Pooled Logistic Regression; Logistic Regression for Case-Control Studies

March 17, 2020

Biostatistics 208

Pooled logistic regression for grouped event time outcomes

- Application of the logistic regression model to survival (event time) outcomes
- For studies where observations of event times are limited to regular intervals (e.g. visit intervals in prospective studies)
- Data structured in "long" form, with one observation per person-visit, including a binary indicator y of event occurrence (at most once / individual), and time t
- Coefficients approximate log relative hazards from Cox proportional hazards regression when outcomes are rare

- Advantages over Cox regression:
  - Can be applied to when info about event times limited to visit intervals
  - Accepts observation weights - used to fit *marginal structural models* for causal inference accounting for time-dependent confounding variables

- Disadvantages:
  - requires explicit model for the time dependence of outcome risk
  - works best for regularly-spaced intervals with complete data
  - often requires modeling assumptions about event occurrences and predictor values within observation intervals

Pooled logistic regression for grouped event time outcomes

- Time dependence of outcome risk is modeled by the intercept part of the model:

$$\log\left(\frac{P}{1-P}\right) = \beta_0(t) + \beta_1 x_1 + \ldots + \beta_p x_p$$

- $\beta_0(t)$ represents the *baseline* log outcome odds at time t, among individuals with all predictors set to zero

  - No time dependence: constant $\beta_0(t) = \beta_0$,
  - including time (t) as an untransformed predictor equivalent to assuming linearity of log outcome odds with increasing time
  - flexible time dependence can be modeled by a restricted cubic spline for t

- The pooled logistic is a proportional odds model

  - Approximates the proportional hazards model for rare events
  - Used for control of predictors acting as time-dependent confounder/mediators in causal inference (including inverse probability weights & robust SEs) (Biostat 215)

References:

1. D'Agostino RB, Lee ML, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. Stat Med. 1990;9:1501-15.
2. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology. 2000;11:561-70.

Pooled logistic regression for grouped event time outcomes

- Example: incident HSV-2 infection in the MIRA study

  listing of data for two individuals (one with incident HSV-2):

```
. list id hsv2 mos agecat stihx newparts if id==28 | id==27
     +-----------------------------------------------+
     | id   hsv2   mos   agecat   stihx   newparts |
     |-----------------------------------------------|
202. | 27      0     3         1       0           0 |
203. | 27      0     6         1       0           0 |
204. | 27      0     9         1       0           0 |
205. | 27      0    12         1       0           0 |
206. | 27      0    15         1       0           0 |
     |-----------------------------------------------|
207. | 27      0    18         1       0           0 |
208. | 27      0    21         1       0           0 |
209. | 27      0    24         1       0           0 |
210. | 28      0     3         1       1           0 |
211. | 28      0     6         1       1           0 |
     |-----------------------------------------------|
212. | 28      0     9         1       1           0 |
213. | 28      0    12         1       1           1 |
214. | 28      0    15         1       1           0 |
215. | 28      1    18         1       1           0 |
     +-----------------------------------------------+
```

time varying predictor

individuals with the outcome stop contributing data after interval of outcome occurrence

mos: months on study, **hsv2:** indicator of incident HSV-2 infection,
agecat: age at baseline, **stihx:** baseline STI history,
newparts: recent sexual partners (Y/N)

Pooled logistic regression for grouped event time outcomes

- Example: incident HSV-2 infection in the MIRA study

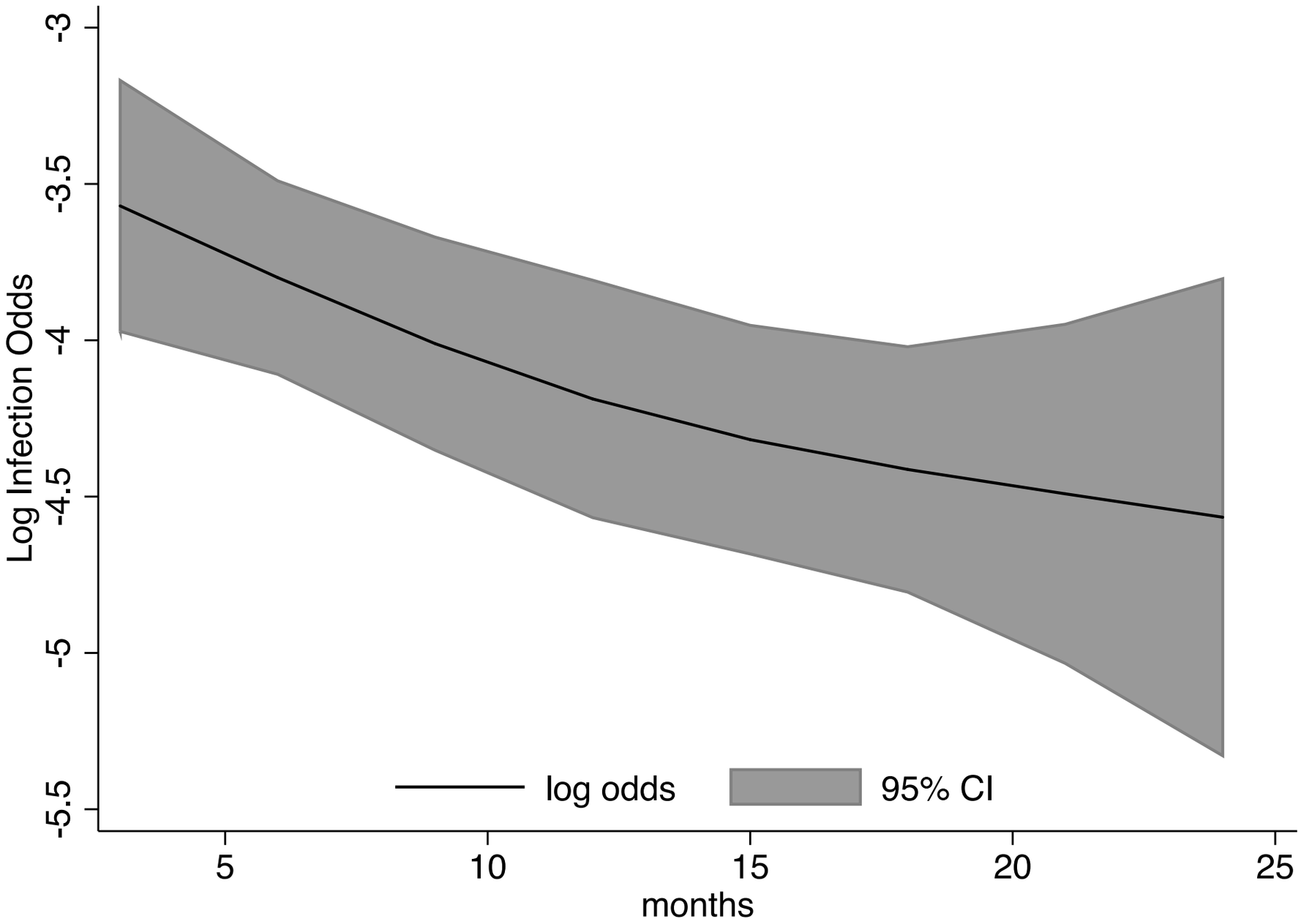
```
. generate spline basis to allow for potentially nonlinear effects of time
. mkspline spl = mos, cubic nknots(3)
. logistic hsv2 spl* i.agecat i.stihx newparts
Logistic regression                                     Number of obs   =       6069
                                                        LR chi2(6)      =      33.69
                                                        Prob > chi2     =     0.0000
Log likelihood = -509.18109                             Pseudo R2       =     0.0320

------------------------------------------------------------------------------
        hsv2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        spl1 |   .9256021   .0379368    -1.89   0.059     .8541555    1.003025
        spl2 |   1.035493   .0556198     0.65   0.516     .9320218     1.15045
             |
      agecat |
          2  |   .6090384   .1369677    -2.20   0.027     .3919372     .946396
          3  |   .5522162   .2213153    -1.48   0.138     .2517489    1.211297
             |
     1.stihx |    1.92962   .4938946     2.57   0.010     1.168431    3.186695
    newparts |   1.826313   .4372661     2.52   0.012     1.142288    2.919945
       _cons |   .0354939   .0105938   -11.19   0.000     .0197741    .0637103
------------------------------------------------------------------------------
```

odds ratios for `spl*` predictors represent time variation (in months) in baseline HSV-2 risk; interest usually focuses on the other predictors in the model

Pooled logistic regression for grouped event time outcomes

• Example: Baseline log infection odds for HSV-2 example on previous slide (plotted using `xblc` as in lab 8)



Linearity of log odds with months of follow-up appears reasonable

Cox regression for HSV-2 example

```
. stset mos hsv2, id(id)
. stcox i.agecat i.stihx newparts, efron

        failure _d:  hsv2
  analysis time _t:  mos
                id:  id
Cox regression -- Efron method for ties

No. of subjects =         1000                     Number of obs    =      6069
No. of failures =          104
Time at risk    =        18207
                                                   LR chi2(4)       =     20.55
Log likelihood  =   -691.38351                     Prob > chi2      =    0.0004
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
           2 |   .6131285   .1365595    -2.20   0.028     .3962479    .9487156
           3 |   .5554717   .2209092    -1.48   0.139     .2547663    1.211105
             |
     1.stihx |   1.918896   .4819664     2.59   0.009     1.172888    3.139399
    newparts |   1.793361   .4225623     2.48   0.013     1.130063    2.845987
------------------------------------------------------------------------------
```

- Results quite similar to pooled logistic regression results
- Modeling of the effect of time is *not required* for the Cox proportional hazards model (courtesy of the *proportional hazards assumption* and *partial likelihood*)
- Cox regression model doesn't accept weights (can't be used to control for time dependent confounder/mediators)

Unmatched Case-Control Studies

Definition

Basic design principle: Sample individuals with outcome (cases) and without outcome (controls), determine exposure retrospectively.

- controls must be representative of target population
- quality of exposure information from cases & controls must be equally good

Advantages of case-control approach:

- appropriate for rare outcomes
- often require smaller samples; less expensive than comparable prospective studies
- odds ratios approximate relative risk (for rare outcomes)

Disadvantages:

- recall bias a factor is measuring exposures/risk factors
- temporal ordering of exposure & disease (i.e. causal links) not always clear
- estimates effects have a conditional (not marginal) causal interpretation

Example: Ille-et-Vilaine study of esophogeal cancer

Sample of 200 adult males with esophogeal cancer from regional hospitals in Ille-et-Vilaine (Brittany); sample of 778 adult male controls drawn from election records in same region.

```
. use esoph
(Ille-et-Vilaine case-control study of esophageal cancer)
. describe
Contains data from esoph.dta
  obs:           975                          Ille-et-Vilaine case-control
                                                study of esophageal cancer
 vars:             7                          7 Feb 2001 08:40
 size:        31,200 (96.9% of memory free)
-------------------------------------------------------------------------------
   1. case      float  %9.0g                  Case status (1=case, 0=control)
   2. age       float  %9.0g
   3. agegp     float  %9.0g       agelab     Age group
   4. tob       float  %9.0g                  Tobacco consumption gm/day
   5. tobgp     float  %10.0g      toblab     Grouped tobacco consum.
   6. alc       float  %9.0g                  Alcohol consumption gm/day
   7. alcgp     float  %11.0g      alclab     Grouped alcohol consum.
-------------------------------------------------------------------------------
```

Example: Ille-et-Vilaine study of esophogeal cancer

```
. tabulate case
Case status |
   (1=case, |
 0=control) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        775       79.49       79.49
          1 |        200       20.51      100.00
------------+-----------------------------------
      Total |        975       100.0
```

```
. table agegp
----------+-----------
Age group |      Freq.
----------+-----------
    25-34 |         116
    35-44 |         199
    45-54 |         213
    55-64 |         242
    65-74 |         161
      75+ |          44
----------+-----------
```

```
. table tobgp
-----------+-----------
Grouped    |
tobacco    |
consum.    |      Freq.
-----------+-----------
0-9 gm/day |         526
     10-19 |         236
     20-29 |         131
       30+ |          82
-----------+-----------
```

```
. table alcgp
------------+-----------
  Grouped   |
  alcohol   |
consum.     |      Freq.
------------+-----------
0-39 gm/day |         414
      40-79 |         355
     80-119 |         139
       120+ |          67
------------+-----------
```

Are odds ratios valid measures of association for the effect of predictors on the outcome in case-control studies?

Example: association between outcome (case/control status) and binary predictor for smoking in Ille-et-Vilaine study

Odds ratio for smoking status (outcome) by cases/control status (predictor):

```
. tabodds ditob case, or
------------------------------------------------------------------------------
        case |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           0 |    1.000000          .            .                .           .
           1 |   10.407051       64.89       0.0000        5.119049   21.157585
------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =    64.89
                                  Pr>chi2  =   0.0000
Score test for trend of odds:     chi2(1)  =    64.89
                                  Pr>chi2  =   0.0000
```

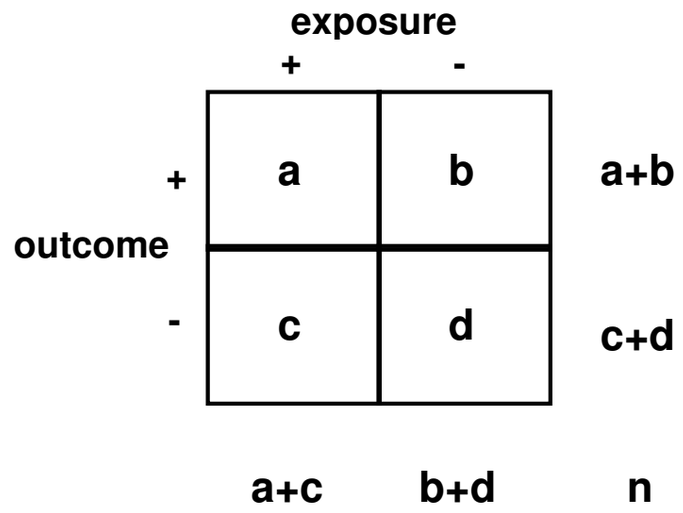
Example: association between outcome (case/control status) and binary predictor for smoking in Ille-et-Vilaine study

Odds ratio for disease by smoking status:

```
. tabodds case ditob, or
------------------------------------------------------------------------------
       ditob |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   0-9 gm/~y |    1.000000          .          .               .           .
   10+ gm/~y |   10.407051      64.89     0.0000        5.119049   21.157585
------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =    64.89
                                  Pr>chi2  =   0.0000
Score test for trend of odds:     chi2(1)  =    64.89
                                  Pr>chi2  =   0.0000
```

- odds ratios in last two slides are identical (i.e. due to symmetry between exposure risk in cases compared to controls and disease risk in exposed compared to unexposed when risk is measured using the *OR*)

Interpretation of odds ratios for case-control studies



- For any binary outcome and a single binary predictor:

- For prospective and cross-sectional studies, odds ratios are estimated with the odds of the outcome in the "exposed" and "unexposed" groups defined the the predictor:

$$\text{odds for exposed} = \frac{a/(a+c)}{c/(a+c)}, \text{ odds for unexposed} = \frac{b/(b+d)}{d/(b+d)}, \quad OR = \frac{\frac{a/(a+c)}{c/(a+c)}}{\frac{b/(b+d)}{d/(b+d)}} = \frac{ad}{bc}$$

- For **case-control** studies, odds ratios are estimated with the odds of exposure among those with and without the outcome:

$$\text{odds for outcomes} = \frac{a/(a+b)}{c/(a+b)}, \text{ odds for non-outcomes} = \frac{c/(c+d)}{d/(c+d)}, \quad OR = \frac{\frac{a/(a+b)}{b/(a+b)}}{\frac{c/(c+d)}{d/(c+d)}} = \frac{ad}{bc}$$

- ORs are the same under reversal of "roles" of outcome and predictor

Correspondence between odds ratios and relative risks

For rare outcomes, the OR approximates RR *(see figure from lecture 10, slide 50)*

- Unless the approximation is known to be satisfactory, it's good practice to avoid referring to odds ratios as relative risks

Summary odds ratios for case-control studies

Standard frequency table methods to estimate odds ratios for associations controlled for confounding variables also remain valid for case-control study data

Example: association between outcome (case/control status) and binary predictor for heavy alcohol consumption in Ille-et-Vilaine study controlled for age

```
. cc case dialc, by(agegp)
       Age group |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
           25-34 |          .             0          .             0 (exact)
           35-44 |   5.046154     .9268664   24.86538      .6532663 (exact)
           45-54 |   5.665025     2.632894   12.16536      2.859155 (exact)
           55-64 |   6.088235     3.172432   11.70469      3.933884 (exact)
           65-74 |   2.580247     1.131489   5.857261      4.024845 (exact)
             75+ |          .     4.388738          .             0 (exact)
-----------------+-------------------------------------------------
           Crude |    5.58042     3.897681   7.972556                (exact)
    M-H combined |   5.079324     3.510674   7.348884
-----------------------------------------------------------------------
Test of homogeneity (B-D)      chi2(5) =     9.17  Pr>chi2 = 0.1025

                    Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =     83.59
                                                Pr>chi2 =    0.0000
```

Definition of logistic regression model for unmatched case-control studies

If $P(x)$ is the probability of outcome occurrence among individuals with a single risk predictor at level x , the logistic regression model for a case-control sample is:

$$\log\left[\frac{P(x)}{1-P(x)}\right] = \beta_0^* + \beta_1 x$$

- β_0^* *Not interpretable* as the log odds of outcome occurrence among indiv. with $x = 0$. It carries, no information about this risk in the absence of additional population-level information about case & control sample selection probabilities:

$$\beta_0{}^* = \beta_0 + \log\left(\pi_1/\pi_0\right)$$

$$\pi_1 = \Pr\left[\text{sampled}|y=1\right] \;\; ; \pi_0 = \Pr\left[\text{sampled}=1|y=0\right]$$

- Even though study participants are sampled based on their outcome status, the conventional logistic model can be used to analyze data from case-control studies

  - the regression coefficients *retain the odds ratio interpretation* familiar from prospective and cross-sectional studies
  - the intercept coefficient *cannot* be related to outcome risk/odds (unless population sampling probabilities are known)
  - other regression models for binary outcomes *not appropriate* for case-control data

Example Ille-et-Vilaine study of esophageal cancer *(continued)*

Effect of alcohol and tobacco use controlled for age

```
. logistic case i.alcgp i.tobgp i.agegp
Logistic regression                               Number of obs   =        975
                                                  LR chi2(11)     =     285.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -352.13648                       Pseudo R2       =     0.2882
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       alcgp |
          2  |   4.212281   1.053849     5.75   0.000     2.579644    6.878204
          3  |   7.145746   2.031704     6.92   0.000     4.092898    12.47568
          4  |   36.71131   14.14065     9.35   0.000     17.25559     78.1034
             |
       tobgp |
          2  |   1.557431   .3552951     1.94   0.052     .9959212    2.435524
          3  |   1.674762   .4571307     1.89   0.059     .9808788    2.859504
          4  |   5.177771   1.781262     4.78   0.000     2.638224    10.16188
       agegp |
          2  |   7.165294    7.91497     1.78   0.075     .8221946    62.44438
          3  |   43.48488   46.47178     3.53   0.000     5.353888    353.1891
          4  |   75.81783   80.80026     4.06   0.000     9.389283    612.2238
          5  |   133.3119   143.5817     4.54   0.000      16.1471    1100.635
          6  |   123.9913    139.105     4.30   0.000     13.75417    1117.759
             |
       _cons |   .0010163   .0011043    -6.34   0.000     .0001208    .0085484
------------------------------------------------------------------------------
```

Example: Ille-et-Vilaine study of esophageal cancer *(continued)*

Use of likelihood ratio test to evaluate independent effects of alcohol, tobacco and age:

```
* save results of previously fitted model
. estimates store M1
* fit model excluding tobacco effects
. quietly logistic case i.alcgp i.agegp
* test contribution of tobacco effects to joint model using lrtest
. lrtest M1
likelihood-ratio test                                 LR chi2(3)  =     23.69
(Assumption: . nested in M1)                          Prob > chi2 =    0.0000

* repeat for alcohol & age effects
. quietly logistic case i.tobgp i.agegp
. lrtest M1
likelihood-ratio test                                 LR chi2(3)  =    127.52
(Assumption: . nested in M1)                          Prob > chi2 =    0.0000

. quietly logistic case i.tobgp i.alcgp
. lrtest M1
likelihood-ratio test                                 LR chi2(5)  =    126.52
(Assumption: . nested in M1)                          Prob > chi2 =    0.0000
```

- ***Note***: A complete analysis would also consider possible interaction and mediation, as well as considering alcohol and tobacco as continuous measures.

Matched Case-Control Studies

Basic design principle: Sample individuals with outcome (cases) and without outcome (controls). Pair each case with one or more controls according to levels of specified predictors. Determine exposure retrospectively.

- 1:1, or *matched pairs* design matches each case with exactly one control
- 1:m design matches m controls with each case
- *frequency matching* fixes ratio of cases/controls within strata formed by selected confounding variables

Advantages:

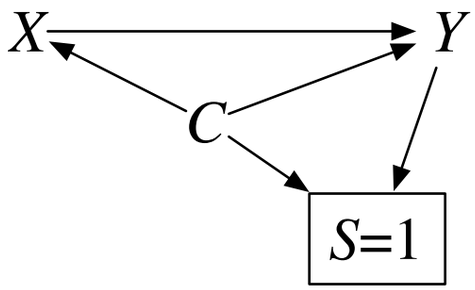
- can be more efficient than unmatched studies (by providing balance of cases & controls over strata defined by matching variables)
- can help to reduce confounding bias
- applies to confounders that are hard to quantify (e.g. geographic location, familial traits)

Disadvantages:

- more complicated and expensive than unmatched designs
- inefficient when matching variables not strongly related to outcome/exposure
- may require a specialized matched analysis
- analysis using methods for unmatched design may yield same conclusions
- recent work indicates that obtaining unbiased effect estimates from matched studies can be challenging unless matching variables are explicitly controlled for in analyses

Matched Case-Control Studies and confounding

Example DAG from a matched case-control study:



- Y: outcome
- X: exposure
- C: confounder
- S: sample inclusion indicator (the box indicates selection based on matching on C) - represents a collider on the backdoor path between X and Y

Issues:

- Matching does not necessarily control confounding bias due to the matching variables
  - matched controls may not be representative of target population
  - matching may make controls and cases similar for the exposure/risk factor of interest
- An analysis accounting only for matching (S), but ignoring adjustment for C can lead to bias (via the collider role of S)
- In general, analyses for matched studies should control for matching variables when possible

References

- Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams. Int J Epidemiol. 2013;42(3):860-9.
- Mansournia MA, Jewell NP, Greenland S. Case-control matching: effects, misconceptions, and recommendations. Eur J Epidemiol. 2018;33(1):5-14.
- Pearce N. Analysis of matched case-control studies. BMJ. 2016;352

Example: matched case-control study of study of the association between use of estrogens and endometrial cancer (Mack, *et al*, 1976, NEJM)

```
. use endom
(Leisure World matched case-control study of endometrial cancer)
. describe
Contains data from endom.dta
  obs:           394                          Leisure World matche
                                                case-control study of
                                                endometrial cancer
 vars:             8                          12 Feb 2001 13:17
-------------------------------------------------------------------------------
   1. set          float  %9.0g                 Matched set indicator
   2. case         float  %9.0g                 Case status (1=case, 0=control)
   3. age          float  %9.0g                 Age in years
   4. gall         float  %9.0g       vlab      Gallbladder disease
   5. hyp          float  %9.0g       vlab      Hypertension
   6. ob           float  %9.0g       vlab      Obesity
   7. est          float  %9.0g       vlab      Estrogen usage
   8. dose         float  %11.0g      doslab    Dose of conjugated estrogen
-------------------------------------------------------------------------------
. * n = 315 (4 controls matched to each case by age & marital status; 63 matched sets)
. list set case age est if _n<=10
          set       case        age        est
  1.        1          1         74        Yes
  2.        1          0         75         No
  3.        1          0         74         No
  4.        1          0         74         No
  5.        1          0         75        Yes
  6.        2          1         67        Yes
  7.        2          0         67        Yes
  8.        2          0         67         No
  9.        2          0         67        Yes
 10.        2          0         68        Yes
 .* Note: the marital status variables is not available in this dataset
```

Example: matched case-control study of endometrial cancer (*continued*)

Frequency table methods for unmatched data can be used for single categorical predictors.

In Stata, the `mhodds` command estimates the Mantel-Haenszel odds ratio for the association between outcome and predictor controlling for the matched sets, and gives 95% confidence limits:

```
. mhodds case est set

Mantel-Haenszel estimate of the odds ratio
Comparing est==1 vs est==0, controlling for set

Note: only 58 of the 63 strata formed in this analysis
contribute information about the effect of the explanatory variable

------------------------------------------------------------------
 Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------
   8.461538      31.16        0.0000         3.437773  20.826746
------------------------------------------------------------------
```

- When additional predictors are involved, logistic regression is required.

Conditional logistic model for matched case-control studies

- Logistic regression model for matched case-control data is defined and applied similarly to the model for unmatched data. However, the version for matched data is based on a "conditional" model for the relationship between the outcome and predictors.

- Motivation for conditional analysis is to avoid estimating effects of matching variables by conditioning on them in estimation (can yield more efficient estimates)

- In cases where there are many matched sets characterized by unique values of the matching variables (e.g. twin studies), it may be infeasible to control for matching in a conventional logistic model

- In cases where matching is based on measured characteristics (e.g. age), conventional (unmatched) logistic regression controlling for matching variables will generally provide valid results

- The conditional approach only estimates conditional ORs - marginal effect estimates are not possible (reference: Rose S, van der Laan MJ. Why match? Investigating matched case-control study designs with causal effect estimation. Int J Biostat. 2009; 5:Article 1.)

Example: matched case-control study of endometrial cancer (*continued*)

Effect of estrogen use on endometrial cancer risk estimated using the conditional logistic model

```
. clogit case est, group(set) 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

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         est |   7.954681   3.347525     4.93   0.000     3.486714    18.14802
------------------------------------------------------------------------------
```

- `or` option requests that odds ratios be printed rather than coefficients
- model does not include a constant term

Example: matched case-control study of endometrial cancer (*continued*)

Effect of estrogen use controlled for history of gall bladder disease:

```
. clogit case est gall, group(set) or
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
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         est |   8.287802   3.644927     4.81   0.000     3.500144    19.62424
        gall |   3.577465   1.469865     3.10   0.002     1.598984    8.003994
------------------------------------------------------------------------------
```

"Full" model for the joint effect of estrogen and history of gall bladder disease:

```
. clogit case i.est##i.gall, group(set) or
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
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.est  |   14.88179   9.104216     4.41   0.000     4.486594    49.36211
     1.gall  |   18.07166   15.95823     3.28   0.001     3.201415    102.0127
             |
   est#gall  |
       1 1   |   .1283818   .1277365    -2.06   0.039     .0182633     .902457
------------------------------------------------------------------------------
. * evidence for multiplicative interaction between est & gall
```

Example: matched case-control study of endometrial cancer (*continued*)

Use of `lincom` to get component odds ratios for estrogen use among women w/wo gall bladder disease:

```
. lincom 1.est + 1.est#1.gall, or

 ( 1)  [case]1.est + [case]1.est#1.gall = 0
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.910551   1.517452     0.82   0.415     .4028036    9.061997
------------------------------------------------------------------------------
```

```
. lincom 1.est, or

 ( 1)  [case]1.est = 0
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   14.88179   9.104216     4.41   0.000     4.486594    49.36211
------------------------------------------------------------------------------
```

Note: the `or` option is required for odds ratios when `lincom` is used after `clogit`.

Example: matched case-control study of endometrial cancer (*continued*)

Compare to results from corresponding unmatched analysis:

```
. logistic case i.est##i.gall age
Logistic regression                               Number of obs     =        315
                                                  LR chi2(4)        =      49.91
                                                  Prob > chi2       =     0.0000
Log likelihood = -132.67224                       Pseudo R2         =     0.1583

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         est |
         Yes |   15.46892   9.492168     4.46   0.000     4.646637    51.49694
             |
        gall |
         Yes |   19.22229   16.29279     3.49   0.000     3.650294    101.2238
             |
    est#gall |
    Yes#Yes |   .1085787   .1024346    -2.35   0.019     .0170887    .6898893
             |
         age |   1.013276   .0250903     0.53   0.594     .9652738    1.063664
       _cons |    .009936   .0186745    -2.45   0.014     .0002497     .395389
------------------------------------------------------------------------------
```

- Estimates fairly close to conditional logistic results in this case
- In this example, a conventional unmatched analysis might be preferred if the marital status matching variable was available and included in the model.

Regression analysis for matched case-control studies

Conclusions

- matching should be accounted for in analysis if it is used in study design
  - conventional unmatched analyses may be used in many cases, but matching variables should be included for adjustment
  - in cases where there are many matched sets, a conditional logistic regression analysis may be needed - adjustment for known confounders (including matching factors) is often advisable (if feasible).
- inappropriate matching (”overmatching”) can lead to inefficiency (i.e. higher variances) than corresponding unmatched designs
  - in general matching should be restricted to variables with known and strong confounding roles