Alternative Binary Regression Models; Logistic Regression for Case-Control Studies

March 17, 2015

Biostatistics 208

Alternative regression models for binary outcomes

As discussed previously, there are situations where an alternative model for a binary outcome may be appropriate:

- When risk is thought to be related to predictors in a fashion other than assumed in the logistic model (i.e. linear and additive in the log odds)
  - For example, relative risk and risk difference-based models
- When risk of the outcome varies with time of follow-up in a prospective study
  - Survival analysis methods may be more appropriate in such situations
- In cases where external information suggests a specific formulation
  - HIV transmission example discussed in section 5.5.1 of text

Example: relationship between probability of coronary heart disease age and BMI in the WCGS study (excluding the extremely low value of BMI noted in the previous lab)

Logistic regression via `binreg`:

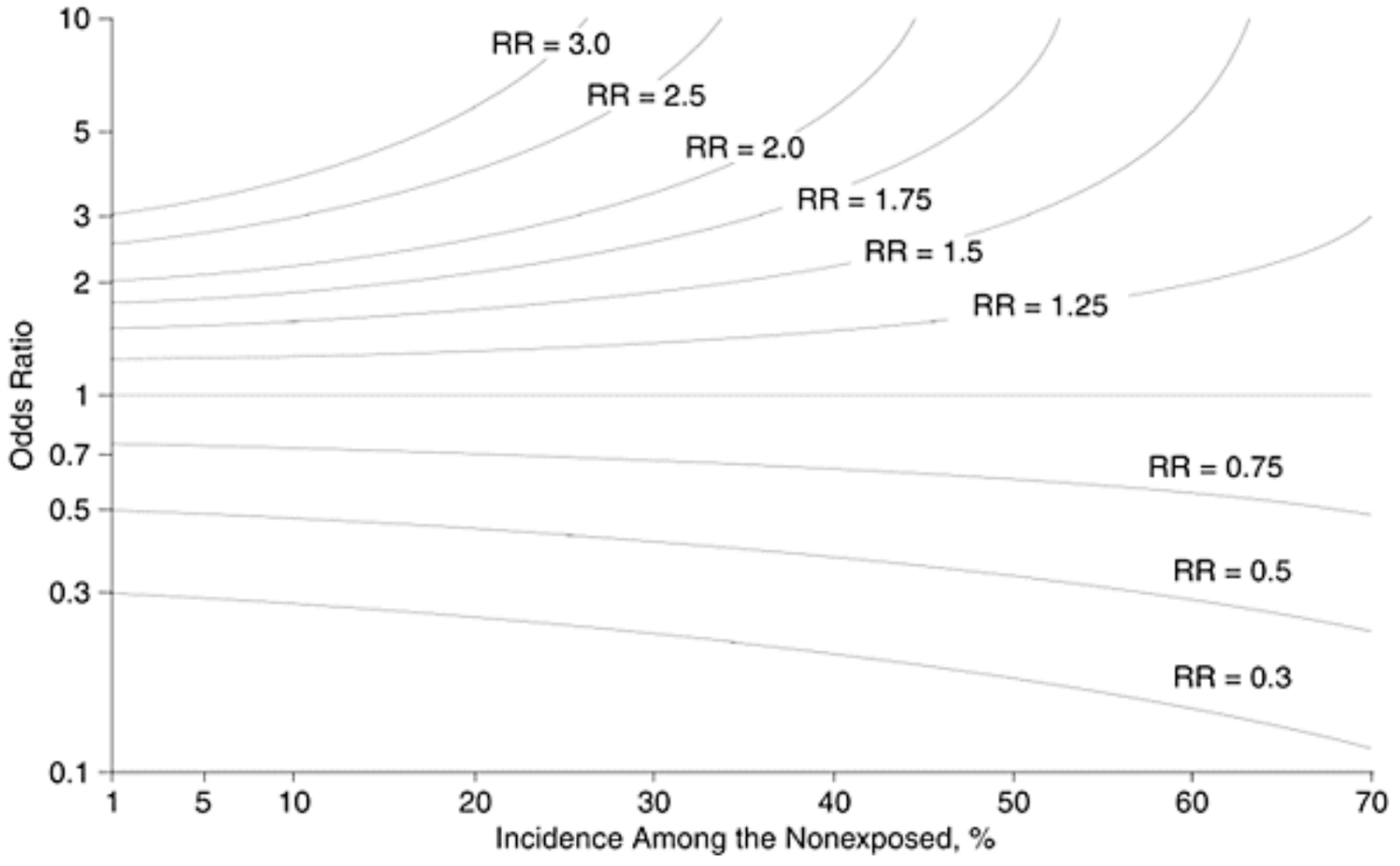
```
. binreg chd69 age bmi if bmi>12 & bmi!=., or

Generalized linear models                         No. of obs      =      3153
Optimization     : MQL Fisher scoring             Residual df     =      3150
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  1727.411122                   (1/df) Deviance =  .5483845
Pearson          =   3155.69054                   (1/df) Pearson  =  1.001807

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  BIC             = -23649.33
------------------------------------------------------------------------------
             |                 EIM
      chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.076649   .0122104     6.51   0.000     1.052981    1.100849
         bmi |    1.08468   .0264501     3.33   0.001     1.034058     1.13778
       _cons |   .0003617   .0002961    -9.68   0.000     .0000727    .0017999
------------------------------------------------------------------------------
```

Note: The `binreg` comand in Stata allows fitting a number of alternative *generalized linear regression* models for binary outcomes, including *logit* (i.e. log odds), *log* (log probability), and *linear* (i.e. linear in probability).

Problem: Want to summarize age adjusted effect of BMI using a *relative risk* rather than an odds ratio

- *Why not use the logistic model and either assume that estimated ORs approximate RRs, or use the probability form of the model to estimate RRs?*

  - Odds ratios only approximate relatives risks in the rare outcome setting (see *figure*)
  - Relative risks estimated using the probability form of the logistic model depend on other predictors in the model



Problem: Want to summarize age adjusted effect of BMI using a *relative risk* rather than an odds ratio

- *Solution*: Use a *log relative risk regression* model based on the log of the outcome probability:

$$\log P(\texttt{age},\texttt{bmi}) = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{bmi}$$

- The coefficient β_2 is interpreted as the log relative risk associated with a unit change in `bmi`, holding `age` fixed.

- The coefficient β_1 is interpreted as the log relative risk associated with a unit change in `age`, holding `bmi` fixed.

- *Note*: recall that the coefficients for this model have to be estimated so that the left-hand side specifies a valid outcome probability (i.e. $0 \leq P \leq 1$, or equivalently $-\infty < \log P \leq 0$) for all values of `bmi` and `age`. This *constraint sometimes causes fitting problems* not encountered with the logistic model.

- Similar constraints and numerical difficulties arise in fitting additive risk difference models

Relative risk regression (*continued*)

- log relative risk models can be fitted in Stata using the `binreg` command:

```
. binreg chd69 age bmi if bmi>12 & bmi!=., rr iterate(100)
Generalized linear models                         No. of obs      =      3153
Optimization     : MQL Fisher scoring             Residual df     =      3150
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  1727.536638                   (1/df) Deviance =  .5484243
Pearson          =  3149.490834                   (1/df) Pearson  =  .9998384

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u)                   [Log]

                                                  BIC             = -23649.21
------------------------------------------------------------------------------
             |                 EIM
       chd69 | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.068756   .0107595     6.61   0.000     1.047874    1.090053
         bmi |   1.073924   .0228021     3.36   0.001      1.03015    1.119558
       _cons |   .0005968   .0004258   -10.41   0.000     .0001474    .0024165
------------------------------------------------------------------------------
```

- Results quite similar to logistic model in this case

- Because of the constraint mentioned above, sometimes Stata will fail to estimate the parameters from a log rel. risk model. The `iterate(100)` option stops estimation after 100 steps if convergence problems are encountered

Relative risk regression (*continued*)

- Approximate log relative risk regression for binary outcomes using the `poisson` regression command:

```
. poisson chd69 age bmi if bmi>12 & bmi!=., irr robust

Poisson regression                                Number of obs   =       3153
                                                  Wald chi2(2)    =      57.00
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -876.93885                 Pseudo R2       =     0.0270
------------------------------------------------------------------------------
             |               Robust
       chd69 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.068913   .0106896     6.66   0.000     1.048165     1.09007
         bmi |   1.075168   .0218344     3.57   0.000     1.033214    1.118825
       _cons |   .0005759   .0004083   -10.52   0.000     .0001435    .0023113
------------------------------------------------------------------------------
```

- This approach is generally recommended for models involving multiple predictors (especially those including continuous variables)

- Results very close to the estimates on the previous slide
- `robust` variance option *required* to obtain appropriate standard error estimates

Reference: Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. Amer J Epidemiol 2004; 159:702-6.

Example: coronary heart disease, age and BMI in the WCGS study (*continued*)

Problem: Want to summarize age adjusted effect of BMI using a *risk difference* (or excess risk) rather than an odds ratio

- *Solution*: *risk difference regression* model based directly on the outcome probability P:

$$P(\texttt{age},\texttt{bmi}) = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{bmi}$$

- The coefficient β_2 is interpreted as the risk difference associated with a unit change in `bmi`, holding `age` fixed.

- The coefficient β_1 is interpreted as the risk difference associated with a unit change in `age`, holding `bmi` fixed.

- *Note*: model coefficients have to be estimated so that the left-hand side still defines a valid outcome probability (i.e. $0 \leq P \leq 1$) for all values of `bmi` and `age`. Fitting problems are common, especially for continuous predictors.

Risk difference regression (*continued*)

- Here is the `binreg` risk difference model fit of the last example:

```
. binreg chd69 age bmi if bmi>12 & bmi!=., rd iterate(100)
Iteration 1:   deviance =  1731.716
Iteration 2:   deviance =  1731.192
Iteration 3:   deviance =  1731.189
Iteration 4:   deviance =  1731.189
Iteration 5:   deviance =  1731.189
.
.
.
Iteration 95:  deviance =  1731.182
Iteration 96:  deviance =  1731.182
Iteration 97:  deviance =  1731.182
Iteration 98:  deviance =  1731.181
Iteration 99:  deviance =  1731.181
Iteration 100: deviance =  1731.181
convergence not achieved
r(430);
```

- No convergence of the estimation process in 100 iterations (convergence is never achieved in this example, even if the iteration limit is increased to >1000).

- Model excluding BMI or using a categorical version converges (see next page):

Risk difference regression (*continued*)

- Using bmi recoded into 4 groups based on quartiles:

```
. xtile bmicat=bmi if bmi>12,nq(4)
. binreg chd69 age i.bmicat, rd
```

```
Generalized linear models                         No. of obs      =      3153
Optimization     : MQL Fisher scoring             Residual df     =      3148
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  1730.073128                   (1/df) Deviance =  .5495785
Pearson          =  3153.048459                   (1/df) Pearson  =  1.001604

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = u                       [Identity]
                                                  BIC             = -23630.56
```

chd69	Risk Diff.	EIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0051902	.000893	5.81	0.000	.00344	.0069405
bmicat						
2	.0170412	.0118722	1.44	0.151	-.0062278	.0403102
3	.0255155	.0121295	2.10	0.035	.001742	.049289
4	.0380856	.0129088	2.95	0.003	.0127848	.0633865
_cons	-.1784397	.0398784	-4.47	0.000	-.2565999	-.1002795

- Categorization leads to reasonable estimates, but a somewhat different interpretation

Conclusions about alternate binary regression approaches:

- Alternative models (e.g. relative risk / risk difference) may be preferable in some instances
  - requested by reviewers
  - scientific grounds
  - marked improvement in fit over the logistic

- Alternate models models may raise numerical estimation problems that typically don't affect logistic regression
  - The Poisson regression approach illustrated on slide 7 is the recommended way to obtain estimated relative risks for binary outcomes and multiple predictors (*not valid for case-control studies*)

- The logistic model should generally be the default choice unless these other factors are compelling
  - recall that the logistic model can also be used as a basis for causal inference (e.g. for estimating average causal effects, propensity scores, marginal structural models, etc.)

Alternatives to logistic regression for small samples & rare outcomes

As sample size shrinks:

- Reliability of inferences from standard methods (Wald (Z) test, 95% confidence intervals and even likelihood ratios diminishes
- difficulties in estimating coefficients become more common
- *Exact logistic regression* and *bootstrap methods* can both help
- What is a "small" sample?

- A related issue is *rare outcomes*, which can be an issue even in large samples
  - can cause estimation issues (hence the 10 outcomes per predictor rule of thumb)
  - frequently results in infinite OR estimates

Example: relationship between transmission of HIV to female partners of HIV infected men in CDC transmission study

Problem: Sample of $n = 31$ - calls validity of approx. inferences from `logistic` into question

```
. logistic hivp sexd60 aids
Logistic regression                                     Number of obs   =         31
                                                        LR chi2(2)      =       6.81
                                                        Prob > chi2     =     0.0331
Log likelihood = -13.151565                             Pseudo R2       =     0.2058
------------------------------------------------------------------------------
        hivp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sexd60 |   7.081807     9.7318      1.42   0.154      .479099    104.6798
        aids |   12.80809    16.1979      2.02   0.044      1.073995    152.7449
------------------------------------------------------------------------------
```

```
. exlogistic hivp sexd60 aids    (Note: i. notation doesn't work)
Exact logistic regression                               Number of obs =         31
                                                        Model score   =   6.582907
                                                        Pr >= score   =     0.0258
------------------------------------------------------------------------------
        hivp | Odds Ratio       Suff.  2*Pr(Suff.)     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.sexd60 |   5.403325           6      0.3047      .4320873    336.6659
     1.aids |    10.2224           3      0.1021      .7104864    589.9442
------------------------------------------------------------------------------
```

Modifications of logistic regression for rare outcomes

Even in large samples, outcomes may be rare

- can cause estimation issues for the logistic model (hence the 10 outcomes per predictor rule of thumb),
- can lead to biased/infinite OR estimates (bootstrap not very helpful in these cases)

- When the exact logistic approach becomes computationally infeasible, the *penalized likelihood* version of the logistic model can be used to obtain more reliable inferences

- In Stata, the downloadable `firthlogit` package can be used to fit penalized logistic models
  - *Note*: does not accept the `i.` syntax for categorical predictors and will not work with some post-estimation commands like `margins`

Reference: Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med. 2002;21:2409-19.

Example: ICU data from homework #4

Problem: Sample of $n = 200$ with 40 outcomes. Model in question #3 involved 11 predictors:

```
. logistic sta i.crn i.typ i.can i.loc i.agecat i.syscat
note: 1.loc != 0 predicts success perfectly; 1.loc dropped and 5 obs not used
Logistic regression                               Number of obs   =        195
                                                  LR chi2(10)     =      62.06
                                                  Prob > chi2     =     0.0000
Log likelihood = -60.742305                       Pseudo R2       =     0.3381
------------------------------------------------------------------------------
         sta | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.crn |   2.840999   2.165964     1.37   0.171     .6375571    12.65969
       1.typ |   52.63315   69.20641     3.01   0.003     3.999717     692.611
       1.can |   13.62877   13.28441     2.68   0.007     2.017263     92.0769
             |
       1.loc |
         1   |          1  (empty)
         2   |   16.38254   15.17629     3.02   0.003     2.665963     100.672
             |
     1.agecat |
         2   |   4.240071   3.244213     1.89   0.059     .9464375    18.99566
         3   |   5.588847   4.478942     2.15   0.032     1.161885    26.88322
         4   |   13.20398   9.938337     3.43   0.001     3.020121    57.72781
             |
     1.syscat |
         2   |   .7691673   .4733606    -0.43   0.670     .2302368    2.569608
         3   |   1.059359   .6293668     0.10   0.923     .3306305    3.394246
         4   |   .0768537   .0754065    -2.62   0.009     .0112329     .525821
             |
       _cons |   .0011421   .0018182    -4.26   0.000     .0000504    .0258683
------------------------------------------------------------------------------
```

No inference about level 1 of `loc`

Example: ICU data from homework #4

Solution: Use `firthlogit`, after generating binary predictors for levels of `loc`, `agecat` & `syscat`:

```
. firthlogit sta crn typ can locc2 locc3 agec2 agec3 agec4 sysc2 sysc3 sysc4, or

initial:       penalized log likelihood = -91.758995
```

… (additional iteration output omitted)

```
Iteration 5:   penalized log likelihood = -56.752069

                                                Number of obs    =        200
                                                Wald chi2(11)    =      30.23
Penalized log likelihood = -56.752069           Prob > chi2      =     0.0015
------------------------------------------------------------------------------
             sta | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
             crn |   2.585408   1.824617     1.35   0.178      .6483431    10.30987
             typ |   23.81719   26.04578     2.90   0.004      2.792809    203.1139
             can |   9.392172   8.038851     2.62   0.009      1.754747    50.27102
           locc2 |   1622.537   3584.976     3.35   0.001      21.35382      123286
           locc3 |   11.43597   9.426443     2.96   0.003      2.273238    57.53094
           agec2 |   3.620715   2.560872     1.82   0.069      .9052191    14.48222
           agec3 |   4.738263   3.503944     2.10   0.035      1.112131     20.1875
           agec4 |   10.26244     7.1259     3.35   0.001      2.631508    40.02182
           sysc2 |   .7950167   .4641699    -0.39   0.694      .2531645    2.496604
           sysc3 |   1.059447   .5989035     0.10   0.919      .3498615    3.208204
           sysc4 |   .1157282   .1005024    -2.48   0.013      .0210972     .634823
           _cons |   .0032221   .0043901    -4.21   0.000       .000223    .0465493
------------------------------------------------------------------------------
```

Estimated OR for level 1 of `loc`

Logistic regression for small samples/rare outcomes

Conclusion:

- exact logistic results preferred for small samples ($n < 50$), especially when standard approach fails to yield estimates
- penalized likelihood approach useful for problems with rare outcomes, and in cases where the exact approach is computationally infeasible
- bootstrap inference useful for small-moderate samples with adequate number of outcomes, and where exact methods are computationally infeasible
- conventional logistic regression preferred for moderate-large samples with adequate numbers of outcomes

Pooled logistic regression for grouped event time outcomes

- Application of the logistic regression model to survival (event time) outcomes
- For event times observed at regular intervals (e.g. visit intervals in prospective studies)
- Data structured in "long" form, with one observation per person-visit, binary indicator of event occurrence (at most once / individual)
- Coefficients approximate 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 weights - used to fit *marginal structural models* for causal inference involving time-dependent confounders
- Disadvantage: requires explicit model for the time dependence of outcome risk

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 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 dependence can be modeled by a 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 (via incorporation of inverse probability weights) (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

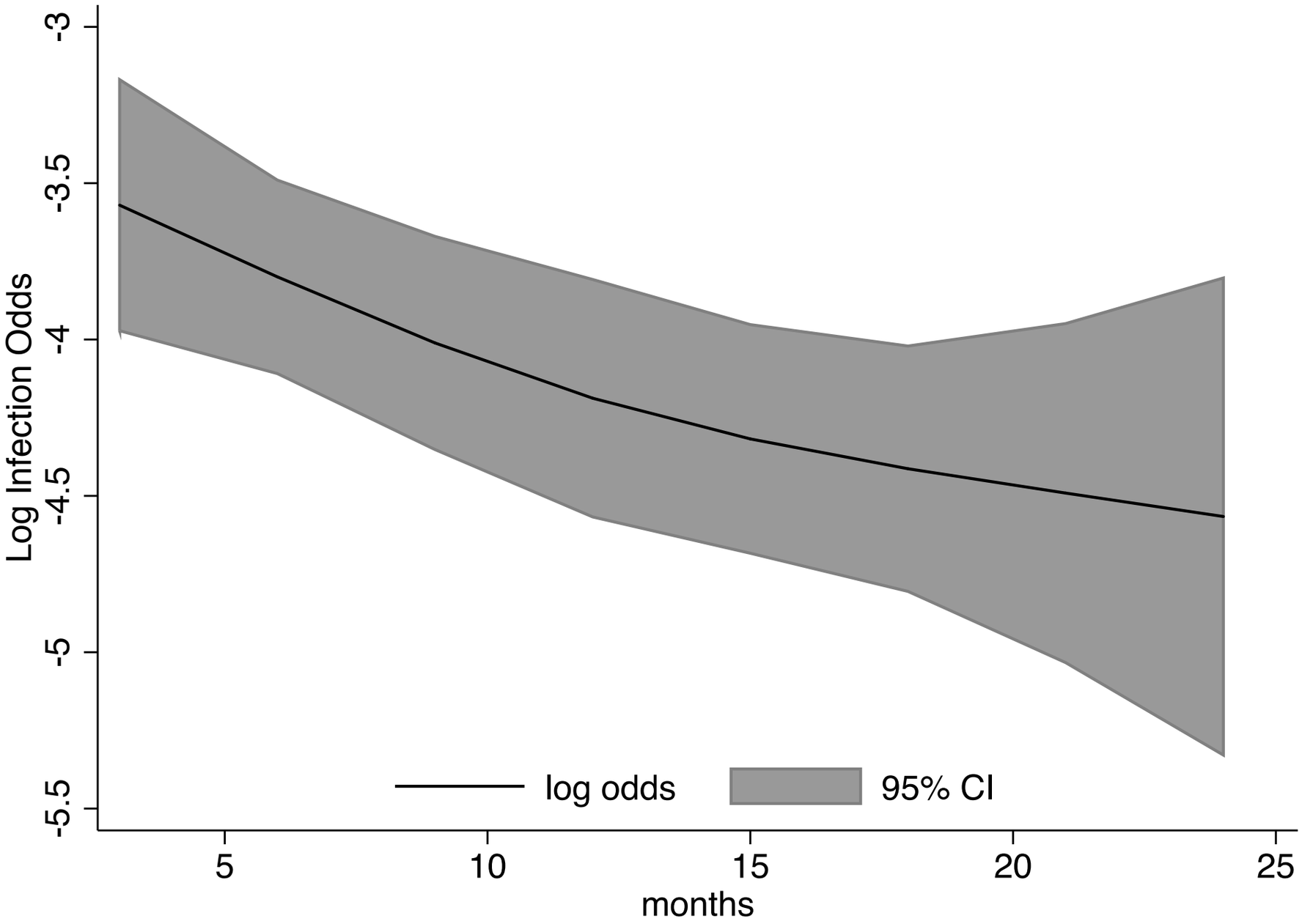
- Example: incident HSV-2 infection in the MIRA study

```
. 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)

Logistic Regression for case-control data

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

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:

outcome \ exposure	+	-	
+	a	b	a+b
-	c	d	c+d
	a+c	b+d	n

- For prospective and cross-sectional studies, odds ratios are estimated with the odds of the outcome among those with and without 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 in slide 4)*

- since the quality of this approximation isn't always known, it's good practice to call an odds ratio an *OR*

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
  - 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:

- "built-in" control for known confounding variables (e.g. age, gender)
- can be more powerful than unmatched studies when confounding is effectively controlled

Disadvantages:

- effects of matching variables cannot be considered in the analysis
- inefficient when matching variables are not strongly related to outcome/exposure
- more complicated and expensive than unmatched designs
- analysis must explicitly account for matching variables
- in many cases appropriate analysis of unmatched design yields same conclusions

Matching works well in situations with identified strong confounders that are hard to quantify (e.g. geographic location, familial traits)

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
                                                d 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
```

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

Frequency table methods for matched data can be used for categorical predictors. In Stata, the `mhodds` command estimates the Mantel-Haenszel odds ratio for the association between outcome and predictor 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
------------------------------------------------------------------
```

- The conditional logistic regression model (`clogit` in Stata) provides a more convenient approach to analyzing multiple predictor data and yields the (preferred) conditional maximum likelihood estimate of the odds ratio.

Conditional logistic model for matched case-control studies

- The logistic regression model for a matched case-control sample 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.

- For given risk factor(s) the model for each matched set is expressed as the probability of exposure to the observed risk factor(s) for the case *conditional* on the possible predictor profiles of all individuals in the set.

- motivation for conditional analysis is to concentrate on effects of predictors of interest and avoid estimating effects of matching variables (accounted for in the study design, and treated as a "nuisance" in the analysis)

- special software programs are needed to fit *conditional logistic regression* models

- analyzing data from a matched study using standard (unconditional) logistic regression may result in biased estimates and incorrect inferences

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

Effect of estrogen use on endometrial cancer risk

```
. 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
Logistic regression                               Number of obs   =        315
                                                  LR chi2(3)      =      49.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -132.81405                       Pseudo R2       =     0.1574

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.est |   15.10811   9.241238     4.44   0.000     4.555669    50.10349
     1.gall |       19.5   16.51041     3.51   0.000     3.709647     102.503
            |
   est#gall |
     1 1    |   .1075581   .1013885    -2.37   0.018     .0169538    .6823686
            |
      _cons |    .025641   .0149924    -6.27   0.000     .0081514    .0806561
------------------------------------------------------------------------------
```

- Estimates fairly close to conditional logistic results in this case

Conditional logistic model for matched case-control studies

Conclusions

- matching should be accounted for in analysis if it is used in study design (except possibly when sizes of matched sets is very large)
  - inappropriate use of unmatched analyses can lead to bias (usually attenuation) in estimated regression coefficients
  - for matched studies, results of (incorrect) unmatched analysis are often close to (correct) matched results
- inappropriate matching (”overmatching”) can lead to inefficiency (i.e. higher variances) than corresponding unmatched designs