GENERALIZED LINEAR MODELS
Matched Case-control studies

Γιώτα Τουλούμη
Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής
Ιατρική Σχολή Πανεπιστημίου Αθήνας

gtouloum@med.uoa.gr

Matched case-control studies

- *What is matching?*
  - Cases are matched with controls according to the levels of one or more *strong* confounders
- *Type of matching*
  - ***Frequency***: Cases and controls have the same (similar) distribution of the confounder; e.g. frequency matching for age: for broad categories of age the same numbers of cases and controls
  - ***One to K***: Each case is matched with K controls for the levels of the confounder; e.g. age and sex matching: for each case with age x a control of the same gender and of the same age (± 1-5 years depending on the availability of controls) is chosen.
- *Choice of controls*
  - Hospital controls, population based controls (for more on the choice and the disadvantages and advantages of each choice see epidemiology course)

Design of case-control studies

- *Why marching?*
  - As a technique for control of confounding, *stratification* may be introduced either at the *design stage* (*matched* case-control studies) or during the *analysis* of results (*unmatched* case-control studies).
- *Gain of matching*
  - With matching greater efficiency is gained by keeping a constant ratio of cases and controls in each stratum of the confounder and thus avoiding inefficiencies resulting from having some strata with a gross imbalance of cases and controls.

Relative efficiency

- One to one pair matching provides the most cost-effective design when cases and controls are equally "scarce"
- When control subjects are more readily obtained than cases (often the case in cancer studies) a 1:M design is more efficient

The theoretical efficiency of a 1:M design for estimating a relative risk of about one, relative to having complete information on the control population (M=∞) is M/(M+1). Thus 1:1 is 50% efficient, 1:4 80%. It is clear that increasing the ratio beyond 5-10 is not worthwhile except if an extreme RR is needed to be estimated.

Analysis of matched case-control studies

- *Frequency matched*
  - As for unmatched case-control studies
    Matching factors included in the model
- 1:M matched
  - Special analysis that takes into account matching. Avoiding it, results in biased results

General rule

Either use individual case-control matching in the design and conditional likelihood (condition on matching) in the analysis **OR** the stratum size for an unconditional analysis should be kept relative large, whether the strata are formed at the design stage or post hoc

Example Data

The study of the exogenous estrogens on the risk of endometrial cancer (Breslow and Day, Statistical methods in cancer research, Volume 1: The analysis of case-control studies). Each case was matched to a 4 control women who were alive and living in the same community (Los Angeles) at the same time the case was diagnosed, who were born within one year of the case, had the same marital status and entered the community at approximately the same time.

Data format

Apart from values for covariates the data should include:

1. An id number, the same for case and controls (identifier of matching)
2. An identifier of case and controls (1 for case, 0 for controls)
3. A counter for the controls (in our case 1-4).

Sample of the data

id	case-control	control No	age	estrogen	dose
1	1	0	74	1	3
1	0	1	75	0	0
1	0	2	74	0	0
1	0	3	74	0	0
1	0	4	75	1	1
2	1	0	65	1	3
2	0	1	67	1	3
2	0	2	67	0	0
2	0	3	67	1	2
2	0	4	68	1	2

Covariate in the data

1. Age (in years)
2. Gall-Bladder disease (Yes:1; No:0)
3. Hypertension (Yes:1; No:0)
4. Obesity (Yes:1; No:0; Unknown: .)
5. Other drugs (non-estrogen) (Yes:1; No:0)
6. Estrogens (Yes:1; No:0)
7. Conjugated estrogen: amount in mg/day (None:0 0.1-0.299:1; 0.3-0.625:2; 0.626+:3; Unknown: .)
8. Conjugated estrogen: duration in months.

Classical analysis of 1:1 matched case-control studies

Data presentation:

		Controls: Exposure +	Controls: Exposure -	Total
Cases	+	a	b	M_1
	-	c	d	M_0
Total		N_1	N_0	N

To analyse 1:1 matched data the paired X^2 (McNemar's) test is used:

We are interested only on the discordant pairs. The rest do not contribute any information:

$$X^2 = \frac{(|b-c|-1)^2}{b+c}$$

Note: The above formula is after the continuity correction.

Under the null hypothesis of no association, the above quantity follows the X^2 with 1 df.

		Controls		
		Exposure		Total
		+	-	
Cases	+	a	b	M_1
	-	c	d	M_0
Total		N_1	N_0	N

Estimated $OR = \hat{y} = \frac{b}{c} = \frac{\text{cases exposed controls not exposed}}{\text{controls exposed cases not exposed}}$

For more details on the rationality of classical analysis of matched case-control studies see Breslow and Day, Statistical methods in cancer research, Volume 1: The analysis of case-control studies, chapter 5.

Analysis using STATA

Lets forget for now the 3 of the 4 controls in the endometrial cancer study. That's the design is 1:1. To do that in stata:

Drop if conno>=1 (deletes controls 2,3,4).

To analyse the data as matched the format should be wide rather than long:

reshape wide age bladder hyper obesity estrogen dose dur nonestr conno , i(id) j(casecon)

(note: j = 0 1)

Data	long	->	wide
Number of obs.	126	->	63
Number of variables	11	->	19
j variable (2 values)	casecon	->	(dropped)

xij variables:

age	->	age0 age1
gall	->	gall0 gall1
hyper	->	hyper0 hyper1
obesity	->	obesity0 obesity1
estrogen	->	estrogen0 estrogen1
dose	->	dose0 dose1
dur	->	dur0 dur1
nonestr	->	nonestr0 nonestr1
conno	->	conno0 conno1

Analysis using STATA (continue)

With the wide version of the data we can use the command mcc (matched case-control):

```
mcc  estrogen1 estrogen0
  Controls
Cases          Exposed  Unexposed      Total

Exposed           27          29          56
Unexposed          3           4           7

Total             30          33          63

McNemar's chi2(1) =     21.13   Prob > chi2 = 0.0000
Exact McNemar significance probability      = 0.0000

Proportion with factor
Cases       .8888889
Controls    .4761905    [95% Conf. Interval]
---------   --------------------
difference  .4126984      .253346   .5720509
ratio       1.866667     1.424262   2.446492
rel. diff.  .7878788     .6331393   .9426183
odds ratio  9.666667     2.996311   49.58254      (exact)
```

It gives the McNemar's test result and the odds ratio (95% CI)

Analysis using STATA (continued)

Alternatively, all the analysis can be done using the commands for the unmatched case-control studies, but using the identifier for case and controls as stratifying variable. For that the data should be in the usual long format.

cc casecon estrogen, by(id)

id	OR	[95% Conf.	Interval]	M-H Weight
1	.	0	.	0 (exact)
2	.	0	.	0 (exact)
3	.	0	.	0 (exact)
4	.	0	.	0 (exact)
5	.	0	.	0 (exact)
.				
.				
.				
61	.	0	.	0 (exact)
62	.	0	.	0 (exact)
63	.	0	.	0 (exact)

Crude	8.8	3.26336	26.00512	(exact)
M-H combined	9.666667	2.944702	31.73307	

Test of homogeneity (B-D) chi2(62) = 37.31 Pr>chi2 = 0.9945

Test that combined OR = 1:

Mantel-Haenszel chi2(1) =21.13

Pr>chi2 = 0.0000

Conditional likelihood

The likelihood which is used in matched case-control studies is not the usual one for the logistic regression. It is the *conditional likelihood*, that is conditional on the fixed values for the marginal totals n_{0i}, n_{1i}, m_{0i}, m_{1i} in each table i where i indicates the i^{th} matched set. That is the analysis follows the same concepts as the stratified analysis.

Conditional likelihood (continue)

Suppose that the i^{th} of I matched sets contains K_i controls in addition to the case. $\mathbf{X}_{io}$ the p-vector of covariates for the case and $\mathbf{X}_{ij}$ the corresponding vector for the j^{th} control (j=1, …, K_i). The conditional likelihood can be written in the form (Liddell, McDonald and Tomas, 1977; Breslow et al., 1978):

$$\prod_{i=1}^{I} \frac{\exp(\sum_{p=1}^{P} \beta_p X_{i0p})}{\sum_{j=0}^{K_i} \exp(\sum_{p=1}^{P} \beta_p X_{ijp})} = \prod_{i=1}^{I} \frac{1}{1+\sum_{j=1}^{K_i} \exp\{\sum_{p=1}^{P} \beta_p (X_{ijp} - X_{iop})\}}$$

It can be seen that the contribution of the matching variates to the likelihood is zero (i.e. the same value for case and control) and the corresponding β cannot be estimated. This means that effects of matching variables cannot be examined. Interactions though with the matching variables can be estimated.

Analysis using conditional logistic regression

Conditional logistic regression can be fitted in STATA using clogit. Data should be in the usual (long) format.

```
clogit casecon estrogen, group(id) or
Iteration 0:   log likelihood =  -38.37664
Iteration 1:   log likelihood = -31.955426
Iteration 2:   log likelihood =   -31.4587
Iteration 3:   log likelihood = -31.443719
Iteration 4:   log likelihood = -31.443696

Conditional (fixed-effects) logistic regression      Number of obs   =126
      LR chi2(1)     =     24.45
      Prob > chi2    =     0.0000
Log likelihood = -31.443696          Pseudo R2      =     0.2799

------------------------------------------------------------------------------
    casecon | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
   estrogen |   9.666667   5.862608     3.74   0.000     2.944712    31.73296
------------------------------------------------------------------------------
```

Results are similar to that from the classical analysis. However, logistic regression is more flexible to analyse matched data, when more than one covariate is going to be analysed. The interpretation of the results is the same as in the unmatched logistic regression. Constant is not reported as now is considered as a nuisance parameter. P value for the OR (Wald test) is similar to that from the M-H test.

Analysis of 1:K matched case-control studies with conditional logistic regression

Conditional logistic regression can be used without any change for any 1:K design. Lets switch to the 1:4 data of endomitrial cancer.

```
clogit  casecon estrogen, group(id) or

Iteration 0:   log likelihood = -96.870519
Iteration 1:   log likelihood = -84.288122
Iteration 2:   log likelihood = -83.728296
Iteration 3:   log likelihood = -83.721592
Iteration 4:   log likelihood =  -83.72159

Conditional (fixed-effects) logistic regression       Number of obs   =315
      LR chi2(1)     =     35.35
      Prob > chi2    =     0.0000
Log likelihood =  -83.72159           Pseudo R2      =     0.1743

casecon  Odds Ratio   Std. Err.      z    P>|z|     [95% Conf.  Interval]
estrogen 7.954681   3.347525     4.93   0.000      3.48671      18.14802
```

Results are similar. Subjects exposed to estrogens are 7.95 times more likely to be cases than controls (95% CI: 3.5 to 18.15) *or* subjects exposed to estrogens have almost 8 times higher risk to develop endometrial cancer than unexposed subjects.

Statistics for testing null hypothesis

In conditional logistic regression the same tests as for unconditional logistic regression can be used:

- Likelihood ratio test
- Wald test
- Score test (for more information on this test see the book of Breslow and Day).

- All tests will give similar results, although with some small differences due to different approximations.

Model checking

The underlying theory for model checking, especially in a 1:M design goes beyond our scope. In general, model checking though leverage, standardized residuals and the rest of diagnostic test becomes more difficult. Especially for 1:1 design simplified formulas have been developed by the extension of Pregibon ideas. For more details on this issue see the book of Hosmer and Lemeshow, applied logistic regression, chapter 7.

Interactions of estrogens with age

While the main effects of age cannot be tested (matched variable) interactions of estrogen with age CAN BE TESTED

```
.  clogit  casecon estrog age32est age33est,group(id)

Iteration 0:   log likelihood = -96.773979
Iteration 1:   log likelihood = -84.029832
Iteration 2:   log likelihood = -83.395607
Iteration 3:   log likelihood = -83.380176
Iteration 4:   log likelihood = -83.380155

Conditional (fixed-effects) logistic regression   Number of obs    =        315
                                                  LR chi2(3)       =      36.03
                                                  Prob > chi2      =     0.0000
Log likelihood = -83.380155                       Pseudo R2        =     0.1777

------------------------------------------------------------------------------
     casecon |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      estrog |   1.430828   .8256894     1.73   0.083    -.1874938    3.049149
    age32est |   .8474007   1.033769     0.82   0.412     -1.17875    2.873551
    age33est |   .7801406    1.15423     0.68   0.499    -1.482108    3.042389
------------------------------------------------------------------------------
```

In this model for women with age 55-64 years the OR is exp(1.430828)=4.182, for women with age 65-74 OR=exp(1.430828+0.8474007)=9.759 and for women with age 75+ years OR=exp(1.430828+0.7801406)=9.125

Test for interaction

Are the differences in the OR's by age group significant?

```
. lrtest,saving(1)

. lrtest, model(0) using(1)
Clogit:  likelihood-ratio test                    chi2(2)    =      0.68
                                                  Prob > chi2 =     0.7107
```

The p=0.71 indicating that the differences by age group ARE not statistically significant. Therefore separate OR's by age groups should not be reported.

Other covariates: Gall-blaster disease

clogit casecon estrogen gall, group(id)

Iteration 0: log likelihood = -95.427631

Iteration 1: log likelihood = -79.81569

Iteration 2: log likelihood = -78.888139

Iteration 3: log likelihood = -78.871318

Iteration 4: log likelihood = -78.871308

Conditional logistic regression Number of obs =315

LR chi2(2) = 45.05

Prob > chi2 = 0.0000

Log likelihood = -78.871308 Pseudo R2 = 0.2221

casecon	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
estrogen	2.114785	.439794	4.81	0.000	1.25280-	2.976765
gall	1.274654	.410868	3.10	0.002	.469368-	2.079941

Other covariates: Gall-blaster disease

- Gall disease is a significant predictor of endometrial cancer:
- OR:exp(1.274654)=3.58. That is, women with Gall disease have 3.58 (95% CI: 1.59 – 8.0) times higher probability (odds) to develop endometrial cancer than women without Gall disease. According to the Wald test: P=0.002.
- The OR for estrogens has not been substantially changed (OR=8.29; 95% CI: 3.50-19.62).

Interactions between estrogens and Gall disease

clogit casecon estrogen gall estgall, group(id)

Iteration 0: log likelihood = -95.292155

Iteration 1: log likelihood = -78.632104

Iteration 2: log likelihood = -76.855555

Iteration 3: log likelihood = -76.7319

Iteration 4: log likelihood = -76.730576

Iteration 5: log likelihood = -76.730576

Conditional logistic regression Number of obs =315

LR chi2(3) = 49.33

Prob > chi2 = 0.0000

Log likelihood = -76.730576 Pseudo R2 = 0.2432

casecon	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
estrogen	2.700139	.6117687	4.41	0.000	1.501094	3.899183
gall	2.894345	.883053	3.28	0.001	1.163593	4.625097
estgall	-2.052747	.9949737	-2.06	0.039	-4.002859	-.1026342

According to the Wald test interaction is significant (P=0.039). **NOTE**: We have **negative interaction** (i.e., the interaction term is negative)

Report of interactions

Estrogens

		Yes	No
Gall	Yes	OR=exp(2.70+2.89-2.05) =14.88x18.67x0.128=34.53	OR=exp(2.89)= =18.07
Disease	No	OR=exp(2.70)= =14.88	1

NOTE: The model suggest that the effects of estrogen use are more likely to be additively combined rather than multiplicatively with those of Gall disease. In other words, in the absence of interactions: effect of using estrogens (OR_1) and having Gall disease OR_2: OR_1*OR_2. Here is more close to $OR_1+OR_2=14.88+18.07=32.95$.

Gall: OR for Estrogens: exp(2.7-2.05)=1.91
Estrogens: OR for Gall: exp(2.89-2.05)=2.32

Finding OR and 95% CI in the presence of significant interactions

clogit casecon estrogen gall estgall, group(id) or

Iteration 0: log likelihood = -95.292155

Iteration 1: log likelihood = -78.632104

Iteration 2: log likelihood = -76.855555

Iteration 3: log likelihood = -76.7319

Iteration 4: log likelihood = -76.730576

Iteration 5: log likelihood = -76.730576

Conditional logistic regression Number of obs =315

LR chi2(3) = 49.33

Prob > chi2 = 0.0000

Log likelihood = -76.730576 Pseudo R2 = 0.2432

casecon	Odds Ratio	Std. Err.	Z	P>z	[95% Conf.	Interval]
estrogen	14.88179	9.104216	4.41	0.000	4.486595	49.36211
gall	18.07166	15.95823	3.28	0.001	3.201415	102.0127
estgall	.1283818	.1277365	-2.06	0.039	.0182633	.902457

lincom estrogen+gall+estgall

(1) estrogen + gall + estgall = 0.0

casecon	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
(1)	3.541737	.7232228	4.90	0.000	2.124246	4.959227

For estrogen and Gall disease: exp(3.54)=34.53. 95% CI: exp(2.12) - exp(4.96)=8.37 – 142.48