park, NC. Research Triangle Institude, 1996.