Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 16 :

Matched Designs :

Classical Analysis & Conditional Logistic Regression

Matching

Lets consider a study of cancer patients designed to compare 5 year survival rates with 2 treatments: chemotherapy (C) and surgery (S). The patients were grouped into pairs based on characteristics thought to be possible confounders: age, gender and clinical condition. Then, within each pair, one patient receives C and one receives S; this assignment made at random.

Comparing within pairs: Intra-pair comparisons

Then, after 5 years, the survival of each patient is determined. We will assume there was no censoring and no loss to follow up and, further, we will assume that just whether or not a patient has died is of interest.

Then, for each pair, we will know whether the C patient lived or died and whether the S patient lived or died.

For a given pair, age, gender and condition are the same.

For a given pair, if both patients lived, we learn nothing about the difference in survival rates for 2 patients with the same age, gender and condition. Similiarly for a pair in which both patients died. Pairs of these types are called concordant pairs.

If, for a given pair, one patient lived and other patient died, then we have a [so-called] discordant pair.

Discordant pairs

If there is a higher proportion of discordant pairs where S died and C lived compared the pairs where S lived and C died, then we would judge C to have a better survival rate than S.

For every discordant pair, the comparison is based entirely on 2 patients with the same age, gender and condition. In a real way, we have “adjusted” for age, gender and condition as [potential] confounders by a design method as opposed to an analysis method.

We then count up the number of pairs where S lived and C died and compare this number to the number of pairs where S died and C lived

Independence and correlation

It is then assumed that the set of determinations for every discordant pair of patients is like a collection of independent "trials". One trial for each discordant pair.

While, since we matched on age, gender and condition, the 2 outcomes in a given pair are most likely correlated, this correlation has no bearing on the assessment of whether a pair is discordant of one form or the other.

An example

Consider a study where 621 pairs of patients were followed for 5 years.

If we were to incorrectly ignore the matching process in the design , we would get...

...an incorrect unmatched analysis

	Survival		survival rate
	Y	N	
A	106	515	106/621 = 0.171
B	95	526	95/621 = 0.153

p-value: 2 sided Fisher's exact test = 0.441

$\hat{OR}$ = 1.14 ; CI for OR is [0.83, 1.57]

[The odds of survival for those receiving chemotherapy is estimated to be only 1.14 times the odds of survival for those receiving surgery]

...but the correct analysis is to construct:

		Survival for Patient S	
		Y	N
Survival	Y	90	16
for Patient C	N	5	510

The number of [independent] trials is 21.

The estimated odds ratio is 16/5 = 3.2

The p-value is from the binomial distribution.

Correct p-value and correct CI for OR

p-value = 2* P(16 or more successes | n=21 p=0.5)

where p= P(discordant pair has C surviving and S dying)

```
. bitesti 21 16 0.5

      N   Observed k   Expected k   Assumed p   Observed p

      21          16         10.5       0.50000      0.76190

  Pr(k <= 5 or k >= 16) = 0.026604  (two-sided test)


. cii 21 16

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]

             |         21    .7619048    .0929429        .5283402    .9178241

. disp .7619048/(1-.7619048)= 3.20000

. disp .5283402/(1-.5283402)= 1.12017

. disp .9178241/(1-.9178241)= 11.1690
```

$\hat{\mathrm{OR}} = 3.2$ and CI for OR is [1.12, 11.17]

Reshaping a data set

Datasets for studies with matching can be configured 2 distinct ways.

If each row contains the data for an individual, we call the dataset long.

If each row contains the data for a [matched] pair, we call the dataset wide.

Long datasets require a variable to determine the pair while wide datasets require an exposure for each member of the pair.

Long dataset : One row per person

Row	pair	chemo	surv
1	1	0	0
2	1	1	0
3	2	0	1
4	2	1	0
.			
.			
.			
1242	621		

Wide Dataset : One row per matched set

Row	pair	surv0	surv1
1	1	0	0
2	2	1	0
3	3		
.			
.			
.			
621	621		

Choice depends on analyst's requirements

Both wide and long formats contain the same information.

The wide format has data in which each row is a matched set

The long format retains to “usual” approach of having one row for each member of a set.

The long/wide format option is also seen in longitudinal studies, studies with clustering and , indeed, the entire group of study types with repeated outcomes

Switching from long to wide and switching from wide to long

In Stata, if we have a long format file as illustrated, then:

reshape wide surv, i(pair) j(chemo)

changes the dataset to wide format.

i(pair) identifies the pair

j(chemo) instructs Stata to replace the variable surv with surv0 [surv for those with chemo=0] and with surv1 [surv for those with chemo=1]

If the dataset is in long format, we can carry out the incorrect analysis

```
. cc surv chemo,exact

                                                        Proportion
               |   Exposed   Unexposed  |      Total     Exposed
         Cases |       106          95  |        201       0.5274
      Controls |       515         526  |       1041       0.4947
         Total |       621         621  |       1242       0.5000
               |                        |
               |      Point estimate    |    [95% Conf. Interval]
     Odds ratio |         1.139622       |    .8327338    1.560771 (exact)
                                 2-sided Fisher's exact P = 0.4411
```

The mcc command [part of the epitab group of Stata commands] wants the dataset in wide format so...

...we reshape the dataset

```
. reshape wide surv,i(pair) j(chemo)
(note: j = 0 1)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                     1242   ->     621
Number of variables                   3   ->       3
j variable (2 values)             chemo   ->   (dropped)
xij variables:
                                   surv   ->   surv0 surv1
-----------------------------------------------------------------------------
```

surv0 -> survival for those receiving surgery [chemo=0]

surv1 -> survival for those receiving chemo [chemo=1]

...then try the mcc command

```
. mcc surv1 surv0

                 | Controls                |
Cases            |   Exposed   Unexposed   |      Total
-----------------+-------------------------+-----------
         Exposed |        90          16   |        106
       Unexposed |         5         510   |        515
-----------------+-------------------------+-----------
           Total |        95         526   |        621

McNemar's chi2(1) =      5.76    Prob > chi2 = 0.0164 ** chi2 is approx **
Exact McNemar significance probability       = 0.0266 ** exact is Binomial **
** some rows deleted**
        odds ratio       3.2      1.120172   11.16902   (exact)
```


...giving us the correct analysis

Decoding the output from mcc

The table from mcc is generic:

Notice that, here,

those 'exposed' are those that survived

while those 'not exposed' did not survive.

The 'cases' are those receiving chemo

while the 'controls' are those receiving surgery.

Approximation to binomial ?

The normal approximation to the binomial is: $z = \frac{\hat{p} - p}{se(\hat{p})}$

which, for this data, is: $\frac{16/21 - 1/2}{\sqrt{1/2 * 1/2 * 1/21}} \approx 2.4$

$P(|z| > 2.4) \approx 0.0164$

which is the same as: $P(\chi^2_1 > (2.4)^2 = 5.76) \approx 0.0164$

Keep in mind that the exact [and correct] p-value is based on the binomial distribution.

Stata calls this exact p-value the "exact McNemar significance probability"

The so-called McNemar χ^2 is nothing more than the square of the uncorrected normal approximation to the binomial distribution.

A matched case-control study: 4 controls per case

A study of endometrial cancer was designed with 63 sets of 5 participants. Each set of 5 had 1 case [with endometrial cancer] and 4 matched controls [each without this cancer]. Then, the researchers determined the presence or absence of a number of different exposures by looking at records from the past including 'use of estrogens', hypertension, obesity and others...

Incorrect analysis

```
. cc cc est,exact

                                                      Proportion
                 |   Exposed   Unexposed  |     Total     Exposed
-----------------+------------------------+----------------------
           Cases |        56           7  |        63      0.8889
        Controls |       127         125  |       252      0.5040
-----------------+------------------------+----------------------
           Total |       183         132  |       315      0.5810
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         7.874016       |   3.38601    21.15736  (exact)
 Attr. frac. ex. |             .873       |  .7046671    .9527351  (exact)
 Attr. frac. pop |             .776       |
                 +-----------------------------------------------
                               1-sided Fisher's exact P = 0.0000
                               2-sided Fisher's exact P = 0.0000
```

Classical analysis

```
. drop row

. sort quint cc

. by quint: gen otf=_n

. reshape wide  cc age gbd hyp obe est conj dur ned , i(quint) j(otf)
(note: j = 1 2 3 4 5)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                      315   ->      63
Number of variables                  11   ->      46
j variable (5 values)               otf   ->   (dropped)
xij variables:
                                     cc   ->   cc1 cc2 ... cc5
                                    age   ->   age1 age2 ... age5
                                    gbd   ->   gbd1 gbd2 ... gbd5
                                    hyp   ->   hyp1 hyp2 ... hyp5
                                    obe   ->   obe1 obe2 ... obe5
                                    est   ->   est1 est2 ... est5
                                   conj   ->   conj1 conj2 ... conj5
                                    dur   ->   dur1 dur2 ... dur5
                                    ned   ->   ned1 ned2 ... ned5
-----------------------------------------------------------------------------
```

A new table

```
. gen sumcon=est1+est2+est3+est4

. gen sumcas=est5
. table sumcas sumcon
```

sumcas	sumcon 0	1	2	3	4
0		4	1	1	1
1	3	17	16	15	5

There are 5 concordant matched sets. [sumcas=1 and sumcon=4]
Exact p-values are based on the Binomial p= 1/5, 2/5, 3/5 and 4/5

Components to the p-value

```
. bitesti 7 3 .2

         N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
         7           3          1.4     0.20000      0.42857

  Pr(k >= 3) = 0.148032  (one-sided test)
  Pr(k <= 3) = 0.966656  (one-sided test)
  Pr(k >= 3) = 0.148032  (two-sided test)

  note: lower tail of two-sided p-value is empty

. bitesti 18 17 .4

         N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
        18          17          7.2     0.40000      0.94444

  Pr(k >= 17) = 0.000002  (one-sided test)
  Pr(k <= 17) = 1.000000  (one-sided test)
  Pr(k >= 17) = 0.000002  (two-sided test)

  note: lower tail of two-sided p-value is empty

return list
scalars:
                  r(p) =  1.92414534861e-06
```

Next 2 p-values

```
. bitesti 17 16 .6

         N   Observed k    Expected k   Assumed p   Observed p
------------------------------------------------------------
        17           16          10.2         0.6     0.94118

  Pr(k >= 16)            = 0.002088  (one-sided test)
  Pr(k <= 16)            = 0.999831  (one-sided test)
  Pr(k <= 3 or k >= 16)  = 0.002539  (two-sided test)

. bitesti 16 15 .8

         N   Observed k    Expected k   Assumed p   Observed p
------------------------------------------------------------
        16           15          12.8         0.8     0.93750

  Pr(k >= 15)            = 0.140737  (one-sided test)
  Pr(k <= 15)            = 0.971853  (one-sided test)
  Pr(k <= 10 or k >= 15)  = 0.222425  (two-sided test)
```

Correct p-value

TITLE
STB-49 sbe28. Meta-analysis of p values.

DESCRIPTION/AUTHOR(S)
STB insert by Aurelio Tobias, Statistical Consultant, Madrid, Spain.
Support: bledatobias@ctv.es
After installation, see help metap.

INSTALLATION FILES (click here to install)
sbe28/metap.ado
sbe28/metap.hlp

ANCILLARY FILES (click here to get)
sbe28/fleiss.dta

An unweighted combining of p-values

```
. input pvar

         pvar
  1. 0.148032
  2. 1.92414534861e-06
  3. 0.002539
  4. 0.222425
  5. end

. metap pvar

Meta-analysis of p_values

------------------------------------------------------------
 Method              |   chi2          p_value      studies
---------------------+--------------------------------------
 Fisher              |   45.101012     3.521e-07    4
------------------------------------------------------------
```

Estimating the Odds Ratio

The tables:

		Number of Controls Exposed:						
Number		0	1	2	3	4	5	6
of Cases	0		f_{01}	f_{02}	f_{03}	f_{04}	f_{05}	f_{06}
Exposed:	1	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	

f_{0i} is the number of matched sets where the control is exposed and the number exposed is exactly i

f_{1i} is the number of matched sets where the case is exposed and the number exposed is exactly i

If there are M matched controls and exactly i are exposed in the matched set, then

Null Probabilities and Odds:

M	1	2		3			4			
i	1	1	2	1	2	3	1	2	3	4
p	1/2	1/3	2/3	1/4	2/4	3/4	1/5	2/5	3/5	4/5
p/(1-p)	1	1/2	2	1/3	1	3	1/4	2/3	3/2	4

M	5					6					
i	1	2	3	4	5	1	2	3	4	5	6
p	1/6	2/6	3/6	4/6	5/6	1/7	2/7	3/7	4/7	5/7	6/7
p/(1-p)	1/5	2/3	1	3/2	5	1/6	2/5	3/4	4/3	5/2	6

The logarithm of the conditional likelihood

So if p=i/(M+1) and p/(1-p)= i/(M+1-i) = w_i And if we let

$$a_i = 1/(M+1-i)$$

It can then be shown that the log likelihood is:

$$l(\beta) = \sum_{i=1}^{M} \{ f_{1i} \log\left(\frac{a_i e^{\beta}}{(w_i e^{\beta}+1)}\right) + f_{0i} \log\left(\frac{a_i}{(w_i e^{\beta}+1)}\right) \}$$

This looks kinda complicated but the likelihood can be easily graphed in this case (there is only one β , here the log(OR)). See the graph later....

For 4 to 1 matched design we get:

In general, with: $a = e^{\beta}$

$$
\begin{aligned}
l(\beta) = {} & f_{11}\log(a/(a+4)) + f_{01}\log(1/(a+4)) \\
& + f_{12}\log(a/(2a+3)) + f_{02}\log(1/(2a+3)) \\
& + f_{13}\log(a/(3a+2)) + f_{03}\log(1/(3a+2)) \\
& + f_{14}\log(a/(4a+1)) + f_{04}\log(1/(4a+1))
\end{aligned}
$$

...and, for the endometrial cancer study, we get

$$
\begin{aligned}
l(\beta) = {} & 3\log(a/(a+4)) + 4\log(1/(a+4)) \\
& + 17\log(a/(2a+3)) + \log(1/(2a+3)) \\
& + 16\log(a/(3a+2)) + \log(1/(3a+2)) \\
& + 15\log(a/(4a+1)) + \log(1/(4a+1))
\end{aligned}
$$

The equation

The maximum of the conditional likelihood is the solution to this equation

$$\sum \{ \frac{w_i a}{w_i a + 1} (f_{01} + f_{1i}) - f_{1i} \} = 0$$

For a 4 to 1 design, we get:

$$\frac{\frac{1}{4}a}{\frac{1}{4}a+1}(f_{01}+f_{11})-f_{11}+\frac{\frac{2}{3}a}{\frac{2}{3}a+1}(f_{02}+f_{12})-f_{12}+\frac{\frac{3}{2}a}{\frac{3}{2}a+1}(f_{03}+f_{13})-f_{13}+\frac{\frac{4}{1}a}{\frac{4}{1}a+1}(f_{04}+f_{14})-f_{14}=0$$

For the example, we get:

$$\frac{\frac{1}{4}a}{\frac{1}{4}a+1}(7)-3+\frac{\frac{2}{3}a}{\frac{2}{3}a+1}(18)-17+\frac{\frac{3}{2}a}{\frac{3}{2}a+1}(17)-16+\frac{\frac{4}{1}a}{\frac{4}{1}a+1}(16)-15=0$$

How about logistic regression? You might be thinking..

... that you could try a model like:

$$\log(p/(1-p)) = \beta_0 + \sum \alpha_i \delta_i + \beta_1 D$$

where $\sum \alpha_i \delta_i$ is a sum over (all but one of) the 621 matched sets. The δ_i being the indicator for the ith matched set.

It turns out (math excluded!) that there are WAY too many parameters (unknowns) and further, it has been shown (more math!) that the estimate of β_1 is biased.

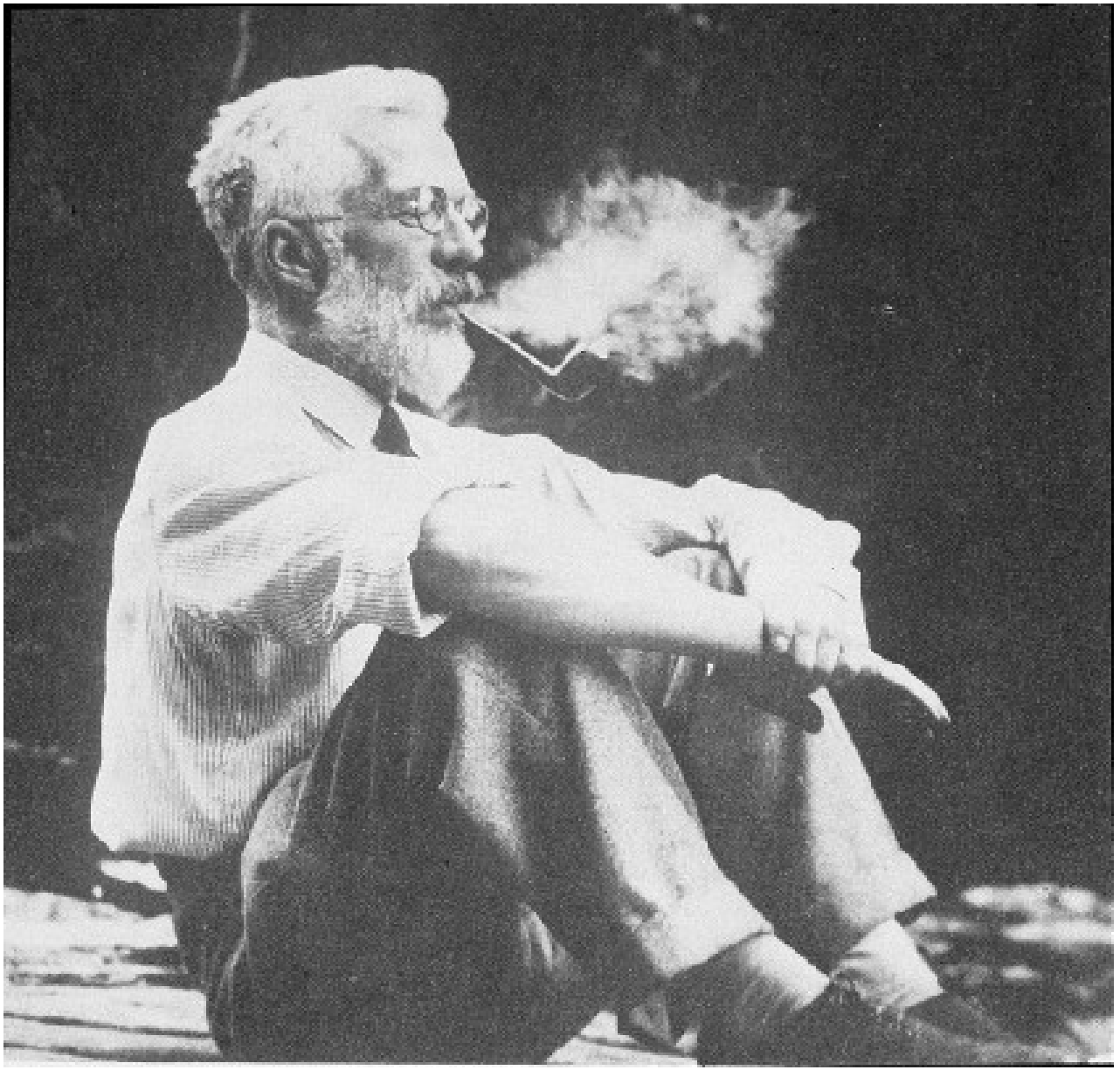
You will recall that, in order to carry out the 'correct analysis', we constructed 'discordant sets' and then made calculations based on conditional probabilities [conditional on being discordant]

A correct model based approach requires a [rather complicated] conditioning process.

We, then, obtain something called a 'conditional' likelihood and the process returns us to the 'familiar' territory in so far as we can then maximize this likelihood, use LR tests, use Wald tests, confidence intervals and much of the related paraphernalia.

Guess who was the innovator of this conditioning argument?

RA Fisher



Conditional Logistic Regression...

.... requires special type of conditional likelihood. A related [but actually quite different] likelihood was developed for survival analysis by DR Cox [1975]. In fact, algorithms for conditional logistic regression were initially developed from the Cox approach. But I digress....

One more digression

At about the same time [mid 1970s], the econometricians [McFadden] were developing a model for an apparently unrelated topic.

The model McFadden developed and applied was the same (!) as the conditional logistic model.

Enough of this history. Lets now see the substance.

Conditional logistic regression version of the correct classical analysis

```
. clogit exp cc,group(pair)
note: multiple positive outcomes within groups encountered.
note: 600 groups (1200 obs) dropped due to all positive or
      all negative outcomes.

Conditional (fixed-effects) logistic regression   Number of obs   =         42
                                                  LR chi2(1)      =       6.06
------------------------------------------------------------------------------
         exp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          cc |   1.163151   .5123475     2.27   0.023     .1589681    2.167334
------------------------------------------------------------------------------

. clogit exp cc,group(pair) or
note: multiple positive outcomes within groups encountered.
note: 600 groups (1200 obs) dropped due to all positive or
      all negative outcomes.

Conditional (fixed-effects) logistic regression   Number of obs   =         42
                                                  LR chi2(1)      =       6.06
------------------------------------------------------------------------------
         exp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          cc |        3.2   1.639512     2.27   0.023     1.172301    8.734961
------------------------------------------------------------------------------
```

P-values / CIs are based on the normal approximation to the binomial.
600 concordant pairs are correctly ‘dropped’

4 matching controls per case

Now let us return to the study of endometrial cancer. We found that, for a crude disease/exposure relationship, the classical analysis provided us with a test, a maximum likelihood estimate but not a direct strategy for analysis beyond the simplest of situations.

More elaborate classical analyses were developed. They are clearly [but technically] explained in Breslow & Day [Volume 1]

Conditional logistic regression now provides all of the analyses.

Conditional logistic regression

```
. clogit est cc,group(quint)
note: multiple positive outcomes within groups encountered.
note: 5 groups (25 obs) dropped due to all positive or
      all negative outcomes.

Conditional (fixed-effects) logistic regression   Number of obs   =        290
                                                  LR chi2(1)      =      35.35
------------------------------------------------------------------------------
         est |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          cc |    2.07376   .4208244     4.93   0.000     1.248959    2.898561
------------------------------------------------------------------------------

. clogit est cc,group(quint) or
note: multiple positive outcomes within groups encountered.
note: 5 groups (25 obs) dropped due to all positive or
      all negative outcomes.

Conditional (fixed-effects) logistic regression   Number of obs   =        290
                                                  LR chi2(1)      =      35.35
                                                  Prob > chi2     =     0.0000
Log likelihood = -99.934552                       Pseudo R2       =     0.1503

------------------------------------------------------------------------------
         est | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          cc |   7.954675   3.347522     4.93   0.000     3.486712      18.148
------------------------------------------------------------------------------
```

The conditional likelihood and the parabolic approximation

In Stata, we can create the actual conditional likelihood and the approximation:

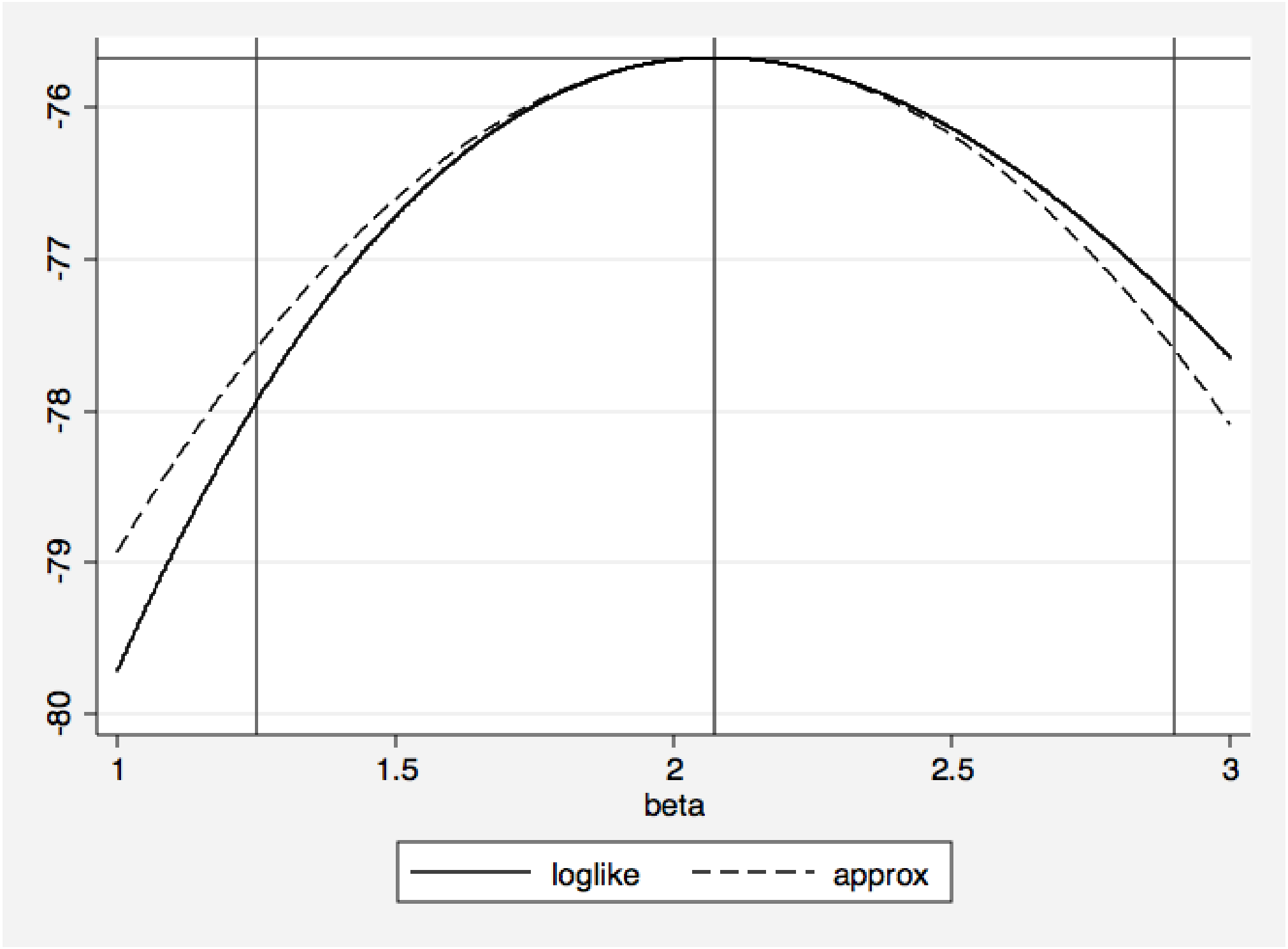
```
. set obs 1001

. range beta 1 3

. gen loglike=3*log(exp(beta)/(exp(beta)+4))+4*log(1/(exp(beta)+4))+17*log(exp(beta)/
    (2*exp(beta)+3))+log(1/(2*exp(beta)+3))+16*log(exp(beta)/(3*exp(beta)+2))+log(1/
    (3*exp(beta)+2))+15*log(exp(beta)/(4*exp(beta)+1))+log(1/(4*exp(beta)+1))

. gen approx= -75.6744- (beta-2.07376)^2 /(2*0.4208244^2)

. twoway (line loglike beta) (line approx beta),xline(2.07376) xline(1.2489442)
    xline(2.8985758) yline(-75.6744) scheme(s2 mono)
```


-76
-77
-78
-79
-80
1
1.5
2
2.5
3
beta
loglike
approx

Assessment of potential confounder

```
. clogit est hyp cc,group(quint) or
note: multiple positive outcomes within groups encountered.
note: 5 groups (25 obs) dropped due to all positive or
      all negative outcomes.

Iteration 0:   log likelihood = -93.816541
Iteration 1:   log likelihood = -93.775297
Iteration 2:   log likelihood = -93.775233
Iteration 3:   log likelihood = -93.775233

Conditional (fixed-effects) logistic regression   Number of obs   =        290
                                                  LR chi2(2)      =      47.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -93.775233                       Pseudo R2       =     0.2026

------------------------------------------------------------------------------
         est | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hyp |   3.175014   1.099597     3.34   0.001     1.610462    6.259518
          cc |   7.423919   3.149142     4.73   0.000     3.232684    17.04917
------------------------------------------------------------------------------
```

Assessment of age as a potential modifier (even though age was a part of the matching criteria)

```
. gen ac=age*cc

. clogit est cc hyp ac,group(quint)
note: multiple positive outcomes within groups encountered.
note: 5 groups (25 obs) dropped because of all positive or
      all negative outcomes.

Conditional (fixed-effects) logistic regression

                                                Number of obs     =        290
                                                LR chi2(3)        =      47.67
                                                Prob > chi2       =     0.0000
Log likelihood = -93.774338                     Pseudo R2         =     0.2027

------------------------------------------------------------------------------
         est |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          cc |   1.805899   4.715429     0.38   0.702    -7.436171    11.04797
         hyp |   1.154741   .3466582     3.33   0.001     .4753032    1.834178
          ac |   .0027714   .0655007     0.04   0.966    -.1256076    .1311505
------------------------------------------------------------------------------
```

Notice…

…that age*cc is included in the model even though age is not included.

This is one of the special cases where we CAN interpret a model with a 'product' term even though one of the constituents of this product is not included in the model.

Matching enables a design-based way to address confounding, but not modification.