SAS seminar
April 29, 2003

Analysing matched case-control studies using PROC PHREG

Anna Johansson
MEB

Slides also available on www.pauldickman.com
Go to SAS seminars

1

Why matching?

Ex. Prostate cancer

Prostate cancer is a disease among old men.
Age is a known confounder for prostate cancer.

We wish to study some exposure for prostate cancer:

If we do an **unmatched case-control study**, then we adjust for age in the model in order to the correct risk estimates.

But if we instead do a **matched case-control study**, then we will get an even better adjustment for age.

2

Matching is the most efficient way to control for a known confounder

Say we just take a random sample of controls (unmatched design). Then if we have age confounding, we are more likely to have more older cases and more younger controls. Cases will be old, and controls will be evenly spread over the ages.

Then, an **age adjusted OR will not be very efficient** in younger ages where the cases are few and the controls are many.

Likewise, in older ages the controls will be few and cases many.

With matching we keep the proportion between cases and controls constant over ages. The data is used more efficiently and we need fewer observations to obtain the same power as an unmatched design.

3

If the design is a matched design, then the analysis must be an appropriate analysis

Otherwise your results will be biased!

To match is to bias the data on purpose.

But we know how many controls we have chosen from each age stratum. So we have control over the age distribution and can make use of the bias by accounting for it in the analysis.

4

What is an appropriate analysis?

Two options:
1. Conditional logistic regression (model is conditioned on age)
2. Unconditional logistic regression (adjust for age in the model)

	Rothman & Greenland	**General Practice**
Fine Matching (individually)	conditional log.regr.	conditional log.regr.
Frequency Matching	conditional log.regr.	unconditional log.regr.

5

You are always safe by choosing conditional logistic regression!

The rest of this seminar will show you how to do **conditional logistic regression in SAS using PROC PHREG**.

6

Logistic regression using SAS

	Unconditional logistic regr.	Conditional logistic regr.
SAS procedure	LOGISTIC, GENMOD	PHREG
Categorical exposures	CLASS statement + format	Dummy variables

7

SAS PHREG procedure

The PHREG procedure is primarily developed for survival analysis and Cox regression modelling.

But the procedure can also be used for conditional logistic regression, i.e. when analysing matched data.

As it happens, the likelihood function for the Cox model, with events and censored observations, is the same as for a conditional logistic model with matched cases and controls.

8

Ex. Hemoglobin (Hb) levels in mother's blood and risk of stillbirth

A population-based matched case-control study.

We had 702 cases of stillbirth and 702 controls.
Matched on delivery hospital (25 hospitals) and year of delivery (1987-1996). Yielding 25*10 matching strata.

We had many cases in each matching stratum, so we have n:m matching, where number of cases and controls vary.

Total number of cases and controls also varied in all strata.

9

Table of PART_YR by CASE

PART_YR(Year of delivery)
CASE(Case indicator (required by PROC PHREG))

Frequency / Col Pct	Control	Case	Total
1987	68 9.69	68 9.69	136
1988	75 10.68	75 10.68	150
1989	82 11.68	82 11.68	164
...	...	...	...
1996	48 6.84	48 6.84	96
Total	702	702	1404

10

Using the PHREG procedure

This is the code for the PHREG procedure:

```
proc phreg data=olofs.stillbirth;
  model time*case(0) = expcat1 /*expcat2*/ expcat3
                              / ties=discrete
                                risklimits;
  strata sjh part_yr;
  exp: test expcat1=/*expcat2=*/expcat3=0;
run;
```

Here we have assumed a three level categorised exposure variable, using a dummy variable for each level.

11

The MODEL statement

In the MODEL statement you must specify the outcome and the exposures.

```
model time*case(0) = expcat1 /*expcat2*/ expcat3
```

In survival analysis you have two outcome variables:

Event (1=death, 0=censored)
Time (time to event)

The corresponding variables for a matched analysis are:

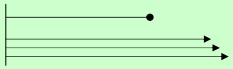
Case (1=case, 0=control)
Time (1=for cases, 2=for controls)

12

All cases in the same stratum should have the same time of event, i.e. time=1. (For practical reasons it is simplest to give all cases the same time regardless of stratum.)

All controls in the same stratum should have the same time, but greater than the cases, i.e. time=2.

The coding is a consequence of that PHREG is designed for survival analysis.



13

So the code

```
time * case(0)
```

gives the information of the cases and controls.

The 0 indicates "censoring" value, or the "control" value of the case variable.

14

```
model time*case(0) = expcat1 /*expcat2*/ expcat3
```

Exposure variables are specified to the right of the "=".

A disadvantage of PHREG is that it does not have CLASS statement like PROC LOGISTIC and PROC GENMOD.

Instead of using the CLASS statement + a format, the exposure variables must be specified with **dummy variables, coded 1 or 0, one dummy per category**.

A structured way to work with dummy variables is to:

- create a dummy for each category (1,0),
- add them all to the model statement, and
- comment out the one you are using as reference group.

15

It is easy to follow which group has been used as reference!

It is easy to change reference without having to create new dummies.

16

Creating dummy variables in the data set

```
data olofs.stillbirth;
  set olofs.iufd;

* Hb : hb value;
* hbcat : categorised hb values;

   if hb=. then hbcat=.;
   else if      hb <= 115 then hbcat=1;
   else if 115< hb <= 125 then hbcat=2;
   else if 125< hb <= 135 then hbcat=3;
   else if 135< hb <= 145 then hbcat=4;
   else if 145< hb        then hbcat=5;
```

17

```
* Create dummies;
* hbcat dummies;

   if hbcat=. then
     do;
       hbcat1=.;
       hbcat2=.;
       hbcat3=.;
       hbcat4=.;
       hbcat5=.;
     end;

   else
     do;
        if hbcat=1 then hbcat1=1;
        else hbcat1=0;

        if hbcat=2 then hbcat2=1;
        else hbcat2=0;
```

18

```
        if hbcat=3 then hbcat3=1;
        else hbcat3=0;

        if hbcat=4 then hbcat4=1;
        else hbcat4=0;

        if hbcat=5 then hbcat5=1;
        else hbcat5=0;
     end;
run;
```

19

Values of the dummies

Obs	HB	HBCAT	HBCAT1	HBCAT2	HBCAT3	HBCAT4	HBCAT5
1	116	2	0	1	0	0	0
2	120	2	0	1	0	0	0
3	126	3	0	0	1	0	0
4	129	3	0	0	1	0	0
5	118	2	0	1	0	0	0
6	133	3	0	0	1	0	0
7	132	3	0	0	1	0	0
8	125	2	0	1	0	0	0
9	141	4	0	0	0	1	0
10	121	2	0	1	0	0	0
11	145	4	0	0	0	1	0

20

This may look like a lot of code, and it is! BUT this code will make it easier to change or collapse categories when we are analysing the data.

We **only need to change the hbcat** variable, and all the dummies will be created "automatically". We need to keep track of which categorisation we are currently using, and what subgroups 1 to 5 stand for. If we for example collapse data into 4 categories, then group 5 is empty due to collapsed data. The dummy5 will be created even though it has a missing value for all. It must be deleted from the MODEL statement.

Using these dummy variables in PHREG

I will choose the middle category, Hb 126-135, as the reference.

```
model time*case(0) = hbcat1 hbcat2 /*hbcat3*/
                     hbcat4 hbcat5
```

21

Options in the MODEL statement

```
model time*case(0) = hbcat1 hbcat2 /*hbcat3*/
                     hbcat4 hbcat5
                           / ties=discrete
                             risklimits;
```

The option **ties=discrete** is needed. The purpose is to replace the proportional hazards model by the discrete logistic model which is needed to get the conditional logistic regression.

The option **risklimits** outputs the confidence intervals for the odds ratios. Default is 95 %.

22

The STRATA statement

```
strata sjh part_yr;
```

The matching variables are specified in the STRATA statement. Here we have matched on delivery hospital (sjh) and year of delivery (part_yr).

It is common to have “individual” matching even when the design is frequency matched.

For practical reasons a control is often chosen individually to a case in the stratum. So we have a variable match_id for the pair. So why not use match_id as the strata variable?

```
strata match_id;
```

23

If we can collapse all cases and controls with the same values on delivery hospital and year, then we gain power.

We have more observations in the stratum.

There is no extra information in the match_id that we don’t have in sjh and part_yr combined.

24

The TEST statement

The TEST statement can be used to create the so-called type 3 tests, i.e. testing if all ORs for an exposure are equal to 1; if the exposure has an over-all effect on the outcome or not.

These tests are Wald tests, an approximation of the likelihood ratio test. To get likelihood ratio tests in PHREG you must fit two models, one with and one without the parameters you wish to test, and compare the log likelihoods by hand.

```
exp: test expcat1=/*expcat2=*/expcat3=0;
```

"exp:" is a label that will show on the output, and you can write anything you like there, just to identify to yourself what variable you are testing.

With Hb as the exposure the test statement would be

```
hb: test hbcat1=hbcat2=/*hbcat3=*/hbcat4=hbcat5=0;
```

So the complete PHREG code is

```
proc phreg data=olofs.stillbirth;
    model time*case(0)= hbcat1 hbcat2 /*hbcat3*/
                        hbcat4 hbcat5
                      / ties=discrete
                        risklimits;
    strata sjh part_yr ;
    hb: test hbcat1=hbcat2=/*hbcat3=*/hbcat4=hbcat5=0;
run;
```

and the output is as follows…

The PHREG Procedure

Model Information

Data Set	OLOFS.STILLBIRTH
Dependent Variable	TIME Survival time (required by PROC PHREG)
Censoring Variable	CASE Case indicator (required by PROC PHREG)
Censoring Value(s)	0
Ties Handling	DISCRETE

Summary of the Number of Event and Censored Values

Stratum	SJH	PART_YR	Total	Event	Censored	Percent Censored
1	BOLLNÄS	1988	2	1	1	50.00
2	BOLLNÄS	1991	2	1	1	50.00
3	BOLLNÄS	1992	2	1	1	50.00
...						
131	NACKA	1987	8	4	4	50.00
132	NACKA	1988	7	4	3	42.86
133	NACKA	1989	4	2	2	50.00
134	NACKA	1990	2	1	1	50.00
135	NACKA	1991	3	1	2	66.67
136	NACKA	1992	1	0	1	100.00
137	NACKA	1993	3	1	2	66.67
138	NACKA	1994	2	1	1	50.00
139	NACKA	1995	6	3	3	50.00
140	NACKA	1996	6	3	3	50.00
...						
207	ÖREBRO	1995	4	2	2	50.00
208	ÖREBRO	1996	2	1	1	50.00
Total			1377	684	693	50.33

The PHREG Procedure

Convergence Status

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

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	1444.091	1432.820
AIC	1444.091	1440.820
SBC	1444.091	1458.932

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.2710	4	0.0237
Score	11.1887	4	0.0245
Wald	11.0107	4	0.0264

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence	Limits
HBCAT1	1	0.56178	0.20872	7.2442	0.0071	1.754	1.165	2.640
HBCAT2	1	0.05658	0.13731	0.1698	0.6803	1.058	0.809	1.385
HBCAT4	1	0.11562	0.14286	0.6550	0.4183	1.123	0.848	1.485
HBCAT5	1	0.60517	0.27787	4.7432	0.0294	1.832	1.062	3.158

Analysis of Maximum Likelihood Estimates

Variable	Variable Label
HBCAT1	First Hb value: <=115
HBCAT2	First Hb value: 116-125
HBCAT4	First Hb value: 136-145
HBCAT5	First Hb value: <=146

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
hb	11.0107	4	0.0264

Summary of the Number of Event and Censored Values

Suppress this list by using option NOSUMMARY in the PROC PHREG statement.

```
proc phreg data=olofs.stillbirth nosummary;
```

Any observation who has missings on any covariate (dummy variable) is excluded automatically.

If percent censored is 0% or 100% then that strata is not used by PHREG in the analyses

If all cases in a stratum (or all controls) have missing on a covariate, which leads to exclusion of all cases, then the number of censored is 100% (or 0% if the controls are excluded due to missingness).

The remaining controls (or cases) will then also be **excluded from the analyses because they are uninformative**. Eg. Nacka 1992.

Be aware that in such case, a **PROC FREQ table won't give correct numbers** even if you exclude those who are missing on the covariate. You must exclude all observations in that **stratum**, also the controls who have a value on the covariate.

This may be tricky. (%subset macro to get correct numbers.)

Analysis of Maximum Likelihood Estimates

The "Hazard Ratio" is the Odds Ratio for conditional logistic regression, remember the PHREG procedure is developed for survival analysis where the hazard ratio is the measure of interest.

Linear Hypothesis Testing Results

Type 3 Wald test p values are in the output. It is a test of over-all homogeneity, testing differences in ORs over categories, I.e. are the ORs equal to 1.

Pr > ChiSq is the p value.

Adding covariates to the model

If you wish to adjust the odds ratios for some other covariate you must create dummies for that variable, if it is a categorical variable, and then add them to the model statement and test statement.

Here I add mother's age (AGEMOM, continuous) and bmi.

```
proc phreg data=olofs.stillbirth nosummary;
    model time*case(0)= hbcat1 hbcat2 /*hbcat3*/ hbcat4 hbcat5
                        agemom
                        bmicat1 /*bmicat2*/ bmicat3 bmicat4
                          / ties=discrete risklimits;
    strata sjh part_yr ;
    hb: test hbcat1=hbcat2=/*hbcat3=*/hbcat4=hbcat5=0;
    age_mother: test agemom=0;
    bmi: test bmicat1=/*bmicat2=*/bmicat3=bmicat4=0;
run;
```


The PHREG Procedure

Model Information

Data Set	WORK.STILLBIRTH
Dependent Variable	TIME
Censoring Variable	CASE
Censoring Value(s)	0
Ties Handling	DISCRETE

Convergence Status

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

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	1402.685	1359.330
AIC	1402.685	1375.330
SBC	1402.685	1411.388

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	43.3555	8	<.0001
Score	42.4804	8	<.0001
Wald	40.8483	8	<.0001

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Variable Label
HBCAT1	1	0.61520	0.21383	8.2775	0.0040	1.850	1.217	2.813	
HBCAT2	1	0.08714	0.14174	0.3779	0.5387	1.091	0.826	1.440	
HBCAT4	1	0.05856	0.14751	0.1576	0.6914	1.060	0.794	1.416	
HBCAT5	1	0.56644	0.28445	3.9654	0.0464	1.762	1.009	3.077	
AGEMOM	1	0.03789	0.01173	10.4314	0.0012	1.039	1.015	1.063	MORALDER
BMICAT1	1	-0.19299	0.17424	1.2268	0.2680	0.824	0.586	1.160	
BMICAT3	1	0.45290	0.15428	8.6174	0.0033	1.573	1.162	2.128	
BMICAT4	1	0.70639	0.23941	8.7054	0.0032	2.027	1.268	3.240	

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
hb	11.4310	4	0.0221
age_mother	10.4314	1	0.0012
bmi	18.8396	3	0.0003

You cannot estimate the effect of the matching variable

You have chosen the distribution of cases and controls to be the same for the matching variable, so the natural difference between them is gone.

If you use unconditional logistic regression and adjust for the matching variable, then the OR should be 1 for the matching variable.

Interaction with the matching variable

However, you can estimate if the matching variable modifies the effect of some other exposure (interaction with the matching variable). This is used in so-called co-twin-control designs, where we match on zygosity.

References

Stephansson O., Dickman P.W., Johansson A., Cnattingius S.; Maternal Hemoglobin Concentration During Pregnancy and Risk of Stillbirth; JAMA 2000; 284:2611-2617

Rothman & Greenland, Modern Epidemiology, 2nd Ed., Lippincott Williams & Wilkins, chapter 10

David Clayton & Michael Hills, Statistical Models in Epidemiology, Oxford University Press, pp178-183

37

Thank you for listening to this seminar

Next seminar May 6

"Data cleaning - tips & tricks"
Paul Dickman

38