The SAS %RRC Macro

Xiaomei Liao, Samuela Pollack and Donna Spiegelman

December 03, 2012

Abstract:

The macro `%rrc` uses the risk set regression calibration (RRC) method to correct the point and interval estimate of the relative risk in the Cox proportional hazard regression model for bias due to measurement error in one or more baseline or time-varying exposures, including time-varying variables that are functions of the exposure history such as the cumulative average. An external validation study must be available to use this macro. Technical details are given in Liao et al. (2011).

Keywords: SAS, macro, measurement error, Cox proportional hazard model, time-varying covariates, cumulatively updated exposure, simple updated exposure.

Contents

1. Description

The RRC method for the Cox proportional hazard model is designed to use under the following circumstances:

1. Four types of time-varying exposure metrics are supported: cumulatively average updated exposure (`type=1`), cumulatively total updated exposure (`type=2`), simple updated exposure (`type=3`), baseline time-independent exposure (`type=4`).
2. Two kinds of measurement error models are supported for cumulatively average updated exposure (`type=1`): general linear error model (`emodel=1`) and classical additive error model (`emodel=2`). The default model is general linear error model.
3. Outcome is time-to-event (left truncation is supported)
4. All model covariates measured with error are continuous (perfectly measured covariates may be either categorical or continuous).
5. Neither main nor external validation datasets can contain any missing data, but you can use missing indicators as usual.

2. Invocation and details

You need to prepare a SAS data set for the main study and another data set for external validation study. Please specify the variable name for "`id=`", which is the subject identification for both the main study and the validation study. Both the main study and validation study must have the same variable names for covariates with measurement error (for "`surrogate=`"). The main study data set must have a variable that defines the occurrence of the event (for "`case=`"). It is optional to have one or more perfectly measured covariates in both main and validation datasets (for "`confounder=`"). The same confounders with the same variable names must be available in both the main study and the validation study. The parameter "`time=`" is the time-to-event variable at which the covariates are measured till the end of follow-up (i.e. got an event or censored). The main study data is in counting process structure (i.e. one person per multiple records indexed by "`time`" variable, please see the example for the data structure).

The `%rrc` macro works for three types of exposures. If the exposure of interest is the cumulatively average updated exposure, please set parameter `Type=1`. In this scenario, you need to provide the individual exposure measurements at each questionnaire cycle; the program will create the cumulatively updated average internally. For fitting the general linear measurement error, i.e. $c = \alpha_0 + \alpha_1 C + e$, where c is the true exposure and C is the surrogate, please set `Type=1, emodel=1`. For fitting the classical measurement error model, i.e. $C = c + e$, which is not a realistic model for most data in nutritional and environmental epidemiology though, please set `Type=1, emodel=2`. In this case, you need to provide a value for parameter `icc(`$i.e.\ \rho_I\ with\ 0 < \rho_I < 1$`)`, which is the correlation of the repeated measurements of the surrogate exposure; and the value for parameter `dampfactor(`$i.e.\ \theta\ with\ 0 < \theta < 1$`)`, which is the dampening factor of the covariance matrix for surrogate exposure, where we assume that the correlation between repeated measures on the same exposure within a study participant, C_{ij} and $C_{ij'}$ is given by $corr\left(C_{ij,}\, C_{ij'}\right) = {\rho_I}^{\theta|j-j'|}$. If $\theta = 0$, the covariance matrix has the compound symmetry structure; if $\theta = 1$, the covariance matrix has AR(1) structure; if $0 < \theta < 1$, the covariance matrix has the damped exponential structure (DEX) (Alvaro Munoz, V. C., et al. *Biometrics*, 48, 733-742, Sep, 1992).

If the exposure of interest is the cumulatively total updated exposure, please set parameter `Type=2`. This scenario is similar as the cumulatively average updated exposure, but we only consider the general linear measurement error here due to the more generality of this type of measurement error model.

If the exposure of interest is the simple updated exposure, please set parameter `Type=3`. In this scenario, the exposure of interest is each study participant's most recent value of the exposure.

If the exposure of interest is the baseline exposure, which is time-independent, please set parameter `Type=4`. In this case, all model covariates must be baseline only, and the main study has one record per person, as does the validation study. You also need to specify the variable name for "`timestart=`", which is the variable for the time (e.g. age) at the start of follow-up, and the variable name for "`timeend=`", which is the variable for the time at the end of follow-up (i.e. got an event or censored). It means that you need to create these two variables in the main study, however, you don't need to create the variable "`time`" in the main study as in options "`type=1,2 or 3`". Put the same name of "`time`" variable as in the validation study

and the macro will generate this variable in the main study according to "timestart" and "timeend". This way will take care of the left truncation in the analysis for time-independent exposure.

The residual method for controlling for confounding is suggested for studies with two or more perfectly measured covariates, in order to reduce the dimension of the models fit to a reasonable size. If Type=1 and emodel=2 and there are multiple covariates in the analysis, you must set "residual=1" for the dimension reduction (this is the only way that works for this scenario at present; more updates maybe come later). If Type=1 and emodel=1, or Type=2, 3,4, and there are multiple covariates in the analysis, theoretically the program can work for the multi-covariate case in these scenarios. However, it won't give stable results for most of the time, due to the limited number of subjects in the validation study. Thus, setting "residual=1" is highly recommended when the instability happens.

The required and optional arguments for %rrc are summarized as follows:

```
%macro rrc
(
id=,      /* Variable name of subject identification;
             Required */
main=,    /* The name of the main study SAS dataset;
             Required */
validation=,   /* The name of the validation study SAS dataset;
                  Required */
surrogate=,    /* The exposure measured with error in main study and
                  validation study; Required */
true=,     /* The exposure measured without error in validation study;
              Required */
case= ,    /* The variable that indicates whether an event occurs or
              not; Required */
time =,   /* The variable for time-to-event outcome;
             Required */
type=,   /* type of data,
            1: Cumulatively average updated exposure (time-varying);
            2: Cumulatively total updated exposure (time-varying);
            3: Simple updated exposure (time-varying);
            4: Baseline exposure only (time-independent);
           Required */
residual=,   /* if residual=1,use residual method for dimension
                reduction of the models; otherwise, residual=0;
                Required */
emodel=1,   /*if emodel=1(default), it assumes general linear
              measurement error model;
              if emodel=2, it assumes classical additive measurement
              error model. Only Required for type=1 */
confounder= , /* The perfectly measured covariates adjusted in the
                 model; Optional */
groupnum=5,  /* The minimum number of subjects in each risk set for
                validation study; Optional, default=5*/
```

```
increments=1, /* The units in which the RR for true covariate will be
                 reported; Optional, default=1*/
icc=,    /* the correlation of repeated measurements of surrogate
            exposure; Only required for type=1, emodel=2 */
dampfactor=,/* the dampening factor of the covariance matrix for
              surrogate exposure; Only required for type=1, emodel=2*/
timestart=, /* the start of follow-up time, for time-independent
               exposure; Only Required for type=4 */
timeend=,  /* the end of follow-up time(an event or censoring), for
              time-independent exposure; Only Required for type=4 */
filename=RRCOutput.txt  /* filename for the output file; Optional,
                           default=RRCOutput.txt */
);
```

3. Examples

Example 1: Low-carbohydrate diet scores and risk of type II diabetes in the Health Professionals Follow-up Study. Three different analysis have been carried out using `%rrc` macro:

1(a). Analysis for cumulatively updated diet scores;

1(b). Analysis for simple updated diet scores;

1(c). Analysis for diet scores at baseline.

Reference: L. de Koning, T. T Fung, X. M. Liao, S. Chiuve, E. B. Rimm, W. C. Willett, D. Spiegelman, F. B. Hu, *American Journal of Clinical Nutrition*, 2011; 93(4): 844-50.

Three low-carbohydrate diet scores were discussed in this paper --- low-carbohydrate diet with high total protein and fat, low-carbohydrate diet with high animal protein and fat, and low-carbohydrate diet with high vegetable protein and fat. We will use the first one, low-carbohydrate diet with high total protein and fat to illustrate the use of `%rrc` macro.

1(a): Analysis for cumulatively updated diet scores.

First, we are interested in the cumulatively updated low-carbohydrate diet with high total protein and fat. The diet score has been calculated according to Halton TL, et al. *N Engl J Med* , 2006, for both the main study and the validation study. BMI is a potential confounder in the analysis.

The main study dataset can be read in as follows:

```
Data all;
infile
"/udd/stxia/RRCmacro/TestforChanning07312012/hpfsMain2_withBMI_p
oint.dat";
input id ageyr allt2d ttlcarb bmi;
run;
```

And the validation dataset can be read in as follows:

```
Data vali;
infile
"/udd/stxia/RRCmacro/TestforChanning07312012/hpfsVal_dscore_bmi.
dat";
input id ageyr ttlcarbdr ttlcarb bmi;
run;
```

Note: In both the main study and the validation study, the input value for exposure of interest (diet score) is the individual diet score based on the FFQ at each questionnaire cycle or based on DR. Both the exposure and covariates could have some values missing and it's suggested to use "Last Value Carried Forward (LVCF) " to fill in the missing values before doing the analysis. The cumulatively updated exposure at time *t* is the average of the individual diet score up to time *t*.

For example, the following are some data records from the main study dataset:

Obs	id	ageyr	allt2d	ttlcarb	bmi
1	100005	73	0	4	22.6
2	100005	74	0	4	21.9
3	100005	76	0	8	22.4
4	100005	78	0	8	22.4
5	100005	80	0	8	23.2
6	100005	82	0	8	23.2
7	100005	84	0	8	23.2
8	100006	63	0	28	26.5
9	100006	65	0	28	24.8
10	100006	67	0	22	26.5
11	100006	70	0	22	26.5
12	100006	71	0	23	26.5
13	100006	73	0	23	25.8
14	100006	75	0	18	25.8
15	100006	77	0	18	25.1
16	100006	79	0	21	25.3
17	100006	81	0	21	25.8
		. . . . .			

Here are some data records from the validation study dataset:

Obs	id	ageyr	ttlcarbdr	ttlcarb	bmi
1	100965	67	23.0	27	22.3510
2	101463	60	5.5	11	21.6997
3	102750	43	12.0	13	22.2422
4	102781	43	6.0	0	21.8556
5	114258	61	15.0	17	27.9793
6	114408	59	17.5	10	25.5231
7	114722	58	20.0	17	25.8270
8	114776	56	9.5	15	26.4689
9	114974	53	7.5	20	30.5152
10	116073	73	4.5	5	22.8072
11	116217	69	15.0	12	27.2520
12	117761	74	14.0	4	23.4931
13	120133	61	18.0	20	23.3521
14	123647	67	11.5	12	28.6967
15	129563	72	20.0	24	26.4563
		. . . . .			

The general linear measurement error is an appropriate measurement error model for dietary exposures, so this measurement error model is always our first choice. We will first illustrate the use of the macro in the crude analysis, which adjusts only for age in years as the time scale. We call the macro as:

```
%rrc
(
id=id,
main=all,
validation=vali,
surrogate=tlcarb,
true=tlcarbdr,
confounder=,  /* leave it blank */
case=allt2d,
time=ageyr,
groupnum=5,
increments=20, /* difference of 90th and 10th percentile */
type=1,        /* For cumulatively updated exposure */
emodel=1,     /* For the general linear measurement error */
timestart=,     /* not needed here, optional */
timeend=,       /* not needed here, optional */
residual=0,
filename=Outputgen_crude);
```

Then the text file "Outputgen_crude" showed:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
This analysis is for cumulatively average updated exposure
with general linear measurement error model.

The increment for each exposure is: 20.0

There is only one exposure in the analysis and no covariate.
The main exposure is measured with error.

# of subjects in the main study: 41539
# of person-year observations in the main study: 380337
# of cases in the main study: 2790
# of unique failure time: 53

The unique failure times are:
39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
89.0 90.0 92.0

# of subjects in the validation study: 127
# of person-year observations in the validation study: 127
```

```
 # of risk sets in the validation study: 19

 ---------------------------------------------
 Information about the measurement error model
 fitting in the validation study:
 ---------------------------------------------
```

Risk_set	Age_low	Age_up	Person_year	Intercept	slope	R^2
1	39.0	40.0	9	0.425	0.496	0.195
2	41.0	42.0	12	0.427	0.471	0.414
3	43.0	43.0	5	0.367	0.395	0.713
4	44.0	44.0	6	0.394	0.491	0.601
5	45.0	46.0	7	0.515	0.772	0.464
6	47.0	48.0	5	0.228	0.598	0.789
7	49.0	50.0	5	0.718	0.179	0.101
8	51.0	53.0	6	0.839	-0.029	0.000
9	54.0	55.0	8	0.099	0.846	0.963
10	56.0	57.0	6	-0.139	0.959	0.876
11	58.0	59.0	11	0.635	0.314	0.320
12	60.0	61.0	7	0.321	0.479	0.218
13	62.0	62.0	5	0.831	-0.178	0.041
14	63.0	63.0	5	0.087	0.835	0.556
15	64.0	65.0	5	0.493	0.167	0.101
16	66.0	67.0	8	0.500	0.488	0.297
17	68.0	68.0	5	0.062	0.554	0.240
18	69.0	71.0	6	0.816	-0.093	0.007
19	72.0	92.0	6	0.352	0.444	0.585

```
 ---------------------------------------------
 Results for uncorrected estimates (naive cox)
 and corrected estimates (RRC):
 ---------------------------------------------

 Uncorrected result:
 Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
 ttlcarb        0.817   0.062  0.000  2.264  [  2.01,  2.56]
 ---------------------------------------------
 RRC result:
 Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
 ttlcarb        1.509   0.237  0.000  4.523  [  2.84,  7.19]

 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

To control for confounding by BMI, we call the macro as follows:

```
%rrc
(
id=id,
main=all,
validation=vali,
surrogate=tlcarb,
```

```
true=tlcarbdr,
confounder=bmi,  /* BMI is a confounder in this analysis*/
case=allt2d,
time=ageyr,
groupnum=5,
increments=20, /* difference of 90th and 10th percentile */
type=1,        /* For cumulatively updated exposure */
emodel=1,     /* For the general linear measurement error */
timestart=,     /* not needed here, optional */
timeend=,       /* not needed here, optional */
residual=1,     /* For dimension reduction */
filename=Outputgen_BMI);
```

Then the text file "Outputgen_BMI" showed:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
This analysis is for cumulatively average updated exposure
with general linear measurement error model.

The increment for each exposure is: 20.0

The residual analysis has been used in this analysis
for the dimension reduction.

# of subjects in the main study: 41539
# of person-year observations in the main study: 380337
# of cases in the main study: 2790
# of unique failure time: 53

The unique failure times are:
39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
89.0 90.0 92.0

# of subjects in the validation study: 127
# of person-year observations in the validation study: 127
# of risk sets in the validation study: 19


-----------------------------------------------
Information about the measurement error model
fitting in the validation study:
-----------------------------------------------

Risk_set  Age_low  Age_up  Person_year  Intercept  slope     R^2
  1        39.0    40.0         9        -0.000    0.399    0.191
  2        41.0    42.0        12        -0.000    0.404    0.297
  3        43.0    43.0         5         0.000    0.493    0.762
```

```
  4        44.0    44.0         6        0.000    0.582    0.229
  5        45.0    46.0         7        0.000    0.717    0.452
  6        47.0    48.0         5       -0.000    0.608    0.814
  7        49.0    50.0         5       -0.000   -0.921    0.639
  8        51.0    53.0         6        0.000   -0.265    0.057
  9        54.0    55.0         8        0.000    0.788    0.935
 10        56.0    57.0         6       -0.000    1.031    0.921
 11        58.0    59.0        11       -0.000    0.313    0.319
 12        60.0    61.0         7       -0.000    0.468    0.308
 13        62.0    62.0         5        0.000   -0.054    0.004
 14        63.0    63.0         5       -0.000    0.493    0.340
 15        64.0    65.0         5        0.000    0.356    0.431
 16        66.0    67.0         8       -0.000    0.898    0.268
 17        68.0    68.0         5       -0.000    0.445    0.467
 18        69.0    71.0         6       -0.000    0.037    0.001
 19        72.0    92.0         6        0.000    0.509    0.383

 ------------------------------------------------
 Results for uncorrected estimates (naive cox)
 and corrected estimates (RRC):
 ------------------------------------------------

 Uncorrected result:
 Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
 ttlcarb        0.312   0.062  0.000  1.367  [  1.21,  1.54]
 ------------------------------------------------
 RRC result:
 Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
 ttlcarb        0.581   0.258  0.024  1.788  [  1.08,  2.97]

 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Although the classical linear measurement error model is not a realistic model for the dietary exposures, we use it below for the illustration of the use of the macro with `type=1, emodel=2` in the crude analysis. We call the macro as follows:

```
%rrc
(
id=id,
main=all,
validation=vali,
surrogate=tlcarb,
true=tlcarbdr,
confounder=,  /* leave it blank */
case=allt2d,
time=ageyr,
groupnum=5,
increments=20, /* difference of 90th and 10th percentile */
type=1,        /* For cumulatively updated exposure*/
emodel=2,      /* For classical measurement error model */
icc=0.6,       /* correlation for the surrogate exposure */
```

```
dampfactor=0.0, /* try compound symmetry covariance */
residual=0,
filename= Outputadd_crude);
```

Then the text file "Outputadd_crude" showed:

```
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
 This analysis is for cumulatively average updated exposure
 with classical additive measurement error model.

 The increment for each exposure is: 20.0

 There is only one exposure in the analysis and no covariate.
 The main exposure is measured with error.
 It used compound symmetry covariance matrix
 with ICC= 0.6 .

 # of subjects in the main study: 41539
 # of person-year observations in the main study: 380337
 # of cases in the main study: 2790
 # of unique failure time: 53

 The unique failure times are:
 39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
 49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
 59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
 69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
 79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
 89.0 90.0 92.0

 # of subjects in the validation study: 127
 # of person-year observations in the validation study: 127
 # of risk sets in the validation study: 19

 ------------------------------------------------
 Information about the measurement error model
 fitting in the validation study:
 ------------------------------------------------
```

Risk_set	Age_low	Age_up	Person_year	Intercept	slope	R^2
1	39.0	40.0	9	0.425	0.496	0.195
2	41.0	42.0	12	0.427	0.471	0.414
3	43.0	43.0	5	0.367	0.395	0.713
4	44.0	44.0	6	0.394	0.491	0.601
5	45.0	46.0	7	0.515	0.772	0.464
6	47.0	48.0	5	0.228	0.598	0.789
7	49.0	50.0	5	0.718	0.179	0.101
8	51.0	53.0	6	0.839	-0.029	0.000
9	54.0	55.0	8	0.099	0.846	0.963
10	56.0	57.0	6	-0.139	0.959	0.876
11	58.0	59.0	11	0.635	0.314	0.320
12	60.0	61.0	7	0.321	0.479	0.218

```
 13        62.0    62.0         5        0.831   -0.178    0.041
 14        63.0    63.0         5        0.087    0.835    0.556
 15        64.0    65.0         5        0.493    0.167    0.101
 16        66.0    67.0         8        0.500    0.488    0.297
 17        68.0    68.0         5        0.062    0.554    0.240
 18        69.0    71.0         6        0.816   -0.093    0.007
 19        72.0    92.0         6        0.352    0.444    0.585

-----------------------------------------------
Results for uncorrected estimates (naive cox)
and corrected estimates (RRC):
-----------------------------------------------

Uncorrected result:
Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
ttlcarb        0.817   0.062  0.000  2.264  [  2.01,  2.56]
-----------------------------------------------
RRC result:
Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
ttlcarb        1.000   0.082  0.000  2.718  [  2.31,  3.19]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

1(b): Analysis for simple updated diet scores.

Here, we look at the effect of the simple updated low-carbohydrate diet with high total protein and fat. BMI is a potential confounder in the analysis.

We can set up the main study dataset and the validation study dataset exactly the same as in Example 1(a). Then call the macro as:

```
%rrc
(
id=id,
main=all,
validation=vali,
surrogate=tlcarb,
true=tlcarbdr,
confounder=bmi,  /* BMI is a confounder */
case=allt2d,
time=ageyr,
groupnum=5,
increments=20,
type=3,       /* For simply updated exposure */
residual=1,    /* For dimension reduction */
filename=OutputSimple_BMI);
```

The text file "`OutputSimple_BMI`" is produced by the call above to `%RRC`:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
This analysis is for simple updated exposure.

The increment for each exposure is: 20.0

The residual analysis has been used in this analysis
for the dimension reduction.

# of subjects in the main study: 41539
# of person-year observations in the main study: 380337
# of cases in the main study: 2790
# of unique failure time: 53

The unique failure times are:
39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
89.0 90.0 92.0

# of subjects in the validation study: 127
# of person-year observations in the validation study: 127
# of risk sets in the validation study: 19

-----------------------------------------------
Information about the measurement error model
fitting in the validation study:
-----------------------------------------------
```

Risk_set	Age_low	Age_up	Person_year	Intercept	slope	R^2
1	39.0	40.0	9	-0.000	0.399	0.191
2	41.0	42.0	12	-0.000	0.404	0.297
3	43.0	43.0	5	0.000	0.493	0.762
4	44.0	44.0	6	0.000	0.582	0.229
5	45.0	46.0	7	0.000	0.717	0.452
6	47.0	48.0	5	-0.000	0.608	0.814
7	49.0	50.0	5	-0.000	-0.921	0.639
8	51.0	53.0	6	0.000	-0.265	0.057
9	54.0	55.0	8	0.000	0.788	0.935
10	56.0	57.0	6	-0.000	1.031	0.921
11	58.0	59.0	11	-0.000	0.313	0.319
12	60.0	61.0	7	-0.000	0.468	0.308
13	62.0	62.0	5	0.000	-0.054	0.004
14	63.0	63.0	5	-0.000	0.493	0.340
15	64.0	65.0	5	0.000	0.356	0.431
16	66.0	67.0	8	-0.000	0.898	0.268
17	68.0	68.0	5	-0.000	0.445	0.467
18	69.0	71.0	6	-0.000	0.037	0.001
19	72.0	92.0	6	0.000	0.509	0.383

```
-----------------------------------------------
Results for uncorrected estimates (naive cox)
```

```
and corrected estimates (RRC):
-----------------------------------------------

Uncorrected result:
Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
ttlcarb        0.216   0.054  0.000  1.242  [  1.12,  1.38]
-----------------------------------------------
RRC result:
Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
ttlcarb        0.269   0.113  0.017  1.309  [  1.05,  1.63]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

1(c): Analysis for diet scores at baseline.

Here, we examine the effect of low-carbohydrate diet with high total protein and fat at baseline. BMI is the potential confounder in the analysis.

You need to prepare the main study dataset and the validation dataset with time-independent baseline covariates. Then read in the main study dataset as follows:

```
Data all_baseline;
infile "hpfsMain2_withBMI_baseline.dat";
input id allt2d ttlcarb bmi ageyr_start ageyr_end;
run;
```

And the validation dataset can be read in as follows:

```
Data vali;
infile "hpfsVal_dscore_bmi.dat";
input id ageyr ttlcarbdr ttlcarb bmi;
run;
```

Note: The validation study dataset is the same as in Example 1(a) and 1(b). However, the main study dataset is different. There is only one observation per subjects. The exposure "`ttlcarb`" and the confounder "`bmi`" are both baseline measurements. The time variable for event or censoring, "`ageyr`", is included in the validation study dataset, but not in the main study dataset. Instead, the main study dataset has another two variables, "ageyr_start" for the age at the start of follow-up, and "ageyr_end" for the age at the end of follow-up (i.e. got an event or censored).

Then we can call the macro as:

```
%rrc
(
id=id,
main=all_baseline,
validation=vali,
surrogate=tlcarb,
```

```
true=tlcarbdr,
confounder=bmi,  /* BMI is a confounder */
case=allt2d,
time=ageyr,   /* the same name as the time variable in the
                 validation study */
groupnum=5,
increments=20,
type=4,       /* For baseline exposure */
timestart=ageyr_start,
timeend=ageyr_end,
residual=0, /* Not using the residual method for dimension
                 reduction */
filename=Outputbaseline_BMI1);
```

Note: Set `residual=0` to directly adjust for confounding in the model here.

The text file "`Outputbaseline_BMI1`" is produced by the call above to `%RRC`:

```
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
 This analysis is for baseline exposure.

 The increment for each exposure is: 20.0

 There is one exposure and one or more covariates in the analysis.
 The exposure is measured with error, but the covariates are error-
free.

 # of subjects in the main study: 41539
 # of person-year observations in the main study: 711494
 # of cases in the main study: 2790
 # of unique failure time: 53

 The unique failure times are:
 39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
 49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
 59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
 69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
 79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
 89.0 90.0 92.0

 # of subjects in the validation study: 127
 # of person-year observations in the validation study: 127
 # of risk sets in the validation study: 19


 ------------------------------------------------
 Information about the measurement error model
 fitting in the validation study:
 ------------------------------------------------

Risk_set  Age_low  Age_up  Person_year  Intercept  slope[1]     R^2
```

1	39.0	40.0	9	-2.492	0.399	0.195
2	41.0	42.0	12	-0.074	0.404	0.414
3	43.0	43.0	5	0.598	0.493	0.713
4	44.0	44.0	6	0.547	0.582	0.601
5	45.0	46.0	7	1.770	0.717	0.464
6	47.0	48.0	5	0.520	0.608	0.789
7	49.0	50.0	5	-5.648	-0.921	0.101
8	51.0	53.0	6	3.166	-0.265	0.000
9	54.0	55.0	8	-0.083	0.788	0.963
10	56.0	57.0	6	0.590	1.031	0.876
11	58.0	59.0	11	0.685	0.313	0.320
12	60.0	61.0	7	-1.089	0.468	0.218
13	62.0	62.0	5	-0.097	-0.054	0.041
14	63.0	63.0	5	-2.032	0.493	0.556
15	64.0	65.0	5	1.428	0.356	0.101
16	66.0	67.0	8	-1.185	0.898	0.297
17	68.0	68.0	5	-3.453	0.445	0.240
18	69.0	71.0	6	-1.370	0.037	0.007
19	72.0	92.0	6	0.539	0.509	0.585

[1] This slope is for the main exposure --- diet score. Since it's a two-dimensional analysis, there is a slope for both diet score and BMI. Due to format limitations, %RRC only displays the slope for the main exposure.

```
-----------------------------------------------------------
Results for uncorrected and corrected relative risks (RRC):
-----------------------------------------------------------

Uncorrected result:
Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
ttlcarb        0.430   0.055  0.000  1.538  [  1.38,  1.71]
bmi            0.087   0.002  0.000  1.091  [  1.09,  1.10]
-----------------------------------------------
RRC result:
Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
ttlcarb        0.023   0.033  0.479  1.024  [  0.96,  1.09]
bmi            0.089   0.001  0.000  1.093  [  1.09,  1.10]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Note: In this two-dimensional analysis, there are few subjects in each risk set, the regression coefficient estimates are unstable. The corrected estimator is de-attenuated to the null.
We next approached the problem using the residual method to control for confounding by BMI.

Call the macro as:

```
%rrc
(
id=id,
main=all_baseline,
validation=vali,
surrogate=tlcarb,
```

```
true=tlcarbdr,
confounder=bmi,  /* BMI is a confounder */
case=allt2d,
time=ageyr,   /* the same name as the time variable in the
                 validation study */
groupnum=5,
increments=20,
type=4,          /* For baseline exposure */
timestart=ageyr_start,
timeend=ageyr_end,
residual=1,      /* For dimension reduction */
filename=Outputbaseline_BMI2);
```

The text file "Outputbaseline_BMI2" is produced by the call above to %RRC:

```
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
 This analysis is for baseline exposure.

 The increment for each exposure is: 20.0

 The residual analysis has been used in this analysis
 for the dimension reduction.

 # of subjects in the main study: 41539
 # of person-year observations in the main study: 711494
 # of cases in the main study: 2790
 # of unique failure time: 53

 The unique failure times are:
 39.0 40.0 41.0 42.0 43.0 44.0 45.0 46.0 47.0 48.0
 49.0 50.0 51.0 52.0 53.0 54.0 55.0 56.0 57.0 58.0
 59.0 60.0 61.0 62.0 63.0 64.0 65.0 66.0 67.0 68.0
 69.0 70.0 71.0 72.0 73.0 74.0 75.0 76.0 77.0 78.0
 79.0 80.0 81.0 82.0 83.0 84.0 85.0 86.0 87.0 88.0
 89.0 90.0 92.0

 # of subjects in the validation study: 127
 # of person-year observations in the validation study: 127
 # of risk sets in the validation study: 19

 ---------------------------------------------
 Information about the measurement error model
 fitting in the validation study:
 ---------------------------------------------

Risk_set  Age_low  Age_up  Person_year  Intercept  slope     R^2
   1        39.0    40.0          9       -0.000    0.399    0.191
   2        41.0    42.0         12       -0.000    0.404    0.297
   3        43.0    43.0          5        0.000    0.493    0.762
   4        44.0    44.0          6        0.000    0.582    0.229
   5        45.0    46.0          7        0.000    0.717    0.452
```

```
  6        47.0    48.0         5       -0.000    0.608    0.814
  7        49.0    50.0         5       -0.000   -0.921    0.639
  8        51.0    53.0         6        0.000   -0.265    0.057
  9        54.0    55.0         8        0.000    0.788    0.935
 10        56.0    57.0         6       -0.000    1.031    0.921
 11        58.0    59.0        11       -0.000    0.313    0.319
 12        60.0    61.0         7       -0.000    0.468    0.308
 13        62.0    62.0         5        0.000   -0.054    0.004
 14        63.0    63.0         5       -0.000    0.493    0.340
 15        64.0    65.0         5        0.000    0.356    0.431
 16        66.0    67.0         8       -0.000    0.898    0.268
 17        68.0    68.0         5       -0.000    0.445    0.467
 18        69.0    71.0         6       -0.000    0.037    0.001
 19        72.0    92.0         6        0.000    0.509    0.383

-----------------------------------------------
Results for uncorrected estimates (naive cox)
and corrected estimates (RRC):
-----------------------------------------------

Uncorrected result:
Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
ttlcarb        0.217   0.055  0.000  1.243  [  1.12,  1.38]
-----------------------------------------------
RRC result:
Variable      Para_es   S.E.  p_val   H.R.        95% C.I.
ttlcarb        0.263   0.111  0.018  1.301  [  1.05,  1.62]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

4. Warnings

Errors in call to %rrc

```
% rrc
(
id = id,
main= one,
validation= vali,
surrogate= surrogate1 surrogate2 surrogate3,
true= true1 true2,
confounder= ,
case= censoring,
time= timetodeath,
groupnum= 5,
increments= 1 ,
type=1,
emodel=1,
icc=,
```

```
dampfactor=,
timestart=,
timeend=,
residual=0,
filename=XXX
);
```

SAS Log:

```
ERROR IN SAS MACRO RRC:    **********************************

****** The number of true predictors and the number of surrogate predictors
should be equal ******
```

More errors report will be updated later.

References

Xiaomei Liao, David M. Zucker, Yi Li and Donna Spiegelman (2011) Survival analysis with error-prone time-varying covariates: a risk set calibration approach. *Biometrics* 67(1), 50-58, 2011, March.

Credits

Written by Xiaomei Liao, Ph.D, Harvard School of Public Health, Boston, MA; Donna Spiegelman, Sc.D., Harvard School of Public Health, Boston, MA. Thank Samuela Pollack for her programming support.

Questions can be directed to Xiaomei Liao (stxia@channing.harvard.edu), Samuela Pollack (spollack@hsph.harvard.edu) and Donna Spiegelman(stdls@hsph.harvard.edu).