Lab session 8
Matched case-control studies
Conditional logistic regression

Open the "endom_ca.dta" file and list the first 20 records. Notice the structure of the dataset and the coding of groups along with the coding of the four controls per group

```
. sort id conno
. li id casecon conno age estro in 1/20

           id       casecon            conno          age       estrog
  1.        1          Case                0           74          Yes
  2.        1       Control                1           75           No
  3.        1       Control                2           74           No
  4.        1       Control                3           74           No
  5.        1       Control                4           75          Yes
  6.        2          Case                0           67          Yes
  7.        2       Control                1           67          Yes
  8.        2       Control                2           67           No
  9.        2       Control                3           67          Yes
 10.        2       Control                4           68          Yes
 11.        3          Case                0           76          Yes
 12.        3       Control                1           76          Yes
 13.        3       Control                2           76          Yes
 14.        3       Control                3           76          Yes
 15.        3       Control                4           77          Yes
 16.        4          Case                0           71          Yes
 17.        4       Control                1           70          Yes
 18.        4       Control                2           70          Yes
 19.        4       Control                3           71          Yes
 20.        4       Control                4           70          Yes
```

This is a 1:4 matched case-control study but for a minute we will pretend that the matching scheme is 1:1. Thus we will keep only the first in each group.

```
. drop if conno>1
(189 observations deleted)

. li id casecon conno age estro in 1/8

           id       casecon            conno          age       estrog
  1.        1          Case                0           74          Yes
  2.        1       Control                1           75           No
  3.        2          Case                0           67          Yes
  4.        2       Control                1           67          Yes
  5.        3          Case                0           76          Yes
  6.        3       Control                1           76          Yes
  7.        4          Case                0           71          Yes
  8.        4       Control                1           70          Yes
```

To analyze this dataset using Mc Nemar's test (paired χ^2) data need to be in wide instead of long format. This can be easily obtained using Stata's reshape command.

```
. reshape wide age gall hyper obesity estrog dose dur nonestr conno, i(id) j(ca
> secon)
(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
                                 estrog   ->   estrog0 estrog1
                                   dose   ->   dose0 dose1
                                    dur   ->   dur0 dur1
                                nonestr   ->   nonestr0 nonestr1
                                  conno   ->   conno0 conno1
-----------------------------------------------------------------------------
```

Now list the first 4 records to check the wide format and compare with the previous listing

```
. li id  estrog* age* in 1/4

          id      estrog0      estrog1         age0         age1
  1.       1           No          Yes           75           74
  2.       2          Yes          Yes           67           67
  3.       3          Yes          Yes           76           76
  4.       4          Yes          Yes           70           71
```

Data are now in a suitable format to perform the Mc Nemar's test using Stata's `mcc` command. We first check the effect of oestrogen use on the risk of endometrial cancer

```
. mcc estrog1 estrog0

                  | Controls               |
Cases             |   Exposed   Unexposed  |     Total
------------------+------------------------+----------
         Exposed |        27          29  |        56
       Unexposed |         3           4  |         7
------------------+------------------------+----------
           Total |        30          33  |        63

McNemar's chi2(1) =     21.12        Pr>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)
```

Similarly we can test for the effect of gall-bladder disease

```
. mcc  gall1 gall0

                 | Controls               |
Cases            |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
         Exposed |         4          13  |        17
       Unexposed |         5          41  |        46
-----------------+------------------------+----------
           Total |         9          54  |        63

McNemar's chi2(1) =      3.56        Pr>chi2 = 0.0593
Exact McNemar significance probability       = 0.0963

Proportion with factor
        Cases       .2698413
        Controls    .1428571     [95% conf. interval]
                  ---------     --------------------
        difference  .1269841      -.017101   .2710693
        ratio       1.888889      .9643767   3.699697
        rel. diff.  .1481481      .0060224   .2902738

        odds ratio       2.6      .8698097   9.315215   (exact)
```

a) How is the Odds Ratio and the χ^2 statistic calculated?

Dividing the pairs according to the age of the case (under 70 – 70 and older) we find

```
. gen age70=age1>=70
```

Under 70 years

```
mcc  gall1 gall0 if age70==0

                 | Controls               |
Cases            |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
         Exposed |         2           7  |         9
       Unexposed |         1          18  |        19
-----------------+------------------------+----------
           Total |         3          25  |        28

McNemar's chi2(1) =      4.50        Pr>chi2 = 0.0339
Exact McNemar significance probability       = 0.0703

Proportion with factor
        Cases       .3214286
        Controls    .1071429     [95% conf. interval]
                  ---------     --------------------
        difference  .2142857     -.0028088   .4313802
        ratio               3      1.032252   8.718799
        rel. diff.        .24      .0466875   .4333125

        odds ratio          7      .8993189    315.599   (exact)
```

70 years and older

```
. mcc  gall1 gall0 if age70==1

                 | Controls               |
Cases            |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
         Exposed |         2           6  |         8
       Unexposed |         4          23  |        27
-----------------+------------------------+----------
           Total |         6          29  |        35

McNemar's chi2(1) =      0.40        Pr>chi2 = 0.5271
Exact McNemar significance probability       = 0.7539

Proportion with factor
        Cases       .2285714
        Controls    .1714286     [95% conf. interval]
                   ---------     --------------------
        difference  .0571429      -.147498   .2617838
        ratio       1.333333      .5450297     3.2618
        rel. diff.  .0689655     -.1372555   .2751866

        odds ratio       1.5      .3557302   7.226552   (exact)
```

To test for the homogeneity of the two OR's in the different age groups we can form the following 2x2 table and calculate the usual χ^2 statistic or most preferably the Fisher's exact significance.

```
. cci 7  6 1 4,exact

                                                      Proportion
                 |   Exposed   Unexposed  |     Total     Exposed
-----------------+------------------------+----------------------
            Cases |         7           6  |        13      0.5385
        Controls |         1           4  |         5      0.2000
-----------------+------------------------+----------------------
           Total |         8          10  |        18      0.4444
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         4.666667       |  .5013338           .   (Cornfield)
 Attr. frac. ex. |         .7857143       | -.9946789           .   (Cornfield)
 Attr. frac. pop |         .4230769       |
                 +-----------------------------------------------
                              1-sided Fisher's exact P = 0.2255
                              2-sided Fisher's exact P = 0.3137
```

Thus there is no evidence for a modifying effect of age on the OR for gall-bladder disease.

If we try to evaluate hypertensive disease as a confounding or modifying factor in similar fashion, we find that there is severe loss of data because of the restriction to case-control pairs, which are homogeneous for hypertension

```
. count
   63
. count if hyper1==hyper0
   32
. di (1-32/63)*100
49.206349
```

Almost half of the 63 pairs will not be used in such a case.
A more flexible way to analyze this sort of datasets is by using the conditonal logistic regression method as implemented in Stata's `clogit` command
In order to use the `clogit` command data should be in the usual long format. The syntax of the `clogit` command is similar to the logit's one except for the required option `group` (or `strata`) to define the grouping variable.

```
. use endom_ca,clear

. drop if conno>1
(189 observations deleted)

. clogit  casecon estrog,group(id)  nolog

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 |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  estrog |   2.268684   .6064767      3.741   0.000       1.080011    3.457356
------------------------------------------------------------------------------
```

Notice that the output does not include an estimate for the models constant (b_0) since this is treated as nuisance parameter.
Adding the option `or` after clogit Stata reports OR's instead of regression coefficients

```
. clogit  casecon estrog,group(id)  nolog or

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]
---------+--------------------------------------------------------------------
  estrog |   9.666667   5.862608      3.741   0.000       2.944712    31.73296
------------------------------------------------------------------------------
```

b) Compare with Mc Nemar's test output

Clogit can accommodate various matching schemes like 1:1, 1:M or even K:M. Moreover this ratio does not have to be constant for all groups in the dataset. Thus from now on we will work with the entire dataset.

```
. use endom_ca,clear

. clogit  casecon estrog,group(id)  nolog or

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     4.928   0.000     3.486714    18.14802
------------------------------------------------------------------------------
. est store A
```

Notice the reduced estimate for the OR along with the reduced value of its Standard error.

c) How is the "LR chi2(1)=35.35" statistic calculated? What is the meaning of this result?

Interactions of estrogens with age

While the main effects of age cannot be tested (matched variable) interactions of estrogen with age CAN BE TESTED
Before we proceed we must create dummy variables for two of the three age groups (1:55-64, 2:65-74, 3:75+) and their interactions with the "estrog" variable. We will use as age for controls the age of the corresponding control pretending that we have exact age matching (in fact differences do not exceed 2 yrs).

```
. gsort id -casecon
. qui by id:gen age32=age[1]>=65 & age[1]<=74
. qui by id:gen age33=age[1]>=75
. gen age32est=age32*estrog
. gen age33est=age33*estrog
```

Now we will fit the model:

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

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 |   1.430827   .8256893     1.73   0.083    -.1874939    3.049149
    age32est |    .847401    1.03377     0.82   0.412    -1.178751    2.873553
    age33est |   .7801409   1.154229     0.68   0.499    -1.482107    3.042389
------------------------------------------------------------------------------

. est store B
```

We can check the significance of the interaction as follows:

```
. lrtest A B

likelihood-ratio test                                 LR chi2(2)  =      0.68
(Assumption: A nested in B)                           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.

Alternative way of doing the same thing

```
gsort id  -casecon
by id:gen case_age=age[1]
egen agegr=cut(case_age),at(0,65,75,200) label
xi i.estrog*i.agegr
clogit casecon  _Iestrog_1 _IestXage_1_1 _IestXage_1_2,group(id)
```

Other covariates: Gall-blaster disease

```
. clogit casecon estrog gall,group(id) nolog

Conditional (fixed-effects) 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]
---------+--------------------------------------------------------------------
  estrog |   2.114785    .439794      4.809   0.000       1.252804    2.976765
    gall |   1.274654   .4108678      3.102   0.002       .4693684    2.079941
------------------------------------------------------------------------------
. est store C
```

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

We can check for interaction between gall-bladder disease and oestrogen use

```
. gen estgall=estrog*gall

. clogit casecon estrog gall estgall,group(id) nolog

Conditional (fixed-effects) 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]
---------+--------------------------------------------------------------------
  estrog |   2.700139   .6117687      4.414   0.000       1.501094    3.899183
    gall |   2.894345    .883053      3.278   0.001       1.163593    4.625097
 estgall |  -2.052747   .9949737     -2.063   0.039      -4.002859   -.1026342
```

```
------------------------------------------------------------------------------
. est store D
```

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

```
. lrtest C D

likelihood-ratio test                                 LR chi2(1)  =      4.28
(Assumption: C nested in D)                           Prob > chi2 =    0.0385
```

The two tests give almost identical results

d) Fill the table below. Are the effects of estrogen use more likely to be additively combined or multiplicatively with those of Gall disease?

Report of interactions

Estrogens

		Yes	No
Gall	Yes		
Disease	No		

We can get the same results using the lincom command and then exponentiating the results

```
. lincom estrog
 ( 1)  estrog = 0.0
------------------------------------------------------------------------------
 casecon |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   2.700139   .6117687      4.414   0.000       1.501094    3.899183
------------------------------------------------------------------------------
. lincom gall
 ( 1)  gall = 0.0
------------------------------------------------------------------------------
 casecon |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   2.894345    .883053      3.278   0.001       1.163593    4.625097
------------------------------------------------------------------------------
. lincom gall+estrog+estgall
 ( 1)  estrog + gall + estgall = 0.0
------------------------------------------------------------------------------
 casecon |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   3.541737   .7232228      4.897   0.000       2.124246    4.959227
------------------------------------------------------------------------------
```

```
. di "[OR (95% CI) | estrog=Yes, gall=No] = ",exp( 2.700139 ),"("
exp(1.501094), ", "exp(3.899183),")"
[OR (95% CI) | estrog=Yes, gall=No] =  14.8818 (4.4865947 , 49.362104 )

. di "[OR (95% CI) | estrog=No, gall =Yes]= ",exp( 2.894345 ),"(" exp(1.163593
) ,", "exp(4.625097 ),")"
[OR (95% CI) | estrog=No, gall=Yes] =  18.071661 (3.2014153 , 102.01267 )

. di "[OR (95% CI) | estrog=Yes, gall=Yes]= ",exp(3.541737 ),"(" exp( 2.124246
 ),", "exp(4.959227 ),")"
```

```
[OR (95% CI) | estrog=Yes, gall=Yes]=  34.52684 (8.3665867 , 142.48361 )
```