

The SAS SUBTYPE Macro

Aya Kuchiba, Molin Wang, and Donna Spiegelman

April 8, 2014

Abstract

The %SUBTYPE macro examines whether the effects of the exposure(s) vary by subtypes of a disease. It can be applied to data from the cohort studies, nested or matched case-control studies, unmatched case-control studies and case-case studies.

Keywords: SAS macro, etiologic heterogeneity, competing risk analysis, cohort study, case-control study, case-case study, subtypes

Contents

1	Description	2
2	Invocation and Details	3
3	Examples	7
3.1	Example 1. Cohort study analysis with the standard counting process data format	8
3.2	Example 2. Cohort study analysis with the augmented data set	11
3.3	Example 3. Nested or matched case-control study analysis	12
3.4	Example 4. Unmatched case-control study analysis . .	15

3.5	Example 5. Case-case study analysis	19
4	Warnings	22
5	How should I describe this in my Methods section?	22
6	Correspondence	23
7	Other reference	23

1 Description

`%SUBTYPE` is a SAS macro that examines whether the effect of the exposure(s) vary by subtypes of a disease in the cohort studies, matched or unmatched case-control studies or case-case studies. Let β_j be the log relative risks of the exposure for subtype j , $j = 1, 2, \dots, J$. It provides overall heterogeneity test ($H_0 : \beta_1 = \beta_2 = \dots = \beta_J$) and pair-wise heterogeneity tests ($H_{01} : \beta_1 = \beta_2, \beta_1 = \beta_3, \dots, \beta_{J-1} = \beta_J$) performed by the likelihood ratio test or Wald test. It provides the constrained and unconstrained models for adjusting the potential confounders. In the constrained model, the effects of the covariates are assumed to be the same across the subtypes; in the unconstrained model, the effects of the covariates are allowed to be different by the subtypes.

For cohort study, the macro uses Cox proportional hazards model with a data augmentation method. It works with both an augmented data set created by the user and a standard data set, for which the macro creates the augmented data set. It allows the constrained and unconstrained models. The model-based variance-covariance matrix estimate is used, unless the user specifies `COV=YES`, which requests robust sandwich variance-covariance matrix estimates. The heterogeneity test is performed by the likelihood ratio test (by default). The Wald test is available with `WALD=YES`.

For nested or matched case-control study, the macro uses the conditional logistic regression model. It allows the constrained and unconstrained models. The model-based variance-covariance matrix estimate is used, unless the user specifies `COV=YES`, which requests robust sandwich variance-covariance

matrix estimates. The heterogeneity test is performed by the likelihood ratio test (by default). The Wald test is available with `WALD=YES`.

For unmatched case-control study, the macro provides two approaches. By default, it uses unconditional nominal polytomous logistic regression model. It provides the unconstrained analysis and Wald test for the heterogeneity test, using the model-based variance-covariance matrix estimate. The other approach is conducted by conditional logistic regression analysis with a data augmentation method. If the user chooses this approach by specifying `conditional=YES`, the macro creates the augmented data set. It allows the user to request the constrained model for some or all covariates, likelihood ratio test for the heterogeneity test and the robust sandwich variance-covariance matrix estimate, in addition to the analysis options available in the first approach.

For case-case study, the macro uses unconditional nominal polytomous logistic regression model. It provides the unconstrained analysis and Wald test for the heterogeneity test, using the model-based variance-covariance matrix estimate. Note that unlike the above three study designs, the case-case study provides the heterogeneity tests only, not estimating and testing the effects of exposures on the risk on each subtype.

2 Invocation and Details

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

```
options mautosource sasautos=<directories macro is located>;
```

In the Channing servers, the option statements might be

```
options mautosource sasautos='/usr/local/channing/sasautos';
```

In the rest of this section, we will list all the input parameters, some of which are required and some of which are optional.

```
%macro subtype(
```

data=, name of data set on which the analysis is conducted

studydesign=COHORT, COHORT if cohort study, MCACO if matched or nested case-control study, CACO if case-control study, CACA if case-case study (the default value is COHORT)

id=ID, subject IDs; each subject may have multiple entries; required when studydesign=COHORT (the default value is ID)

augmented=YES/NO; YES if the input dataset is augmented for every outcome subtype; applicable only if studydesign=COHORT; the default value is NO

exposure=, the exposure variable(s); the heterogeneity test is for comparing coefficient(s) of this/these variable(s); the macro can handle multiple exposure variables , which can be indicator variables for a categorical exposure, which should be put in curly brackets, or multiple exposures, for each of which the heterogeneity test is performed; for a cohort study, if augmented=YES, the variable names should have the suffix _j indicating subtypes (j=1,2,...,J total subtypes) and the variables should be sorted by subtypes in curly brackets. For example, if you have two exposures, a 3-level categorical exposure alcohol drinking, with indicators, alco2 and alco3, and another binary exposure bmi (body mass index), and J=3, for augmented=YES, this macro parameter should be defined as {alco2_1 alco3_1 alco2_2 alco3_2 alco2_3 alco3_3} {bmi_1 bmi_2 bmi_3}; if the data set is no augmented, this macro parameter should be {alco2 alco3} bmi.

time=, time-to-failure variable used in the model statement of PROC PHREG; a single failure-time variable, or t2 of at-risk intervals (t1,t2] for the counting process format; required if studydesign=COHORT; otherwise not applicable.

entrytime=, entry time variable, t1, of the at-risk intervals (t1,t2], mentioned in the description above for macro parameter time; applicable if studydesign=COHORT; if the user specifies a single failure-time variable, this parameter should be empty.

eventtype=, subtype variable, required for all designs; for a cohort study, if augmented=YES, the specified variable takes on the value j for all person-times for the outcome subtype j (j=1,2,...,J total subtypes) and censoring status will be specified in the parameter censoring; if augmented=NO, the variable specified has value j if the outcome subtype j has occurred by end of follow up or 0 if censored; for a case-control or case-case study, the variable has j for cases with outcome subtype j and 0 for controls (in case-control study)

censoring=, censoring variable. The variable takes on value 0 if censored and 1 if the corresponding outcome subtype contained in eventtype occurs; applicable only if augmented=YES

unconstrvar (optional)= names of covariates, not including the exposure variables, of which the associations with the outcome may be different for different outcome subtypes

constrvar (optional)= names of covariates, not including the exposure variables, of which the associations

with the outcome are forced to be the same across subtypes of outcome

stratavar (optional)= stratification variables; only applicable if studydesign=COHORT, MCACO, or CACO with conditional=YES

matchid= matched set variable code; applicable only if studydesign=MCACO

reftype= reference subtype variable code; applicable only if studydesign=CACA; the default value is 1

conditional= YES/NO; YES if requesting conditional logistic regression analysis for unmatched case-control study; this allows the constrained analysis and heterogeneity test by likelihood ratio test; applicable only if studydesign=CACO; the default value is NO

covs= YES/NO; YES if requesting the robust sandwich covariance matrix estimate; applicable only if studydesign=COHORT, MCACO, or CACO with conditional=YES; the default value is NO

wald= YES/NO; YES if requesting Wald test for the heterogeneity test, in addition to the default likelihood ratio test; only applicable if studydesign=COHORT, MCACO, or CACO with conditional=YES; Wald test is the only heterogeneity test available (and is the default test) for studydesign=CACA and CACO with conditional=NO; the default value is NO

covout= YES/NO; YES if requesting to display the estimated covariance matrix of the parameter estimates; the default value is NO

```

eventtypelabel (optional)= it can be used to define
                           the coding of eventtype; please do not use ',,'
                           here; for example, note = 1=high; 2=low;

paramest (optional)= name of the SAS dataset
                     containing the parameter estimates

heterotest (optional)= name of the SAS dataset
                       containing the results from the
                       heterogeneity tests; if the Wald test is
                       requested with
                       studydesign=COHORT, MCACO, or CACO
                       with conditional=YES, those results are
                       contained in the dataset named heterotest_WT

covest (optional)= name of SAS dataset containing the estimated
                   covariance matrix of the parameter estimates

);

```

3 Examples

The examples below describe the macro calls for each study design, using data from a study of the alcohol effects on LINE-1 methylation subtypes of colon cancer in the Health Professional Follow-up study. The outcome is incidence colon cancer defined by LINE-1 methylation status; there are three subtypes: LINE-1 high, medium and low. The exposure of interest is alcohol intake and we'll focus on the trend test for median alcohol intake at the baseline (0g/day, 1.8g/day, 10.2g/day, 27.5g/day) divided by the standard alcohol serving unit of 12g/day. The potential confounders controlled for in the analysis include current aspirin use, body mass index, history of screening, physical activity, history of prior polyps, family history of colon cancer, pack year of smoking, red meat intake, multivitamin use, calcium intake and folate intake, which are all categorical variables.

All data sets used in the example include the following variables:

id study subject's unique ID
 cancer outcome variable
 (1 for LINE-1 high, 2 for median, 3 for low,
 0 for non-cancer)
 alcohol exposure score for alcohol intake
 (0, 0.15, 0.85, 2.29)

The other design-specific variables will be described in each Example section

3.1 Example 1. Cohort study analysis with the standard counting process data format

The data set, cohort1, below is in the standard counting process data format, where period is questionnaire period, agemo is age in months at the beginning of each questionnaire period, time is the months from the start of the questionnaire cycle until date of colon cancer incidence, date of death, or date of the end of questionnaire period, whichever happens first.

Cohort1:

id	time	cancer	period	agemo	alcohol	OTHER COVARIATES
1	20	0	1	560	0.15	...
1	23	0	2	580	0.15	...
1	16	1	3	603	0.15	...
...						
2	23	0	1	606	0	...
2	21	0	2	623	0	...
2	19	0	3	644	0	...
2	25	0	4	663	0	...
...						

The macro call to apply the unconstrained model for all covariates is:

```

%subtype(data=cohort1, studydesign=cohort, id=id,
exposure=alcohol, augmented=no, time=time, eventtype=cancer,
unconstrvar=ause_p2 screen2 polyps2 cafam2

```

```

py30ct2 py30ct3 py30ct4 py30ct5 py30ctm
actct2 actct3 actct4 actct5 actctm
mvit2 mvitm bmain2 bmain3 bmain4
bmi2 bmi3 bmi4 bmi5 bmim
calcq2 calcq3 calcq4 calcq5 calcqm
folq2 folq3 folq4 folq5, stratavar=agemo period,
eventtypelabel=1=high; 2=medium; 3=low,
heterotest=heterogeneity);

```

For using the constrained models for some or all covariates, those covariates can be placed in CONSTRVAR .

The output is

```

=====
Running on data set COHORT1, Read 47363 observations 52
Tie handling: BRESLOW
CANCER: 1=high; 2=medium; 3=low
Number of cases in each outcome type

      cancer      Frequency
      Count
      1          99
      2         102
      3          67
=====

Running on data set COHORT1, Read 47363 observations 53
Convergence Status
Reason
Convergence criterion (GCONV=1E-8) satisfied.
=====

Running on data set COHORT1, Read 47363 observations 54
Model Fit Statistics

      Criterion      Without      With
      Covariates      Covariates
-2 LOG L      2301.497      2146.860
AIC            2301.497      2350.860
SBC            2301.497      2717.140
=====

Running on data set COHORT1, Read 47363 observations 55
Testing Global Null Hypothesis: BETA=0

      Test      Chi-Square      DF      Pr >
      Chi-Square      Chi-Square
Likelihood Ratio      154.6370      102      0.0006
Score                  152.3984      102      0.0009
Wald                    141.8420      102      0.0056

```

```

=====
Running on data set COHORT1, Read 47363 observations 56
Analysis of Maximum Likelihood Estimates

Label          DF      Parameter      Standard      Hazard
                Estimate      Error      Ratio      lowerCL      upperCL      Pvalue      Parameter
exposure alcohol and cancer 1  1  -0.0007371  0.11743  0.99926  0.79382  1.2579  0.9950  _expND_1_1
exposure alcohol and cancer 2  1   0.44929  0.10814  1.56720  1.26787  1.9372  <.0001  _expND_1_2
exposure alcohol and cancer 3  1   0.30950  0.13467  1.36274  1.04660  1.7744  0.0215  _expND_1_3
ause_p2 and cancer 1          1  -0.11295  0.20992  0.89319  0.59191  1.3478  0.5905  _ucv_1_1
ause_p2 and cancer 2          1  -0.58319  0.21481  0.55811  0.36633  0.8503  0.0066  _ucv_1_2
ause_p2 and cancer 3          1  -0.24737  0.25845  0.78085  0.47051  1.2959  0.3385  _ucv_1_3
... (The rest is omitted)
=====

Running on data set COHORT1, Read 47363 observations 58
Heterogeneity Tests (Likelihood ratio test)

Label          DF      Pvalue
All: alcohol          2  0.01563
Pairwise 1 vs 2: alcohol  1  0.00443
Pairwise 1 vs 3: alcohol  1  0.08233
Pairwise 2 vs 3: alcohol  1  0.41765
=====

```

The titles tell you the name of data set and the number of the observations on which the analysis is conducted. First, the macro tells you the number of events for each subtype and the method of handling ties. Then, you get the results of Cox proportional hazards model. The first table shows Convergence Status, which should be satisfied. The second and third tables show Model Fit Statistics and Testing Global Null Hypothesis, respectively. The table of Analysis of Maximum Likelihood Estimates shows the hazard ratios and confidence intervals of the exposures and covariates, which indicates here the HRs of alcohol for subtype 1, 2 and 3 are 0.999, 1.567 and 1.363, respectively. Note that since the unconstrained model are requested for all covariates, the HRs of covariates for each subtype are shown. Finally, you get the results of heterogeneity test. The rows starting with "All:" and "Pair-wise:" correspond to the results of the overall heterogeneity test across the three subtypes and the pair-wise heterogeneity tests, respectively. Pair-wise 1 vs 2, Pair-wise 1 vs 3, and Pair-wise 2 vs 3 correspond to the comparisons of the effects of alcohol intake between subtype 1 and subtype 2, between subtype 1 and subtype 3 and between subtype 2 and subtype 3, respectively. The data set, heterogeneity, which contains the results of heterogeneity tests is created with using the macro parameter heterotest.

3.2 Example 2. Cohort study analysis with the augmented data set

The data set, cohort2, is the augmented data set for id =1 in cohort1, where the variable censor is a censoring indicator for each subtype which is specified by variable type; it is 1 for censored and 0 if the specific type of cancer is diagnosed in the corresponding block of person-time. The variables alcohol_1, alcohol_2 and alcohol_3 are the subtype-specific exposure variables, which are for subtype 1, 2 and 3, respectively. Note that the data set should have the subtype-specific variables of covariates for which you want to request the unconstrained model, in the same way as the exposure variables.

Cohort2:

id	time	cancer	period	agemo	alcohol	censor	type	alcohol_1	alcohol_2	alcohol_3	OTHER COVARIATES
1	20	0	1	560	0.15	1	1	0.15	0	0	...
1	20	0	1	560	0.15	1	2	0	0.15	0	...
1	20	0	1	560	0.15	1	3	0	0	0.15	...
1	23	0	2	580	0.15	1	1	0.15	0	0	...
1	23	0	2	580	0.15	1	2	0	0.15	0	...
1	23	0	2	580	0.15	1	3	0	0	0.15	...
1	16	1	3	603	0.15	0	1	0.15	0	0	...
1	16	1	3	603	0.15	1	2	0	0.15	0	...
1	16	1	3	603	0.15	1	3	0	0	0.15	...
...											...

The macro call to apply the same model as that used in Example 1 is

```
%subtype(data=cohort2, studydesign=cohort, id=id,
exposure=alcohol_1 alcohol_2 alcohol_3, augmented=yes,
time=time, eventtype=type, censoring=censor,
unconstrvar=ause_p2_1 ause_p2_2 ause_p2_3
screen2_1 screen2_2 screen2_3
polyps2_1 polyps2_2 polyps2_3
cafam2_1 cafam2_2 cafam2_3
py30ct2_1 py30ct2_2 py30ct2_3
py30ct3_1 py30ct3_2 py30ct3_3
py30ct4_1 py30ct4_2 py30ct4_3
py30ct5_1 py30ct5_2 py30ct5_3
py30ctm_1 py30ctm_2 py30ctm_3
actct2_1 actct2_2 actct2_3
actct3_1 actct3_2 actct3_3
```

```

actct4_1 actct4_2 actct4_3
actct5_1 actct5_2 actct5_3
actctm_1 actctm_2 actctm_3
mvit2_1 mvit2_2 mvit2_3
mvitm_1 mvitm_2 mvitm_3
bmain2_1 bmain2_2 bmain2_3
bmain3_1 bmain3_2 bmain3_3
bmain4_1 bmain4_2 bmain4_3
bmi2_1 bmi2_2 bmi2_3
bmi3_1 bmi3_2 bmi3_3
bmi4_1 bmi4_2 bmi4_3
bmi5_1 bmi5_2 bmi5_3
bmim_1 bmim_2 bmim_3
calcq2_1 calcq2_2 calcq2_3
calcq3_1 calcq3_2 calcq3_3
calcq4_1 calcq4_2 calcq4_3
calcq5_1 calcq5_2 calcq5_3
calcqm_1 calcqm_2 calcqm_3
folq2_1 folq2_2 folq2_3
folq3_1 folq3_2 folq3_3
folq4_1 folq4_2 folq4_3
folq5_1 folq5_2 folq5_3,
stratavar=agemo period);

```

The results are the same as those in Example 1.

3.3 Example 3. Nested or matched case-control study analysis

Example 3 use a nested case-control data set, necaco, sampled from the original cohort data set by the risk set sampling with age (years) as time scale and matched on race/ethnicity. There are one cases and two controls in each matching set. The necaco includes the variables matchid which indexes matched set ID.

The macro call is

```

%subtype(data=necaco, studydesign=mcaco, exposure=alcohol,
eventtype=cancer, matchid=matchid,

```

```

constrvar=ause_p2 screen2 polyps2 cafam2
py30ct2 py30ct3 py30ct4 py30ct5 py30ctm
actct2 actct3 actct4 actct5 actctm
mvit2 mvitm
bmain2 bmain3 bmain4
bmi2 bmi3 bmi4 bmi5 bmim
calcq2 calcq3 calcq4 calcq5 calcqm
folq2 folq3 folq4 folq5,
wald=yes
);

```

Note that this macro call requests the constrained models for all covariates and requests Wald test for the heterogeneity test. If you want the unconstrained models for some or all of covariates, those covariates can be placed in the macro parameter unconstrvar.

The output is

```

-----
Running on data set NECACO, Read      268 matched pairs      10
Number of controls and cases in each outcome type
      cancer      Frequency
                   Count
      0           536
      1           99
      2          102
      3           67
-----

Running on data set NECACO, Read      268 matched pairs      11
Convergence Status
Reason
Convergence criterion (GCONV=1E-8) satisfied.
-----

Running on data set NECACO, Read      268 matched pairs      12
Model Fit Statistics
Criterion      Without      With
              Covariates  Covariates
-2 LOG L      588.856    505.805
AIC           588.856    577.805
SBC           588.856    707.081
-----

Running on data set NECACO, Read      268 matched pairs      13
Testing Global Null Hypothesis: BETA=0

```

```

                                Pr >
                                Chi-Square
Test                               Chi-Square    DF
Likelihood Ratio                   83.0512     36    <.0001
Score                               76.8894     36    <.0001
Wald                                65.2835     36    0.0020
=====
Running on data set NECACO, Read      268 matched pairs                                14
Analysis of Maximum Likelihood Estimates
Label                                DF    Parameter Estimate    Standard Error    Hazard Ratio    lowerCL    upperCL    Pvalue    Parameter
exposure alcohol and cancer 1        1    -0.02251              0.14774           0.978           0.73192     1.30613     0.8789    _expND_1_1
exposure alcohol and cancer 2        1     0.35664              0.14972           1.429           1.06524     1.91570     0.0172    _expND_1_2
exposure alcohol and cancer 3        1     0.32872              0.18305           1.389           0.97039     1.98872     0.0725    _expND_1_3
aused_p2                             1    -0.38554              0.17998           0.680           0.47793     0.96774     0.0322
... (The rest is omitted)
=====
Running on data set NECACO, Read      268 matched pairs                                15
Heterogeneity Tests (Likelihood ratio test)
Label                                DF    Pvalue
All: alcohol                         2     0.13649
Pairwise 1 vs 2: alcohol             1     0.06469
Pairwise 1 vs 3: alcohol             1     0.12897
Pairwise 2 vs 3: alcohol             1     0.90352
=====
Running on data set NECACO, Read      268 matched pairs                                16
Heterogeneity Tests (Wald test)
Label                                DF    Pvalue
All: alcohol                         2     0.1390
Pairwise 1 vs 2: alcohol             1     0.0663
Pairwise 1 vs 3: alcohol             1     0.1310
Pairwise 2 vs 3: alcohol             1     0.9036
=====

```

The titles tell you the name of data set and the number of matched pairs on which the analysis is conducted. First, the macro tells you the number of controls and cases for each subtype. Then, you get the results of conditional polytomous logistic regression model. The results are shown in the same way as those in the cohort study analysis. The table of Analysis of Maximum Likelihood Estimates shows the hazard ratios and confidence intervals of the exposures and covariates, which indicates here the HRs of alcohol for subtype 1, 2 and 3 are 0.978, 1.429 and 1.389, respectively. Note that since the constrained model are requested for all covariates, the HRs of covariates for overall colon cancer are shown, assuming the effects of the covariates are the same across the subtypes. Since WALD=yes is specified, you get the results of the heterogeneity test by Wald test, following those by likelihood

ratio test.

3.4 Example 4. Unmatched case-control study analysis

Example 4 analyze the data set used in the Example 3, excluding 3 controls in that data set who were colon cancer cases but in the risk set sampling were sampled as matched controls for ages before the cancer were developed, with adjusting for the matching factors (age and race) by including them as covariates instead of stratified by matcheid. The unconstrained analysis is based on the unconditional nominal polytomous logistic regression model.

The macro call is

```
%subtype(data=necaco, studydesign=caco, exposure=alcohol,
eventtype=cancer,
unconstrvar=ause_p2 screen2 polyps2 cafam2
py30ct2 py30ct3 py30ct4 py30ct5 py30ctm
actct2 actct3 actct4 actct5 actctm
mvit2 mvitm
bmain2 bmain3 bmain4
bmi2 bmi3 bmi4 bmi5 bmim
calcq2 calcq3 calcq4 calcq5 calcqm
folq2 folq3 folq4 folq5
);
```

The output is

```
-----
Running on data set NECACO, Read      801  observations          1
Model: GENERALIZED LOGIT
Number of controls and cases in each outcome type
      cancer      Count
      0          533
      1           99
      2          102
      3           67
-----

Running on data set NECACO, Read      801  observations          2
Convergence Status
Reason
```

Convergence criterion (GCONV=1E-8) satisfied.

Running on data set NECACO, Read 801 observations 3

Model Fit Statistics

Criterion	Intercept Only Model	Model With Intercept and Covariates
AIC	1607.088	1687.278
SC	1621.146	2207.409
-2 Log L	1601.088	1465.278

Running on data set NECACO, Read 801 observations 4

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > Chi-Square
Likelihood Ratio	135.8103	108	0.0363
Score	127.5854	108	0.0961
Wald	113.0353	108	0.3510

Running on data set NECACO, Read 801 observations 5

Type 3 Analysis of Effects

Effect	DF	Wald Chi-square	Pr > Chi-Square
alcohol	3	14.3917	0.0024
ause_p2	3	9.4577	0.0238

... (The rest is omitted)

Running on data set NECACO, Read 801 observations 6

Analysis of Maximum Likelihood Estimates

Variable	outcometype	DF	Estimate	Standard Error	Odds Ratio	lowerCL	upperCL	Pvalue
Intercept	1	1	-0.7457	1.1061	0.47439	0.05428	4.1464	0.5002
Intercept	2	1	-2.5060	1.3155	0.08159	0.00619	1.0750	0.0568
Intercept	3	1	-4.3589	1.8352	0.01279	0.00035	0.4668	0.0175
alcohol	1	1	-0.0422	0.1311	0.95870	0.74150	1.2395	0.7476
alcohol	2	1	0.4382	0.1278	1.54988	1.20641	1.9911	0.0006
alcohol	3	1	0.2660	0.1542	1.30471	0.96443	1.7650	0.0845
ause_p2	1	1	-0.2413	0.2339	0.78564	0.49673	1.2426	0.3023
ause_p2	2	1	-0.7110	0.2444	0.49115	0.30422	0.7929	0.0036
ause_p2	3	1	-0.3766	0.2829	0.68617	0.39410	1.1947	0.1831

... (The rest is omitted)

Running on data set DATASET1, Read 801 observations 8

Heterogeneity Tests (Wald test)

Label	DF	Pvalue
All: alcohol	2	0.0139
Pairwise 1 vs 2: alcohol	1	0.0037
Pairwise 1 vs 3: alcohol	1	0.0988
Pairwise 2 vs 3: alcohol	1	0.3473

The first table shows the number of common controls (533) and subtype specific cancer cases. The results for the association of alcohol intake with high, medium and low LINE-1 colon cancer risk are shown in the table Analysis of Maximum Likelihood Estimates, indicating that odds ratios in unconditional and conditional logistic regression model are 0.96, 1.55 and 1.30, and 0.94, 1.56 and 1.30, respectively. These results suggest that the association of alcohol with LINE-1 tumor risk varies with subtype (p values in unconditional and conditional logistic regression model are 0.014 and 0.023, respectively). Note that, by default, the heterogeneity test was performed using the Wald test in the unconditional nominal polytomous logistic regression model, while the likelihood ratio test was used in the conditional model.

As described above, this approach allow only the unconstrained models for the covariates. A constrained analysis is available with conditional logistic regression model through setting the macro parameter conditional to yes, and place the confounders in the macro parameter constrvar.

The macro call is

```
%subtype(data=necaco, studydesign=caco, exposure=alcohol,
eventtype=cancer, conditional=yes,
constrvar=ause_p2 screen2 polyyps2 cafam2
py30ct2 py30ct3 py30ct4 py30ct5 py30ctm
actct2 actct3 actct4 actct5 actctm
mvit2 mvitm
bmain2 bmain3 bmain4
bmi2 bmi3 bmi4 bmi5 bmim
calcq2 calcq3 calcq4 calcq5 calcqm
folq2 folq3 folq4 folq5,
eventtypelabel =1=high; 2=medium; 3=low
);
```

The main part of the output is

```
-----
Running on data set NECACO, Read      801 observations                               104
Number of controls and cases in each outcome type
CANCER: 1=high; 2=medium; 3=low
Frequency
cancer      Count
```

0	533
1	99
2	102
3	67

```

=====
Running on data set NECACO, Read      801 observations      105
Convergence Status
Reason
Convergence criterion (GCONV=1E-8) satisfied.
=====

```

```

=====
Running on data set NECACO, Read      801 observations      106
Model Fit Statistics
Criterion      Without      With
              Covariates  Covariates
-2 LOG L      1509.867    1399.693
AIC           1509.867    1475.693
SBC           1509.867    1612.151
=====

```

```

=====
Running on data set NECACO, Read      801 observations      107
Testing Global Null Hypothesis: BETA=0
Test           Chi-Square  DF      Pr >
              Chi-Square
Likelihood Ratio  110.1735   38      <.0001
Score            110.3896   38      <.0001
Wald             100.0512   38      <.0001
=====

```

```

=====
Running on data set NECACO, Read      801 observations      108
Analysis of Maximum Likelihood Estimates
Label          DF      Parameter Estimate      Standard Error      Odds Ratio      lowerCL      upperCL      Pvalue      Parameter
exposure alcohol and cancer 1      1      -0.02658      0.12684      0.97377      0.75943      1.24859      0.8340      _expND_1_1
exposure alcohol and cancer 2      1      0.41225      0.12011      1.51021      1.19343      1.91107      0.0006      _expND_1_2
exposure alcohol and cancer 3      1      0.22222      0.14136      1.24884      0.94662      1.64754      0.1160      _expND_1_3
... (The rest is omitted)
=====

```

```

=====
Running on data set NECACO, Read      801 observations      109
Heterogeneity Tests (Likelihood ratio test)
Label          DF      Pvalue
All: alcohol      2      0.03214
Pairwise 1 vs 2: alcohol      1      0.00883
Pairwise 1 vs 3: alcohol      1      0.17575
Pairwise 2 vs 3: alcohol      1      0.28964
=====

```

3.5 Example 5. Case-case study analysis

The example data set consists of all 268 cases from the data set used in Example 1. Unlike the above three study designs, the case-case study allows for testing and estimating of heterogeneity in the exposure associations among subtypes, but cannot estimate the associations of exposures with the risk of each subtype. The Wald test is used for the heterogeneity test.

The data set, caonly is in the standard format, where id, cancer, alcohol and other variables are as described above, and agemo is age in months when the cancer was diagnosed.

```
caonly:
id cancer alcohol agemo Other variables
1 2 0.85 885 ...
2 3 0.85 713 ...
3 1 0 953 ...
...
```

Let the reference level of LINE-1 be the high LINE-1, cancer=1. The macro code that allows the associations of all confounders to be different among subtypes is:

```
%subtype(data=caonly, studydesign=caca, exposure=alcohol,
eventtype=cancer, reftype=1,
unconstrvar=ause_p2 screen2 polyps2 cafam2
py30ct2 py30ct3 py30ct4 py30ct5 py30ctm
actct2 actct3 actct4 actct5 actctm
mvit2 mvitm
bmain2 bmain3 bmain4
bmi2 bmi3 bmi4 bmi5 bmim
calcq2 calcq3 calcq4 calcq5 calcqm
folq2 folq3 folq4 folq5
ageyr,
eventtypelabel = 1 high; 2=medium; 3=low
);
```

The main part of the output is

```

=====
Running on data set CAONLY, Read      268  observations      35
Model: GENERALIZED LOGIT
CANCER: 1=high; 2=medium; 3=low
Number of cases in each outcome type

      cancer      Count
      -----      -----
      1             99
      2            102
      3             67
=====

Running on data set CAONLY, Read      268  observations      36
Convergence Status
Reason
Convergence criterion (GCONV=1E-8) satisfied.
=====

Running on data set CAONLY, Read      268  observations      37
Model Fit Statistics

      Criterion      Intercept      Model With
                   Only Model      Intercept
                   and
                   Covariates
      -----      -----      -----
      AIC            584.012      671.707
      SC             591.194      930.258
      -2 Log L       580.012      527.707
=====

Running on data set CAONLY, Read      268  observations      38
Testing Global Null Hypothesis: BETA=0

      Test            Chi-Square      DF      Pr >
                   Chi-Square
      -----      -----      -----
      Likelihood Ratio      52.3046      70      0.9437
      Score                48.5199      70      0.9765
      Wald                 41.8484      70      0.9970
=====

Running on data set CAONLY, Read      268  observations      39
Type 3 Analysis of Effects

      Effect      DF      Wald      Pr >
                   Chi-square      Chi-Square
      -----      -----      -----
      alcohol      2          8.4864      0.0144
      ause_p2      2          2.2924      0.3178
... (The rest is omitted)
=====

Running on data set CAONLY, Read      268  observations      40
Analysis of Maximum Likelihood Estimates

      Variable      line1      DF      Estimate      Standard      Odds
                   DF      Estimate      Error      Ratio      lowerCL      upperCL      Pvalue
      -----      -----      -----      -----      -----
      Intercept      2          1      -1.4894      1.9294      0.2255      0.00514      9.896      0.4401
      Intercept      3          1      -0.2393      2.0516      0.7872      0.01412      43.901      0.9072
      alcohol        2          1       0.5189      0.1796      1.6802      1.18156      2.389      0.0039
      alcohol        3          1       0.3275      0.1959      1.3874      0.94502      2.037      0.0947
=====

```

```

      ause_p2      2      1      -0.4733      0.3378      0.6229      0.32131      1.208      0.1611
      ause_p2      3      1      -0.0363      0.3652      0.9643      0.47132      1.973      0.9208
... (The rest is omitted)

```

```

Running on data set CAONLY, Read      268 observations

```

42

Heterogeneity Tests (Wald test)

Label	DF	Pvalue
All: alcohol	2	0.0144
Pairwise 1 vs 2: alcohol	1	0.0039
Pairwise 1 vs 3: alcohol	1	0.0947
Pairwise 2 vs 3: alcohol	1	0.3263

The table Heterogeneity Tests (Wald test) shows the results of overall and pair-wise heterogeneity tests in the same way as the other study designs. Pair-wise heterogeneity tests comparing the association of exposure with high LINE-1 to that with medium LINE-1 and low LINE-1 are also provided in the table Analysis of Maximum Likelihood Estimates, since high LINE-1 is the reference group as declared by a macro parameter reftype=1. The respective p-values are $p = 0.0039$ and $p = 0.0947$. Additionally, the result of the overall heterogeneity test is displayed in the table Type 3 Analysis of Effects as $p = 0.0144$. It should be noted that the odds ratios given in this case-case analysis are the ratio of the odds ratio for the alcohol association with each subtype relative to the odds ratio for the alcohol association with reference subtype (i.e., high LINE-1).

Under the assumption of the associations of all confounders to be the same with all subtypes, the macro code can be as follows.

```

%subtype(data=caonly, studydesign=caca, exposure=alcohol,
eventtype=cancer, reftype=1,
constrvar=ause_p2 screen2 polyps2 cafam2
py30ct2 py30ct3 py30ct4 py30ct5 py30ctm
actct2 actct3 actct4 actct5 actctm
mvit2 mvitm
bmain2 bmain3 bmain4
bmi2 bmi3 bmi4 bmi5 bmim
calcq2 calcq3 calcq4 calcq5 calcqm
folq2 folq3 folq4 folq5,
eventtypelabel =1=high; 2=medium; 3=low
);

```

4 Warnings

If the required input is incorrect, the macro will display warnings or errors. For example, if the user specifies `STUDYDESIGN=COHORT` and inputs no variable in `ID` parameter, the macro will display an error as follows.

```
ERROR in macro call:  You did not give a variable name in ID,  
                    as required when you use studydesign=COHORT.
```

If the user specifies `STUDYDESIGN=CACA` and `CONDITIONAL=NO` and gives the variable `age` for a `CONSTRVAR` parameter, the macro will display a warning message as follows.

```
WARNING in macro call:  Your SUBTYPE call have a value for a  
CONSTRVAR parameter,  
but this model does not accept the constrained analysis.  
You may consider using CONDITIONAL=YES option.  
The macro will continue, not adjusting for age.
```

If the data set for a matched case-control study includes the matched sets with only controls or only cases, the macro will display a warning message and exclude those matched sets from the analysis. For example, the warning message below was displayed when `MATCHID=matchid` was specified and the matched sets with `matchid=1` and `16` included only cases.

```
WARNING in macro run:  There are 2 matched sets with control  
or case only  
matchid = 1,16  
will be excluded from a data set used in analysis.
```

5 How should I describe this in my Methods section?

Please refer to the following paper:

Wang M, Spiegelman D, Kuchiba A, Lochhead P, Kim S, Chan AT, Poole EM, Tamimi R, Tworoger SS, Giovannucci E, Rosner B, Ogino S. Statistical methods for studying disease subtype heterogeneity. *Stat Med.* 2016; 35(5): 782-800.

6 Correspondence

Questions should be addressed to Molin Wang via email stmow@channing.harvard.edu.

7 Other reference

Lunn M, McNeil D. Applying Cox regression to competing risks. *Biometrics* 1995;51(2):524-32.