**SAS coding examples for case-cohort designs**

**“Simple” scenario (Table, row 1)** All cases selected, selection probability of sub-cohort = x%

Example: O’Brien et al. (2017) Serum Vitamin D and Risk of Breast Cancer within Five Years. *Environ Health Persp*

cc1 = a data set containing the case-cohort data, including the following variables

subcohort = 1 if in subcohort; 0 if not

case = 1 if a case; 0 if not

age\_enrollment = age at enrollment

age\_eof = age at end of follow-up (e.g. event time or censoring time)

exp = exposure of interest

covar1, covar2, covar3 = covariates of interest (coded as categories)

ID = identification variable

sampling\_rate= number of particants in sub-cohort / number of participants in full eligible cohort

\*test code;

%LET epsilon=0.01; \*or any number smaller than your smallest time unit;

%LET sampling\_rate=0.05; \*for the example data set cc1;

\*restructure data set so that cases in sub-cohort weighted differently according to time (will appear as two entries);

**DATA** ccnew1;

SET wcc.cc1;

\*cases within subcohort - contribute fully until just before diagnosis;

IF subcohort=**1** AND case=**1** THEN DO;

start = age\_enrollment;

stop= age\_eof - &epsilon;

event = **0**; \*considered a censored observation;

wt= **1**/&sampling\_rate; \*inverse probability of sampling weight;

OUTPUT;

END;

\*all cases contribute person-time right before event, count as event;

IF case=**1** THEN DO;

start = age\_eof - &epsilon;

stop = age\_eof;

event = **1**;

wt=**1**;

OUTPUT;

END;

\*non-cases within subcohort - contribute full person time, censored;

ELSE IF subcohort=**1** AND case=**0** THEN DO;

start = age\_enrollment;

stop = age\_eof;

event = **0**;

wt= **1**/&sampling\_rate;

\*inverse probability of sampling weight;

OUTPUT;

END;

**RUN**;

**PROC** **PHREG** DATA=ccnew1 covs(aggregate);

CLASS covar1 covar2 covar3;

MODEL (start,stop)\*event(**0**) = exp covar1 covar2 covar3;

WEIGHT wt;

ID ID;

HAZARDRATIO exp;

**RUN**;

|  |  |  |
| --- | --- | --- |
|  | **Full-cohort,** n=2,983 cases | **Case-cohort,** n=2,983 cases |
| **HR (95% CI)** | 1.14 (1.04, 1.19) | 1.15 (1.00, 1.33) |

**Covariate-stratified case-cohort (Table, row 2)** All cases selected, Sub-cohort selection probabilities of xA% (Group A) and xB% (Group B)

Example: Niehoff et al. (*in review*) Metals and breast cancer risk: a prospective study using toenail biomarkers

cc2 = a data set containing the case-cohort data, including the following variables

subcohort = 1 if in subcohort; 0 if not

case = 1 if a case; 0 if not

age\_enrollment = age at enrollment

age\_eof = age at end of follow-up (e.g. event time or censoring time)

exp = exposure of interest

covar2, covar3 = covariates of interest (coded as categories)

ID = identification variable

groupA=1 if in group A; 0 if in group B

sampling\_rateA= number in sub-cohort from group A / number in full cohort from group A

sampling\_rateB= number in sub-cohort from group B / number in full cohort from group B

%LET sampling\_rateA=0.08; \*for the example data set cc2;

%LET sampling\_rateB=0.15; \*for the example data set cc2;

\*restructure data set so that cases in sub-cohort weighted differently according to time (will appear as two entries);

**DATA** ccnew2;

SET wcc.cc2;

\*cases within subcohort - contribute fully until just before diagnosis;

IF subcohort=**1** AND case=**1** THEN DO;

start = age\_enrollment;

stop= age\_eof - &epsilon;

event = **0**; \*considered a censored observation;

IF groupA=**1** THEN wt= **1**/&sampling\_rateA;

ELSE IF groupA=**0** THEN wt= **1**/&sampling\_rateB;

\*inverse probability of sampling weights;

OUTPUT;

END;

\*all cases contribute person-time right before event, count as event;

IF case=**1** THEN DO;

start = age\_eof - &epsilon;

stop = age\_eof;

event = **1**;

wt=**1**;

OUTPUT;

END;

\*non-cases within subcohort - contribute full person time, c ensored;

ELSE IF subcohort=**1** AND case=**0** THEN DO;

start = age\_enrollment;

stop = age\_eof;

event = **0**;

IF groupA=**1** THEN wt= **1**/&sampling\_rateA;

ELSE IF groupA=**0** THEN wt= **1**/&sampling\_rateB;

\*inverse probability of sampling weights;

OUTPUT;

END;

**RUN**;

**PROC** **PHREG** DATA=ccnew2 covs(aggregate);

CLASS covar2 covar3;

MODEL (start,stop)\*event(**0**) = exp groupA covar2 covar3;

WEIGHT wt;

ID ID;

HAZARDRATIO exp;

**RUN**;

\*group-stratified;

**PROC** **PHREG** DATA=ccnew2 covs(aggregate);

WHERE groupA=**1**;

CLASS covar2 covar3;

MODEL (start,stop)\*event(**0**) = exp covar2 covar3;

WEIGHT wt;

ID ID;

HAZARDRATIO exp;

**RUN**;

**PROC** **PHREG** DATA=ccnew2 covs(aggregate);

WHERE groupA=**0**;

CLASS covar2 covar3;

MODEL (start,stop)\*event(**0**) = exp covar2 covar3;

WEIGHT wt;

ID ID;

HAZARDRATIO exp;

**RUN**;

|  |  |  |
| --- | --- | --- |
|  | **Full-cohort,**  n=2,983 cases  (2,573 Group A,  214 Group B) | **Case-cohort**  n=2,983 cases  (2,573 Group A,  214 Group B) |
| **Overall, HR (95% CI)** | 1.14 (1.04, 1.25) |  |
| *Groups A and B only* | 1.12 (1.02, 1.23) | 1.12 (0.99, 1.27) |
| **Group A, HR (95% CI)** | 1.12 (1.01, 1.23) | 1.12 (0.98, 1.28) |
| **Group B, HR (95% CI)** | 1.11 (0.81, 1.50) | 1.10 (0.76, 1.59) |

**Outcome-stratified case-cohort (Table, row 3)** 100% of type I cases and y% of type 2 cases selected; sub-cohort selection probability x% for all

Example: Sampling 100% of estrogen receptor-negative breast cancers and 50% of estrogen receptor-positive breast cancers, with the desire to look at subtype-specific and overall exposure-disease associations

**Derivation of weights for cases by type**

Need total number of cases of each subtype (i.e., those selected as cases + those in sub-cohort) to weight back to the total number of cases of that subtype in the full cohort (defined as C, with C2=number of cases of type II)

If type 2 cases selected at a probability of y%, there will be **y%\*C2** cases selected as cases

If the sub-cohort is selected with a probability of x%, there will by **x%\*C2** cases selected in the sub-cohort

There will be **y%\*x%\*C2** cases selected into both groups

Therefore, the total number of cases selected is: **y%\*C2 +x%\*C2 - y%\*x%\*C2**

Weight for type 2 cases = C2 / (y%\*C2 +x%\*C2 - y%\*x%\*C2) = 1/ (y%+x% - y%\*x%)

cc3 = a data set containing the case-cohort data, including the following variables

subcohort = 1 if in subcohort; 0 if not

case = 1 if a case; 0 if not

age\_enrollment = age at enrollment

age\_eof = age at end of follow-up (e.g. event time or censoring time)

exp = exposure of interest

covar2, covar3 = covariates of interest (coded as categories)

ID = identification variable

Subtype1=1 if case of disease subtype 1; 0 otherwise

Subtype2=1 if case of disease subtype 2; 0 otherwise

sampling\_rate= number of particants in sub-cohort / number of participants in full eligible cohort

sampling\_rate\_subtype1= number of case of subtype 1 selected / total number of subtype 1 cases

sampling\_rate\_subtype2= number of case of subtype 2 selected / total number of subtype 2 cases

%LET epsilon=0.01; \*or any number less than your smallest time unit;

%LET sampling\_rate=0.05; \*for the example data set cc3;

%LET sampling\_rate\_subtype1=0.50; \*50% of subtype1 selected;

%LET sampling\_rate\_subtype2=1; \*100% of subtype2 selected;

\*restructure data set so that cases in sub-cohort weighted differently according to time (will appear as two entries);

**DATA** ccnew3;

SET wcc.cc3;

\*selected cases within subcohort - contribute fully until just before diagnosis;

IF subcohort=**1** AND (subtype1=**1** | subtype2=**1**) THEN DO;

start = age\_enrollment;

stop= age\_eof - &epsilon;

event = **0**; \*considered a censored observation;

wt= **1**/&sampling\_rate;

\*inverse probability of sampling weight;

OUTPUT;

END;

\*cases contribute person-time right before event only if selected, contribute based on weights;

IF (subtype1=**1** | subtype2=**1**) THEN DO;

start = age\_eof - &epsilon;

stop = age\_eof;

event = **1**;

IF subtype1=**1** THEN

wt=**1**/(&sampling\_rate\_subtype1+&sampling\_rate-

(&sampling\_rate\_subtype1\*&sampling\_rate));

ELSE IF subtype2=**1** THEN

wt=**1**/(&sampling\_rate\_subtype2+&sampling\_rate-

(&sampling\_rate\_subtype2\*&sampling\_rate));

OUTPUT;

END;

\*non-cases within subcohort - contribute full person time, censored;

ELSE IF subcohort=**1** AND subtype1=**0** AND subtype2=**0** THEN DO;

start = age\_enrollment;

stop = age\_eof;

event = **0**;

wt= **1**/&sampling\_rate;

\*inverse probability of sampling weight;

OUTPUT;

END;

**RUN**;

**PROC** **PHREG** DATA=ccnew3 covs(aggregate);

CLASS covar1 covar2 covar3;

MODEL (start,stop)\*event(**0**) = exp covar1 covar2 covar3;

WEIGHT wt;

ID ID;

HAZARDRATIO exp;

**RUN**;

\*subtype 1 only;

**DATA** ccnew3\_1;

SET wcc.cc3;

\*selected cases within subcohort - contribute fully until just before diagnosis;

IF subcohort=**1** AND subtype1=**1** THEN DO;

start = age\_enrollment;

stop= age\_eof - &epsilon;

event = **0**; \*considered a censored observation;

wt= **1**/&sampling\_rate; \*inverse probability of sampling weight;

OUTPUT;

END;

\*cases contribute person-time right before event only if selected, contribute based on weights;

IF subtype1=**1** THEN DO;

start = age\_eof - &epsilon;

stop = age\_eof;

event = **1**;

wt=**1**/(&sampling\_rate\_subtype1+&sampling\_rate-

(&sampling\_rate\_subtype1\*&sampling\_rate));

OUTPUT;

END;

\*non-cases within subcohort - contribute full person time, censored;

ELSE IF subcohort=**1** AND subtype1=**0** THEN DO;

start = age\_enrollment;

stop = age\_eof;

event = **0**;

wt= **1**/&sampling\_rate;

\*inverse probability of sampling weight;

OUTPUT;

END;

**RUN**;

**PROC** **PHREG** DATA=ccnew3\_1 covs(aggregate);

CLASS covar1 covar2 covar3;

MODEL (start,stop)\*event(**0**) = exp covar1 covar2 covar3;

WEIGHT wt;

ID ID;

HAZARDRATIO exp;

**RUN**;

\*subtype 2 only;

**DATA** ccnew3\_2;

SET wcc.cc3;

\*selected cases within subcohort - contribute fully until just before diagnosis;

IF subcohort=**1** AND subtype2=**1** THEN DO;

start = age\_enrollment;

stop= age\_eof - &epsilon;

event = **0**; \*considered a censored observation;

wt= **1**/&sampling\_rate;

\*inverse probability of sampling weight;

OUTPUT;

END;

\*cases contribute person-time right before event only if selected, contribute based on weights;

IF subtype2=**1** THEN DO;

start = age\_eof - &epsilon;

stop = age\_eof;

event = **1**;

wt=**1**/(&sampling\_rate\_subtype2+&sampling\_rate-

(&sampling\_rate\_subtype2\*&sampling\_rate));

OUTPUT;

END;

\*non-cases within subcohort - contribute full person time, censored;

ELSE IF subcohort=**1** AND subtype2=**0** THEN DO;

start = age\_enrollment;

stop = age\_eof;

event = **0**;

wt= **1**/&sampling\_rate;

\*inverse probability of sampling weight;

OUTPUT;

END;

**RUN**;

**PROC** **PHREG** DATA=ccnew3\_2 covs(aggregate);

CLASS covar1 covar2 covar3;

MODEL (start,stop)\*event(**0**) = exp covar1 covar2 covar3;

WEIGHT wt;

ID ID;

HAZARDRATIO exp;

**RUN**;

|  |  |  |
| --- | --- | --- |
|  | **Full-cohort**  n=2,983 cases  (2,197 Subtype 1,  393 Subtype 2) | **Case-cohort**  n=1,349 cases  (956 Subtype 1,  393 Subtype 2) |
| **Overall, HR (95% CI)** | 1.14 (1.04, 1.25) |  |
| *Subtype 1 or 2 only* | 1.11 (1.01, 1.23) | 1.13 (0.96, 1.34) |
| **Subtype 1, HR (95% CI)** | 1.07 (0.96, 1.20) | 1.09 (0.90, 1.31) |
| **Subtype 2, HR (95% CI)** | 1.33 (1.05, 1.68) | 1.35 (1.04, 1.75) |

**Covariate and outcome-stratified case-cohort (Table, row 4)** 100% of type I cases and y% of type 2 cases selected; Sub-cohort selection probabilities of xA% (Group A) and xB% (Group B)

NOTE: This assumes that case status and subgroup status are selected independently; if not, weights can be re-calculated for each combination (= a product of the specified weights)

Example: Oversampling for Black women and estrogen receptor-negative breast cancers

cc4 = a data set containing the case-cohort data, including the following variables

subcohort = 1 if in subcohort; 0 if not

case = 1 if a case; 0 if not

age\_enrollment = age at enrollment

age\_eof = age at end of follow-up (e.g. event time or censoring time)

exp = exposure of interest

covar1, covar2, covar3 = covariates of interest (coded as categories)

ID = identification variable

groupA=1 if in group A; 0 if in group B

Subtype1=1 if case of disease subtype 1; 0 otherwise

Subtype2=1 if case of disease subtype 2; 0 otherwise

sampling\_rateA= number in sub-cohort from group A / number in full cohort from group A

sampling\_rateB= number in sub-cohort from group B / number in full cohort from group B

sampling\_rate\_subtype1= number of case of subtype 1 selected / total number of subtype 1 cases

sampling\_rate\_subtype2= number of case of subtype 2 selected / total number of subtype 2 cases

%LET epsilon=0.01; \*or any number less than your smallest time unit;

%LET sampling\_rateA=0.08; \*for the example data set cc4;

%LET sampling\_rateB=0.15; \*for the example data set cc4;

%LET sampling\_rate\_subtype1=0.50; \*50% of subtype1 selected;

%LET sampling\_rate\_subtype2=1; \*100% of subtype2 selected;

\*restructure data set so that cases in sub-cohort weighted differently according to time (will appear as two entries);

**DATA** ccnew4;

SET wcc.cc4;

\*selected cases within subcohort - contribute fully until just before diagnosis;

IF subcohort=**1** AND (subtype1=**1** | subtype2=**1**) THEN DO;

start = age\_enrollment;

stop= age\_eof - &epsilon;

event = **0**; \*considered a censored observation;

IF groupA=**1** THEN wt= **1**/&sampling\_rateA;

ELSE IF groupA=**0** THEN wt=**1**/&sampling\_rateB;

\*inverse probability of sampling weight;

OUTPUT;

END;

\*cases contribute person-time right before event only if selected, contribute based on weights;

IF (subtype1=**1** | subtype2=**1**) THEN DO;

start = age\_eof - &epsilon;

stop = age\_eof;

event = **1**;

IF subtype1=**1** AND groupA=**1** THEN wt=**1**/(&sampling\_rate\_subtype1+&sampling\_rateA-

(&sampling\_rate\_subtype1\*&sampling\_rateA));

ELSE IF subtype1=**1** AND groupA=**0** THEN

wt=**1**/(&sampling\_rate\_subtype1+&sampling\_rateB-

(&sampling\_rate\_subtype1\*&sampling\_rateB));

ELSE IF subtype2=**1** AND groupA=**1** THEN

wt=**1**/(&sampling\_rate\_subtype2+&sampling\_rateA-

(&sampling\_rate\_subtype2\*&sampling\_rateA));

ELSE IF subtype2=**1** AND groupA=**0** THEN

wt=**1**/(&sampling\_rate\_subtype2+&sampling\_rateB-

(&sampling\_rate\_subtype2\*&sampling\_rateB));

OUTPUT;

END;

\*non-cases within subcohort - contribute full person time, censored;

ELSE IF subcohort=**1** AND subtype1=**0** AND subtype2=**0** THEN DO;

start = age\_enrollment;

stop = age\_eof;

event = **0**;

IF groupA=**1** THEN wt= **1**/&sampling\_rateA;

ELSE IF groupA=**0** THEN wt=**1**/&sampling\_rateB;

\*inverse probability of sampling weight;

OUTPUT;

END;

**RUN**;

**PROC** **PHREG** DATA=ccnew4 covs(aggregate);

CLASS covar2 covar3;

MODEL (start,stop)\*event(**0**) = exp groupA covar2 covar3;

WEIGHT wt;

ID ID;

HAZARDRATIO exp;

**RUN**;

\*subtype-specific estimates computed as for example 3;

|  |  |  |
| --- | --- | --- |
|  | **Full-cohort** | **Case-cohort** |
| **All women** |  |  |
| Overall, HR (95% CI) | 1.14 (1.04, 1.25) |  |
| *Group A or B;*  *Subtype 1 or 2 only* | 1.09 (0.99, 1.21) | 1.05 (0.89, 1.24) |
| Subtype 1, HR (95% CI) | 1.07 (0.96, 1.20) | 1.01 (0.84, 1.22) |
| Subtype 2, HR (95% CI) | 1.33 (1.05, 1.68) | 1.23 (0.94, 1.60) |
| **Group A** |  |  |
| Overall, HR (95% CI) | 1.12 (1.01, 1.23) |  |
| *Subtype 1 or 2 only* | 1.09 (0.98, 1.21) | 1.06 (0.89, 1.25) |
| Subtype 1, HR (95% CI) | 1.07 (0.96, 1.20) | 1.04 (0.86, 1.25) |
| Subtype 2, HR (95% CI) | 1.18 (0.90, 1.54) | 1.15 (0.87, 1.53) |
| **Group B** |  |  |
| Overall, HR (95% CI) | 1.11 (0.81, 1.50) |  |
| *Subtype 1 or 2 only* | 1.10 (0.77, 1.58) | 0.94 (0.53, 1.66) |
| Subtype 1, HR (95% CI) | 0.92 (0.60, 1.40) | 0.68 (0.32, 1.44) |
| Subtype 2, HR (95% CI) | 2.02 (0.99, 4.14) | 1.98 (0.88, 4.47) |

**Case-independent designs (Table, row 5)** v% cases and z% of non-cases included in case-cohort sample; want to measure the association between previously measured exposure (“exp”) and a second exposure (“exp2”), independent of case status

Example: Lawrence et al. (2020) Association of neighborhood deprivation with epigenetic aging using four clock methodologies. *JAMA Open*

Sampling\_rate\_cases= number of selected cases / total number of cases

sampling\_rate\_subcohort= number selected into subcohort / total number in cohort

%LET sampling\_rate\_cases=1; \*for the example data set cc5 (all cases);

%LET sampling\_rate\_subcohort=0.05; \*5% of cohort selected into subcohort;

**DATA** wcc.cc5;

SET wcc.cc5;

IF case=**1** THEN wt= **1**/&sampling\_rate\_cases;

ELSE IF case=**0** THEN wt= **1**/&sampling\_rate\_subcohort;

\*create indicator versions of covariates;

IF covar1=**1** THEN covar1\_1=**1**;

ELSE covar1\_1=**0**;

IF covar1=**2** THEN covar1\_2=**1**;

ELSE covar1\_2=**0**;

IF covar1=**3** THEN covar1\_3=**1**;

ELSE covar1\_3=**0**;

IF covar2=**1** THEN covar2\_1=**1**;

ELSE covar2\_1=**0**;

IF covar2=**2** THEN covar2\_2=**1**;

ELSE covar2\_2=**0**;

IF covar2=**3** THEN covar2\_3=**1**;

ELSE covar2\_3=**0**;

IF covar3=**2** THEN covar3\_2=**1**;

ELSE covar3\_2=**0**;

IF covar3=**3** THEN covar3\_3=**1**;

ELSE covar3\_3=**0**;

**RUN**;

**PROC** **REG** DATA=wcc.cc5;

MODEL exp2 = age\_enrollment covar1\_1 covar1\_2 covar1\_3 covar2\_1 covar2\_2 covar2\_3 covar3\_2 covar3\_3/ CLB;

WEIGHT wt;

**RUN**;

**QUIT**;

|  |  |  |
| --- | --- | --- |
|  | **Full-cohort** | **Case-cohort** |
| ** (95% CI)** | 1.34 (1.21, 1.47) | 1.62 (1.23, 2.02) |