Logistic Regression

March 2, 2021

Biostatistics 208

Review of logistic regression for a single predictor

Example: dependence of CHD risk (P) on the binary predictor `arcus` in WCGS

- `arcus` $= 1$ for those with condition, `arcus` $= 0$ for those without
- `chd69` is a binary indicator of CHD at the end of follow-up

Default output of model - reports exponentiated coefficients (ORs) and 95% CIs:

```
. logistic chd69 i.arcus
Logistic regression                             Number of obs     =      3,152
                                                LR chi2(1)        =      12.98
                                                Prob > chi2       =     0.0003
Log likelihood = -879.10783                     Pseudo R2         =     0.0073

------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       arcus |
     present |   1.635281   .2195036     3.66   0.000     1.257001    2.127399
       _cons |    .074344   .0062298   -31.02   0.000     .0630839    .0876141
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

- default output of Stata logistic regression procedure gives the odds ratio comparing the odds of CHD among men with `arcus` to the odds among men without `arcus`
- the odds of outcome occurrence in the baseline (`arcus` $= 0$) group are also provided

Review of logistic regression for a single predictor

Example: dependence of CHD risk (P) on the binary predictor `arcus` in WCGS

`logistic` output with coef option - reports coefficients and 95% CIs:

```
. logistic chd69 i.arcus, coef
Logistic regression                             Number of obs     =      3,152
                                                LR chi2(1)        =      12.98
                                                Prob > chi2       =     0.0003
Log likelihood = -879.10783                     Pseudo R2         =     0.0073

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       arcus |
    present  |   .4918144   .1342299     3.66   0.000     .2287286    .7549002
       _cons |  -2.599052   .0837965   -31.02   0.000     -2.76329   -2.434814
------------------------------------------------------------------------------
```

- $\beta_0 = -2.599052$, the log odds of CHD among individuals with `arcus` $= 0$
- $\beta_1 = 0.4918141$, the log odds ratio comparing the odds CHD among individuals with `arcus` $= 1$ to the odds of CHD among individuals with `arcus` $= 0$

$$\log\left[\frac{P(\texttt{arcus})}{1-P(\texttt{arcus})}\right] = -2.599052 + 0.4918141\texttt{arcus}$$

Review of logistic regression for a single predictor

Note: the logistic model on the previous slide can be expressed in these equivalent ways

Log-odds :
$$\log\left[\frac{P(\texttt{arcus})}{1-P(\texttt{arcus})}\right] = \beta_0 + \beta_1\texttt{arcus}$$

Odds :
$$\frac{P(\texttt{arcus})}{1-P(\texttt{arcus})} = \exp\left(\beta_0 + \beta_1\texttt{arcus}\right)$$

Probability:
$$P(\texttt{arcus}) = \frac{\exp\left(\beta_0 + \beta_1\texttt{arcus}\right)}{1+\exp\left(\beta_0 + \beta_1\texttt{arcus}\right)}$$

Review of logistic regression for a single predictor

Note: The effect of `arcus` on the outcome is *additive* on the log odds scale, and *multiplicative* on the odds scale

The log odds of CHD for an indiv. with arcus (`arcus` =1) is equal to the log odds for an indiv. without arcus (`arcus` =0) plus β_1:

$$\log\left[\frac{P(\texttt{arcus}=1)}{1-P(\texttt{arcus}=1)}\right]-\log\left[\frac{P(\texttt{arcus}=0)}{1-P(\texttt{arcus}=0)}\right]=(\beta_0+\beta_1)-\beta_0=\beta_1$$

The odds of CHD for an indiv. with arcus (`arcus` = 1) is equal to the odds for an indiv. without arcus (`arcus` = 0) times $\exp(\beta_1)$:

$$\frac{\dfrac{P(\texttt{arcus}=1)}{1-P(\texttt{arcus}=1)}}{\dfrac{P(\texttt{arcus}=0)}{1-P(\texttt{arcus}=0)}}=\frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)}=\frac{\exp(\beta_0)\exp(\beta_1)}{\exp(\beta_0)}=\exp(\beta_1)$$

Example: dependence of CHD risk (P) on the binary predictor `arcus` in WCGS

$$\log\left[\frac{P(\texttt{arcus})}{1-P(\texttt{arcus})}\right] = -2.599052 + 0.4918141\texttt{arcus}$$

among those without arcus ($\texttt{arcus} = 0$):

log-odds of CHD: -2.599052

odds of CHD: exp(-2.599052) = 0.07434402

probability of CHD: 0.07434402 / (1 + 0.07434402) = 0.06919945

among those with arcus ($\texttt{arcus} = 1$):

log-odds of CHD: -2.599052 + 0.4918141 = -2.1072379

odds of CHD: exp(-2.599052 + 0.4918141) = 0.1215733

probability of CHD: 0.1215733 / (1 + 0.1215733) = 0.10839532

comparison of those with arcus ($\texttt{arcus} = 1$) to those without ($\texttt{arcus} = 0$):

log odds ratio : (-2.1072379) - (-2.599052) = 0.4918141

odds ratio : exp(0.4918141) = 1.63528

95% C.I. for odds ratio: (exp(0.2287283) , exp(0.7548999)) ; (1.26, 2.13) (based on the 95% CI for the coefficient in the previous output)

Example: dependence of CHD risk (P) on the binary predictor `arcus` in WCGS (continued)

- The logistic model (below) can also be used to compute other measures of risk:

$$RR = \frac{P(1)}{P(0)} = \frac{\exp(-2.1072379)/[1+\exp(-2.1072379)]}{\exp(-2.599052)/[1+\exp(-2.599052)]} = \frac{0.10839532}{0.06919945} = 1.566419$$

$$RD = P(1) - P(0) = 0.10839532 - 0.06919945 = 0.03919587$$

$$\log\left[\frac{P(\texttt{arcus})}{1-P(\texttt{arcus})}\right] = -2.599052 + 0.4918141\texttt{arcus}$$

- Logistic models frequently used in estimation of marginal causal effects
- *Note*: Estimated *RR* & *RD* from a logistic model won't in general correspond to estimates obtained from a relative risk or risk difference binary regression approach, or to *marginal effects* estimated from the logistic model, especially when additional predictors are included.

Stata alternatives for fitting logistic models (all essentially *equivalent*)

- `logit` command - equivalent to `logistic` with `coef` option:

```
. logit chd69 arcus
------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       arcus |   .4918144   .1342299     3.66   0.000     .2287286    .7549002
       _cons |  -2.599052   .0837965   -31.02   0.000    -2.76329   -2.434814
------------------------------------------------------------------------------
```

Note: Normally not needed - useful to override default `lincom` reporting of ORs after using `logistic`

- `glm` command - umbrella for the whole family of generalized linear models:

```
. glm chd69 arcus, family(binomial)
------------------------------------------------------------------------------
             |                 OIM
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       arcus |   .4918139   .1342299     3.66   0.000     .2287281    .7548997
       _cons |  -2.599052   .0837965   -31.02   0.000    -2.76329   -2.434814
------------------------------------------------------------------------------
```

- `binreg` command - useful for fitting binary regression models with alternate links

```
. binreg chd69 arcus, or
------------------------------------------------------------------------------
             |                 EIM
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       arcus |   .4918141   .1342298     3.66   0.000     .2287285    .7548997
       _cons |  -2.599052   .0837963   -31.02   0.000    -2.76329   -2.434814
------------------------------------------------------------------------------
```

Example: dependence of CHD risk (P) on age (continuous) in WCGS

Assume linearity of log odds with age holds:

```
. logistic chd69 age
Logistic regression                               Number of obs   =       3154
                                                  LR chi2(1)      =      42.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -869.17806                       Pseudo R2       =     0.0241
------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.077262   .0121756     6.58   0.000      1.05366    1.101392
       _cons |   .0026333   .0014465   -10.81   0.000     .0008973    .0077283
------------------------------------------------------------------------------
```

output with `coef` option - reports coefficients and 95% CIs:

```
. logistic chd69 age, coef
Logistic regression                               Number of obs   =       3154
                                                  LR chi2(1)      =      42.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -869.17806                       Pseudo R2       =     0.0241
------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0744226   .0113024     6.58   0.000     .0522703    .0965748
       _cons |  -5.939516    .549322   -10.81   0.000    -7.016167   -4.862865
------------------------------------------------------------------------------
```

Example: dependence of CHD risk (P) on age in WCGS

$$\log\left[\frac{P(\text{age})}{1-P(\text{age})}\right] = \beta_0 + \beta_1 \text{age}$$

$$= -5.939516 \quad +0.0744226 \quad \text{age}$$

- -5.939516 is the log odds of CHD among individuals with age = 0
- 0.0744226 is the log odds ratio comparing the odds CHD among individuals with age = x+1 (in years) to the odds CHD among individuals with age = x

risk associated with a fixed age (age = 55):

log-odds of CHD: -5.939516 + 0.0744226 × 55 = -1.846273

odds of CHD: exp(-1.846273) = 0.15782428

probability of CHD: 0.15782428 / (1 + 0.15782428) = 0.13631108

OR associated with a one year increase in age:

log odds ratio : (-5.939516 + 0.0744226*56) - (-5.939516 + 0.0744226*55) = 0.0744226

odds ratio : exp(0.0744226) = 1.077262

95% C.I. for odds ratio: (exp(0.0522703) , exp(0.0965748)) ; (1.05, 1.10)

OR associated with a ten year increase in age:

log odds ratio : 10 X 0.0744226= 0.744226

odds ratio : exp(0.744226) = 2.1048117

95% C.I. for odds ratio: (exp(10 X 0.0522703) , exp(10 X 0.0965748)) ; (1.69, 2.63)

Logistic models with multiple predictors

Example: Logistic model for dependence of CHD risk on SBP and cholesterol in WCGS

```
. logistic chd69 chol sbp if chol<600
Logistic regression                               Number of obs   =       3141
                                                  LR chi2(2)      =     110.21
                                                  Prob > chi2     =     0.0000
Log likelihood = -831.98952                       Pseudo R2       =     0.0621
------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        chol |   1.011765   .0014981     7.90   0.000     1.008833    1.014706
         sbp |   1.024741   .0039104     6.40   0.000     1.017106    1.032434
       _cons |   .0002261   .0001393   -13.63   0.000     .0000676    .0007562
------------------------------------------------------------------------------
```

Notes:

- Excludes 12 observations with missing `chol` and 1 obs. with `chol` ≥ 600.
- Assumption of linearity for both `chol` and `sbp` in log CHD odds must be checked before model is accepted as reasonable.
- centering & scaling of predictors for interpretability follows same approach used in linear regression

Logistic models with multiple predictors

Example: Logistic model for dependence of CHD risk on SBP and cholesterol in WCGS

```
. logistic chd69 chol sbp if chol<600, coef
Logistic regression                               Number of obs   =       3141
                                                  LR chi2(2)      =     110.21
                                                  Prob > chi2     =     0.0000
Log likelihood = -831.98952                       Pseudo R2       =     0.0621
------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        chol |   .0116966   .0014807     7.90   0.000     .0087945    .0145987
         sbp |   .0244401    .003816     6.40   0.000      .016961    .0319193
       _cons |  -8.394706   .6160962   -13.63   0.000    -9.602233    -7.18718
------------------------------------------------------------------------------
```

Notes:

- 0.0117 is the log-odds ratio for each unit increase in `chol` holding `sbp` fixed
- 0.0244 gives the log-odds ratio for each unit increase in `sbp` holding `chol` fixed
- -8.395 gives the log-odds of CHD among indiv. with `chol=0` & `sbp=0`
- model specifies linearity of log odds in both predictors:
  - the log-odds of CHD is linear in cholesterol for any specified value of SBP
  - the log-odds of CHD is linear in SBP for any specified value of cholesterol

Evaluating confounding using logistic regression

Example: Relationship between hypertension and exercise in HERS (subsample of 1055 women without diabetes at enrollment)

Goal: Investigate possibly causal association between exercise and the occurrence of hypertension at study baseline

- Hypertension measured as a binary indicator of enrollment SBP>140
- Exercise represented as a binary indicator of exercise ≥ 3 times / week in the year before enrollment

Approach:

- Construct a causal model (DAG) representing the relationship between the primary predictor and the outcome, considering the effects of potential confounding (and/or mediating) variables
- Identify a minimally sufficient set of variables (MSAS) to adjust for in assessing the association of interest - use regression to estimate an adjusted association
- Look for possible interactions between key predictors; assess nonlinearity and lack of overlap
- Choose a summary measure of causal association
- Summarize adjusted association

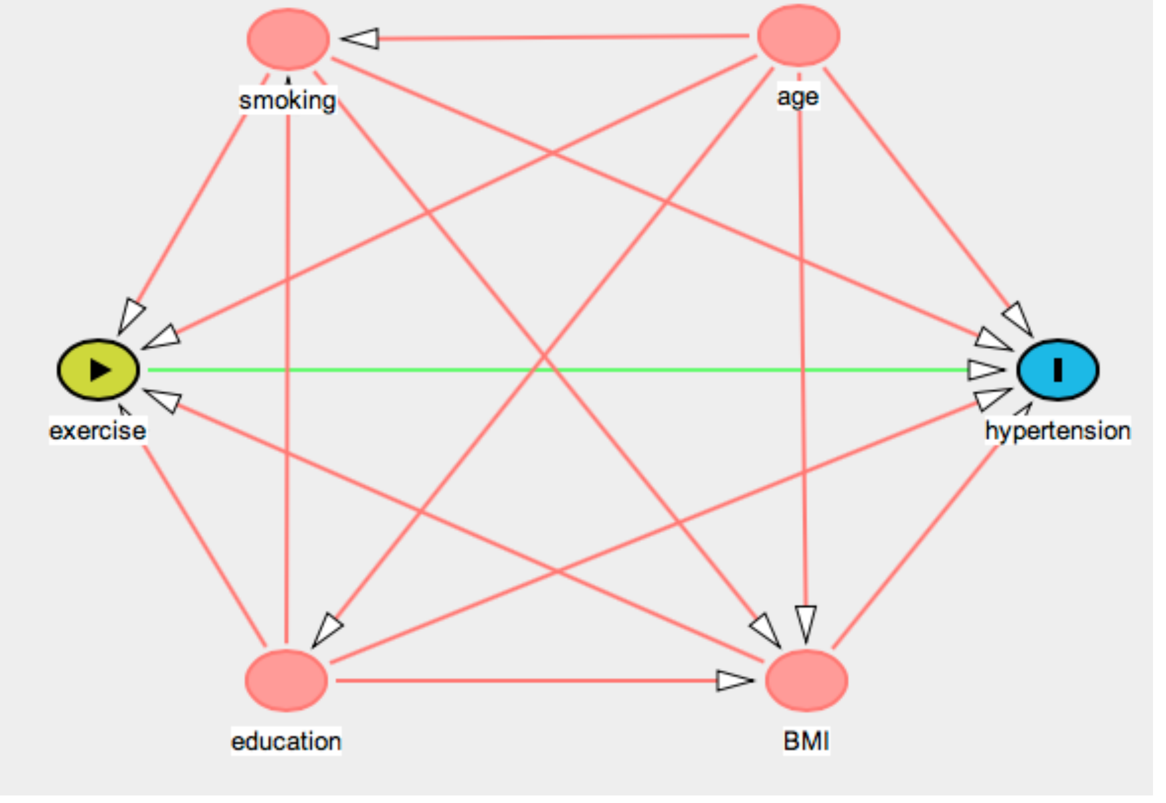
Evaluating confounding using logistic regression

Example: Relationship between hypertension and exercise in HERS

- Additional variables from HERS that may be relevant predictors include:
  - demographic (age, BMI, education)
  - health (cholesterol, glucose)
  - lifestyle (smoking, alcohol, diet, medications)

- Question: Could some or all of the apparent association between hypertension and exercise be explained by adjusting/controlling for these variables?

- Note: omitted redictors such as cholesterol & glucose could be hypothesized to play a mediating role - omitted in DAG



MSAS: smoking, age, education & BMI

Evaluating confounding using logistic regression

Example: Unadjusted association between hypertension and exercise in HERS

```
* unadjusted logistic model
. logistic hypt i.exer3
Logistic regression                                 Number of obs     =      1,055
                                                    LR chi2(1)        =      10.12
                                                    Prob > chi2       =     0.0015
Log likelihood = -667.38065                         Pseudo R2         =     0.0075
------------------------------------------------------------------------------
        hypt | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.exer3 |   .6526043   .0882531    -3.16   0.002     .5006567    .8506675
       _cons |   .5953608    .049477    -6.24   0.000     .5058733    .7006784
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

The unadjusted or *crude* odds ratio OR_{crude} is 0.65 (95% CI: $0.50, 0.85$)

Evaluating confounding using logistic regression

Example: Relationship between hypertension and exercise in HERS

```
* logistic model including MSAS predictors
. logistic hypt i.exer3 i.edu i.csmker agec bmisp*
Logistic regression                               Number of obs   =       1054
                                                  LR chi2(8)      =      53.82
                                                  Prob > chi2     =     0.0000
Log likelihood = -645.12775                       Pseudo R2       =     0.0400
------------------------------------------------------------------------------
        hypt | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.exer3 |   .7251488   .1048338    -2.22   0.026     .5462231    .9626849
       1.edu |   .7278707   .1006952    -2.30   0.022     .5550055    .9545776
    1.csmker |   1.210394   .2428564     0.95   0.341     .8168453    1.793552
        agec |   1.060382   .0113939     5.46   0.000     1.038284    1.082951
      bmisp1 |   .9541977   .0734822    -0.61   0.543     .8205174    1.109657
      bmisp2 |   1.735852   1.197488     0.80   0.424     .4490619    6.709952
      bmisp3 |    .205744   .5854537    -0.56   0.578     .0007784    54.38248
      bmisp4 |   3.052587   10.39188     0.33   0.743     .0038631    2412.131
       _cons |   .3695945   .1793267    -2.05   0.040     .1427986    .9565928
------------------------------------------------------------------------------
```

agec (in years at enrollment, centered) bmi (body mass index (kg/m^2), centered, and represented as a cubic spline)
edu (>12 yrs education) csmker (smoker at enrollment)
Note: HDL cholesterol & impaired fasting glucose (IFG) not included because of their possible mediating effects

The adjusted or *conditional* odds ratio OR_c is 0.73 (95% CI: $0.55, 0.96$)

Note: assume no other significant nonlinearities / interactions / non-overlap detected.

Evaluating confounding using logistic regression

Example: Relationship between hypertension and exercise in HERS

The adjusted *OR* is closer to the null value (one) than the unadjusted (crude) *OR*:

	OR	95% CI	*p*-value (Wald)
OR_{crude}	0.65	0.50, 0.85	0.002
OR_c	0.73	0.55, 0.96	0.026

- Recall that for the linear regression model, differences between an adjusted and unadjusted (crude) regression coefficient could be related directly to confounding, *and:*

- the adjusted (conditional) effect could be equated to the corresponding estimated marginal effect (e.g. obtained using `margins` in Stata)
  - we interpreted equality of the adjusted (marginal) effect and the unadjusted effect to indicate lack of confounding

- *Questions*:
  - can the adjusted/conditional odds ratio OR_c for the effect of exercise in this example be interpreted as accounting for the confounding influence of the other variables in the model?
  - does the conditional OR_c equal the corresponding marginal OR_m?

Evaluating confounding using logistic regression

Collapsibility and odds ratios

- The conditional odds ratio (OR_c) for the effect of exercise, given fixed values of the other included predictors (and assuming no interaction)
  - has a valid, *individual* causal interpretation if
    1. sample is representative of target population
    2. no unmeasured confounders
    3. correctly specified model

- The marginal odds ratio (OR_m) is defined as the odds ratio for the effect of exercise on hypertension obtained by averaging over the values of the model-predicted outcome probabilities (e.g. using `margins` in Stata)
  - has a *population* causal interpretation if
    1. sample is representative of target population
    2. no unmeasured confounders, and acceptable overlap w.r.t. exer. groups
    3. correctly specified model and appropriate analysis

- In general, $OR_c \neq OR_m$ (even when confounding is absent) - reflecting *non-collapsibility* of the odds ratio
  - By contrast, RR & RD are *collapsible* measures
  - Comparison of OR_c & OR_m can be useful in evaluating effects of non-collapsibility

Example: Artificial data illustrating non-collapsibility of OR
no confounding present (C is balanced with X) in this example

$C = 1$	Y=1	Y=0	
X=1	50	50	100
X=0	20	80	100

OR = 4.0

$C = 0$	Y=1	Y=0	
X=1	80	20	100
X=0	50	50	100

OR = 4.0

Pooled (crude)	Y=1	Y=0	
X=1	130	70	200
X=0	70	130	200

OR = 3.45

```
. cs Y X, by(C) or

               C |        OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               0 |          4      2.143427    7.460138            5 (Cornfield)
               1 |          4      2.143427    7.460138            5 (Cornfield)
-----------------+-------------------------------------------------
           Crude |    3.44898      2.288457    5.198021
    M-H combined |          4      2.566617    6.233887
-------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.000  Pr>chi2 = 1.0000

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =     39.36
                                                Pr>chi2 =    0.0000
```

- Non-collapsibility: marginal (crude) *OR* ≠ conditional (combined) *OR*, even when no confounding present
- In the corresponding logistic regression model, same result holds:

Ref: Jewell NP. Statistics for Epidemiology. 2003: Chapman and Hall/CRC.

Example: Previous example analyzed via logistic regression

```
. logistic Y i.X i.C
------------------------------------------------------------------------------
           Y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         1.X |          4   .9055385     6.12   0.000     2.566617    6.233887
         1.C |        .25   .0565962    -6.12   0.000     .1604136     .389618
       _cons |          1     .17943     0.00   1.000     .7035078    1.421448
------------------------------------------------------------------------------
. quietly margins X, post      <---- post option to margins request results be saved for post-estimation
. nlcom (_b[1.X]/(1-_b[1.X])) / (_b[0.X]/(1-_b[0.X]))
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   3.448979    .686414     5.02   0.000     2.103633    4.794326
------------------------------------------------------------------------------
```

$OR_c \neq OR_m$

Note: the use of `margins` will be covered further below

Conditional and marginal effects from the logistic model

The logistic regression model for the dependence of binary predictors X_1 and X_2 on a binary outcome Y is nonlinear on the probability scale:

$$\Pr\left[Y=1|X_1=x_1,\ X_2=x_2\right]=P\left(x_1,\ x_2\right)=\frac{\exp\left[\beta_0+\beta_1x_1+\beta_2x_2\right]}{\left(1+\exp\left[\beta_0+\beta_1x_1+\beta_2x_2\right]\right)}$$

β_1 estimates the conditional log OR_c for the effect of X_1 on Y (i.e. *conditional* on the value of X_2)

The *marginal probability* for Y=1 conditional on X_1 is

$$P(x_1)=\Pr\left[Y=1|X_1=x_1\right]=\sum_{j=0,\,1}\Pr\left[Y=1|X_1=x_1,X_2=j\ \right]\times\Pr\left[X_2=j\right]$$

The corresponding *marginal odds ratio* is $$OR_m=\frac{P(1)}{(1-P(1))}/\frac{P(0)}{(1-P(0))}$$

For the logistic model: $OR_m \neq e_1^{\beta}$, except under certain conditions (e.g. $\beta_2 = 0$)

Example: Artificial example repeated with risk difference regression model

```
. * risk difference regression model (to be discussed further in lecture 11)
. binreg Y i.X i.C, rd
------------------------------------------------------------------------------
             |                 EIM
           Y | Risk Diff.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         1.X |         .3   .0452769     6.63   0.000     .2112589    .3887411
         1.C |        -.3   .0452769    -6.63   0.000    -.3887411   -.2112589
       _cons |         .5    .041687    11.99   0.000      .418295     .581705
------------------------------------------------------------------------------

. quietly margins X, post
. lincom _b[1.R] - _b[0.R]

 ( 1)  - 0bn.R + 1.R = 0

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |         .3   .0452769     6.63   0.000     .2112589    .3887411
------------------------------------------------------------------------------
```

Risk difference is a collapsible measure $RD_c = RD_m$

Note: The linear model may present estimation issues (covered in lecture 11)

Conditional and marginal effects from the risk difference model

The *linear* risk difference regression model for the dependence of a binary outcome Y on binary predictors X_1 and X_2:

$$\Pr\left[Y=1|X_1=x_1,\ X_2=x_2\right]=P\left(x_1,\ x_2\right)=\beta_0+\beta_1x_1+\beta_2x_2$$

β_1 estimates the *conditional risk difference* (RD_c) for the effect of X_1 on Y (*conditional* on the value of X_2)

The *marginal probability* for Y=1 conditional on X_1 is

$$P(x_1)=\Pr\left[Y=1|X_1=x_1\right]=\sum_{j=0,\,1}\Pr\left[Y=1|X_1=x_1,X_2=j\ \right]\times\Pr\left[X_2=j\right]$$

The corresponding *marginal risk difference* is $RD_m=P(1)-P(0)$

For this (linear) model, the conditional and marginal effects coincide: $RD_m=\beta_1$

Note: A similar result holds for the relative risk regression model: $RR_m=e^{\beta_1}$

Example: Artificial example repeated with relative risk regression model

$$\log\left[\Pr\left(Y=1|X_1=x_1,\ X_2=x_2\right)\right]=\beta_0+\beta_1x_1+\beta_2x_2$$

```
. * relative risk regression model (to be discussed further in lecture 11)
. binreg Y i.X i.C, rr
------------------------------------------------------------------------------
             |                 EIM
           Y | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         1.X |   1.778207   .1820147     5.62   0.000     1.454972    2.173252
         1.C |   .5623643   .0575628    -5.62   0.000       .46014    .6872986
       _cons |   .4579834   .0441135    -8.11   0.000     .3791937    .5531443
------------------------------------------------------------------------------
. quietly margins X, post
. nlcom _b[1.X] / _b[0.X]
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   1.778207   .1820147     9.77   0.000     1.421465    2.134949
------------------------------------------------------------------------------
```

Relative risk is a collapsible measure $RR_c = RR_m$

Conditional and marginal effects from the logistic model

The contrast between the crude odds ratio OR_{crude} for the (unadjusted) association between outcome and primary predictor, and the corresponding conditional (adjusted) odds ratio OR_c can be decomposed into non-collapsibility & confounding components :

$$\log OR_{crude} - \log OR_c = [\log OR_m - \log OR_c] + [\log OR_{crude} - \log OR_m]$$

This suggests that computing both marginal and conditional estimates may help in evaluating importance on non-collapsibility (see reference below for details)

Reference: (Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Stat Methods Med Res*; 2013.)

Estimating the marginal effect in the HERS example

- Conditional odds ratio for the effect of exercise from HERS (slide 16)

```
------------------------------------------------------------------------------
        hypt | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.exer3 |   .7251488   .1048338    -2.22   0.026     .5462231    .9626849
```

- Marginal outcome probabilities for exercise/no exercise

```
. margins exer3, post
Predictive margins                                Number of obs   =       1054
Model VCE    : OIM
Expression   : Pr(hypt), predict()
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exer3 |
          0  |   .3624135   .0192471    18.83   0.000     .3246899    .4001371
          1  |   .2946184   .0223288    13.19   0.000     .2508547    .3383821
------------------------------------------------------------------------------
```

- `margins` uses the previously fitted model to estimate the average outcome probabilities for the entire sample under exercise/no exercise
- marginal probabilities can be used to estimate marginal measures of association

Evaluating confounding using logistic regression

Example: Relationship between hypertension and exercise in HERS

- Marginal OR_m, RR_m & RD_m for the effect of exercise (from marginal outcome probabilities estimated under exercise=$1/0$)

```
* marginal OR
. nlcom (_b[1.exer3]/(1-_b[1.exer3])) / (_b[0.exer3]/(1-_b[0.exer3]))
       _nl_1:  (_b[1.exer3]/(1-_b[1.exer3])) / (_b[0.exer3]/(1-_b[0.exer3]))
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .7348022   .1023049     7.18   0.000     .5342881    .9353162
------------------------------------------------------------------------------
* marginal RR
. nlcom (_b[1.exer3] / _b[0.exer3])
       _nl_1:  _b[1.exer3] / _b[0.exer3]
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .8129343   .0769937    10.56   0.000     .6620295    .9638391
------------------------------------------------------------------------------

* marginal RD - equivalent to using command: margins, dydx(exer3)
. lincom (_b[1.exer3] - _b[0.exer3])
 ( 1)  - 0bn.exer3 + 1.exer3 = 0
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0677951    .030205    -2.24   0.025    -.1269959   -.0085943
------------------------------------------------------------------------------
```

Evaluating confounding using logistic regression

Example: Relationship between hypertension and exercise in HERS

Summary:

	OR	95% CI
unadjusted (OR_{crude})	0.65	$0.50, 0.85$
adjusted (OR_c)	0.73	$0.55, 0.96$
marginal (OR_m)	0.74	$0.53, 0.93$

- Observed difference between adjusted and unadjusted ORs for exercise ($0.72, 0.65$) a very small degree of confounding may be operating
- OR_m & OR_c estimates similar in this case (non-collapsibility effect is negligible)
- Note that if BMI was a mediator of this relationship, the adjusted conditional effect of exercise would represent a direct effect

Evaluating confounding using logistic regression

Summary:

- Conditional odds ratio (OR_c) from an adjusted logistic regression model cannot generally be equated with a corresponding average/marginal causal effect measure because of non-collapsibility
  - The OR_c is a valid effect measure, but must be interpreted conditionally
  - Approximates a marginal RR only in the case of rare outcomes
  - Because of non-collapsibility, methods to assess degree of confounding (and mediation effects) based on changes in coefficients unreliable for logistic models

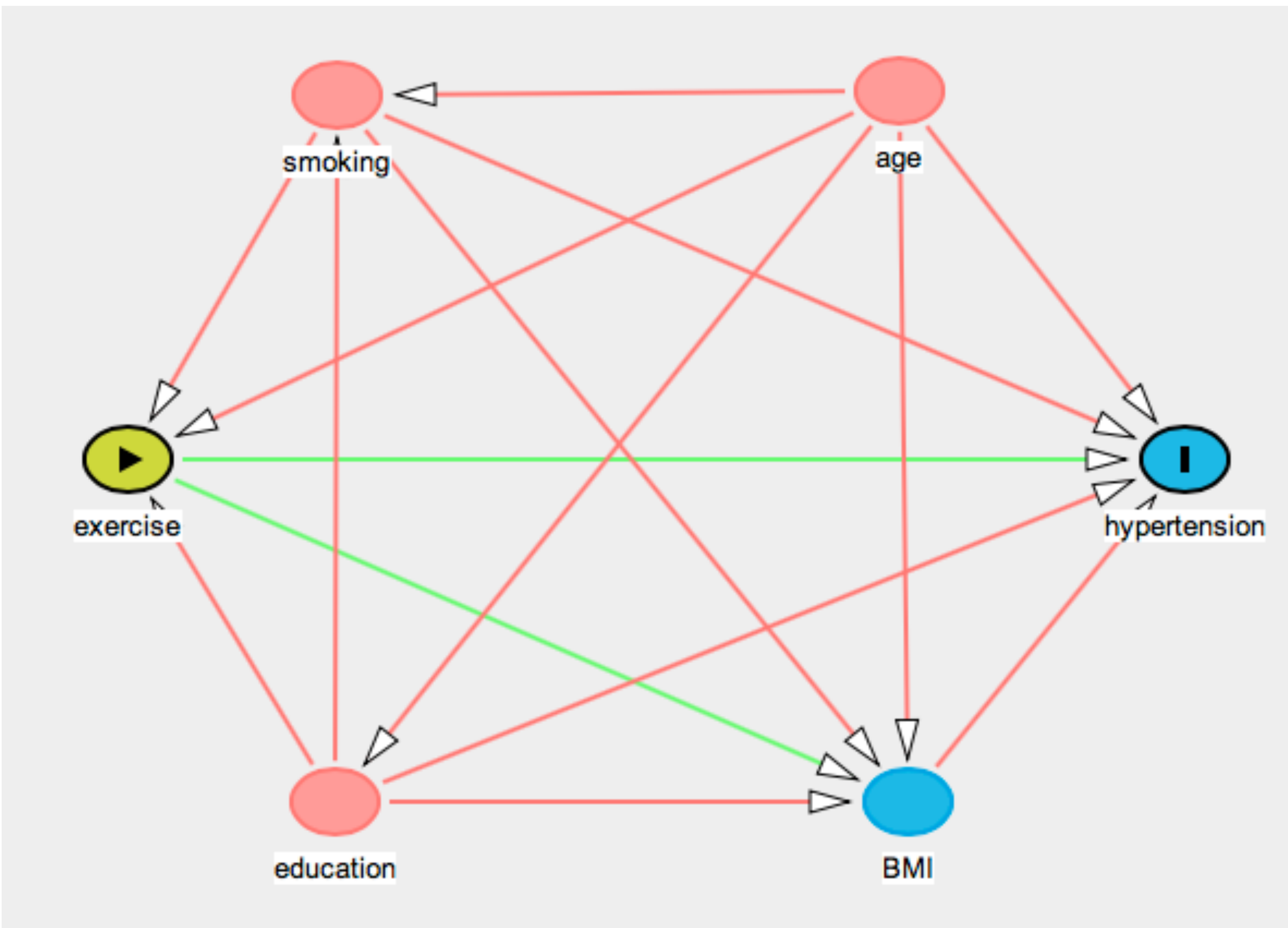
- The `margins` can be used to get estimates of marginal OR, RR & RD for logistic models
  - Log-relative risk and risk difference regression models can be used as well, but attention must be paid to estimation issues (lecture 11)

- Strategies for evaluating confounding otherwise unchanged from recommendations made for linear models

Evaluating mediation using logistic regression

Example: Relationship between hypertension and exercise in HERS

- Consider possible mediation role of BMI



Note: see section 4.5.4 in textbook for discussion of BMI as a possible mediator of exercise in HERS

Evaluating mediation using logistic regression

Example: Relationship between hypertension and exercise in HERS

- Three steps for evaluating mediation of exercise effect by BMI (analysis should be based same set of observations):

  1. evaluate adjusted assoc. between exercise and BMI
  2. evaluate adjusted assoc. between exercise and hypertension
  3. evaluate adjusted assoc. between exercise and hypertension controlling for BMI

- The OR for exercise from the 3rd model estimates a ("controlled") direct effect of exercise not mediated by BMI (and conditional on the other predictors)

- Estimating separate indirect and direct components of mediating effects is problematic due to non-collapsibility of ORs (except for rare outcomes - VanderWeele TJ, Vansteelandt S. Odds Ratios for Mediation Analysis for a Dichotomous Outcome. Am J Epidemiol. 2010; 172:1339–1348.)

- Impact of confounders of the outcome-exposure and outcome-mediator relationships need to be considered (lecture 4).
  - Variables with dual confounding/mediating roles require special methods

Evaluating mediation using logistic regression

Example: Relationship between hypertension and exercise in HERS

- Evaluating conditions for role of BMI as a possible mediator:

```
. regress bmi i.exer3 i.edu i.csmker agec

      Source |       SS       df       MS              Number of obs =    1054
-------------+------------------------------           F(  4,  1049) =   23.12
       Model |  2356.56238     4  589.140594           Prob > F      =  0.0000
    Residual |  26733.6582  1049  25.4848982           R-squared     =  0.0810
-------------+------------------------------           Adj R-squared =  0.0775
       Total |  29090.2206  1053  27.6260404           Root MSE      =  5.0483

------------------------------------------------------------------------------
         bmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.exer3 |  -2.442347   .3238882    -7.54   0.000     -3.07789   -1.806805
       1.edu |  -.3196396   .3169374    -1.01   0.313    -.9415429    .3022638
    1.csmker |   -1.78384   .4640269    -3.84   0.000    -2.694366   -.8733131
        agec |  -.1283467    .023634    -5.43   0.000    -.1747219   -.0819714
       _cons |    29.2754   .2547791   114.91   0.000     28.77547    29.77533
------------------------------------------------------------------------------
```

Evaluating mediation using logistic regression

Example: Relationship between hypertension and exercise in HERS

- Evaluating conditions for role of BMI as a possible mediator:

```
. logistic hypt i.exer3 i.edu i.csmker agec

Logistic regression                               Number of obs   =       1054
                                                  LR chi2(4)      =      41.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -651.35392                       Pseudo R2       =     0.0308

------------------------------------------------------------------------------
        hypt | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.exer3 |   .6491883   .0911016    -3.08   0.002     .4930833    .8547145
       1.edu |   .7138105   .0978588    -2.46   0.014      .545618      .93385
    1.csmker |   1.139427   .2241928     0.66   0.507     .7748269    1.675592
        agec |   1.054058   .0109989     5.05   0.000      1.03272    1.075838
       _cons |    .659734   .0699621    -3.92   0.000      .535923    .8121484
------------------------------------------------------------------------------
```

- *OR* of 0.65 estimates overall effect of exercise

Evaluating mediation using logistic regression

Example: Relationship between hypertension and exercise in HERS

- Evaluating conditions for role of BMI (treated as a linear term) as a possible mediator:

```
. logistic hypt i.exer3 i.edu i.csmker agec bmic

Logistic regression                               Number of obs   =       1054
                                                  LR chi2(5)      =      52.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -645.98359                       Pseudo R2       =     0.0388

------------------------------------------------------------------------------
        hypt | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.exer3 |   .7192284   .1037521    -2.28   0.022     .5420969     .954238
       1.edu |    .722425   .0994817    -2.36   0.018      .551541    .9462539
    1.csmker |   1.227736   .2445182     1.03   0.303     .8309558    1.813977
        agec |   1.060745   .0113766     5.50   0.000      1.03868    1.083279
        bmic |    1.04369   .0136563     3.27   0.001     1.017264    1.070802
       _cons |   .6179738   .0672956    -4.42   0.000     .4992023    .7650037
------------------------------------------------------------------------------
```

- *OR* of 0.72 estimates (conditional) direct effect of exercise *not* mediated through BMI
- Because of non-collapsibility, this can't be equated to a marginal direct effect

Evaluating mediation using logistic regression

Example: Relationship between hypertension and exercise in HERS

- Estimating marginal controlled direct effect (CDE) for individuals with BMI at the mean value, based on results from last model:

```
. * Run margins at a fixed level of BMI (in this case, the mean BMI)
. quietly margins exer3, at(bmic = 0) post
. nlcom (_b[1.exer3]/(1-_b[1.exer3])) / (_b[0.exer3]/(1-_b[0.exer3]))
------------------------------------------------------------------------------
     OR      |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .7273468   .1015144     7.16   0.000     .5283822    .9263113
------------------------------------------------------------------------------

. nlcom (_b[1.exer3] / _b[0.exer3])
------------------------------------------------------------------------------
     RR      |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .8071434   .0767328    10.52   0.000     .6567499    .9575369
------------------------------------------------------------------------------

. lincom (_b[1.exer3] - _b[0.exer3])
------------------------------------------------------------------------------
     RD      |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0699291   .0302295    -2.31   0.021    -.1291778   -.0106803
```

- CDE *OR* (marginal) estimate is similar (in this case) to conditional direct effect *OR* on previous slide

Causal mediation analysis: introduction

- Causal definitions of direct and indirect effects are based on averages of potential (counterfactual) outcomes indexed by both the intervention/exposure *and* the mediator: $Y(e, m)$

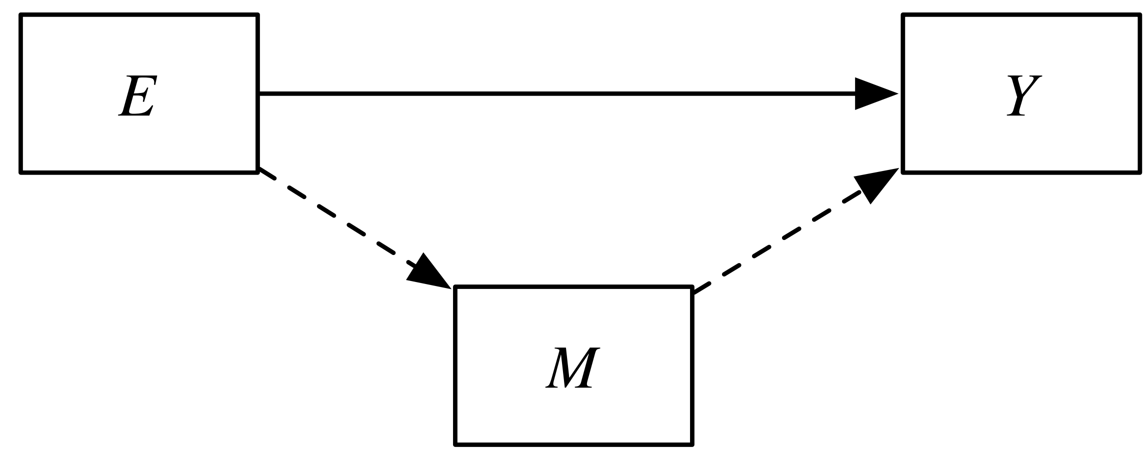
- $Y(e, m)$ is the potential outcome observed in an individual with intervention/exposure set at E=e and mediator set at M=m

- Examples of causal effect measures for mediation:

  - controlled direct effect

  - natural direct effect

  - natural indirect effect



- Assumptions for valid estimation address confounding of the relationships of both E & M to Y

Causal mediation analysis

- The causal controlled direct effect (CDE) is the average causal effect of exposure (E) on outcome (Y) for a fixed value m of a mediator M
  - Causal controlled direct effect evaluated at M=m :

$$E[Y(1,m)] - E[Y(0,m)]$$

  - Can be estimated with `logistic` & `margins` as illustrated above

- The natural direct effect (NDE) is the average causal effect of exposure, with the "potential" mediator set at "natural" values under exposure/no exposure for everyone in the population:

$$E[Y(1,m_e)] - E[Y(0,m_e)]$$

  - Requires definition of a *potential mediator* $M(e) = m_e$ for the value of M under a particular assigned level of exposure e $(0, 1)$
  - NDE & CDE are equivalent in the case where exposure does not interact with the mediator
  - Estimation requires special software (`medeff` package in Stata)

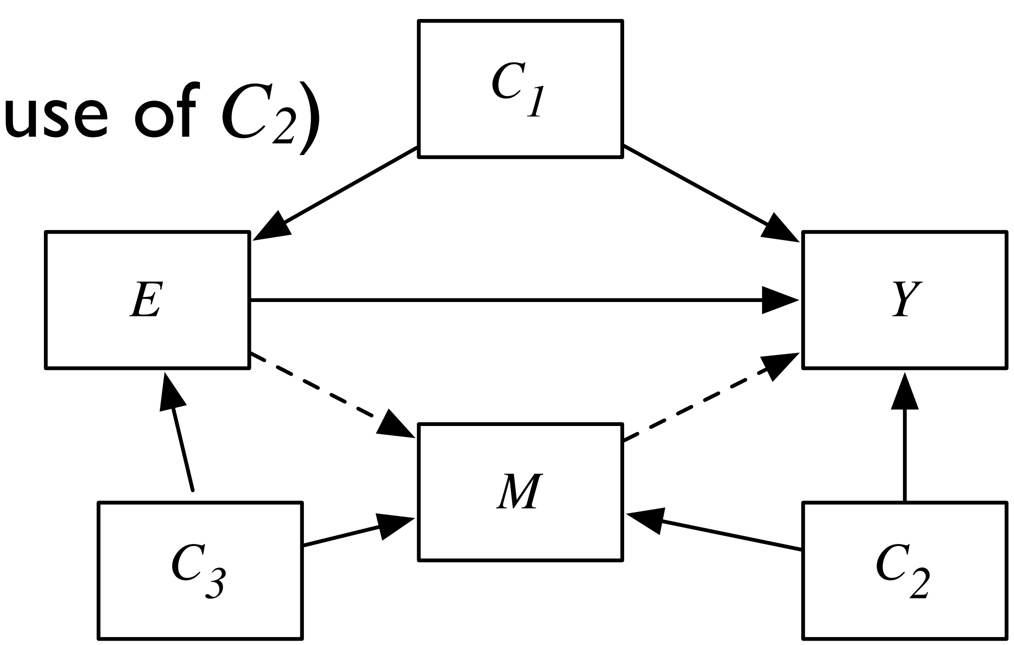
- The natural indirect effect (NIE) compares average potential outcomes for the population assuming all are exposed under two conditions: (*i*) the mediator set to its value under exposure, (*ii*) the mediator is set to its value under no exposure.
    - NIE for effect of mediator under exposure (e=1):

$$E[Y(1,m_1)] - E[Y(1,m_0)]$$

    - Can also be estimated under the *no* exposure condition (e=0)
- The NIE is also called the average causal mediated effect (ACME)
- The total causal effect is the sum of the NDE and NIE

Assumptions required for CDE, NDE & NIE

- The following two assumptions required for CDE, NDE & NIE:
  - No unobserved confounders C_1 of the exposure-outcome relationship
  - No unobserved confounders C_2 of the mediator-outcome relationship
- These additional assumptions are also required for NDE & NIE:
  - No unobserved confounders C_3 of the exposure-mediator relationship
  - No causal intermediates (e.g. E also a cause of C_2)



Using `medeff` to assess mediation for a binary outcome and binary predictor (effects are estimated as risk differences)

```
. medeff (regress bmic exer3 edu csmker agec) (logit hypt exer3 edu csmker agec bmic),
mediate(bmic) treat(exer3)
Using 0 and 1 as treatment values

------------------------------------------------------------------------------------
       Effect                   |  Mean             [95% Conf. Interval]
--------------------------------+---------------------------------------------------
       ACME1                    | -.0209207      -.0355808      -.0080683
       ACME0          NIE       | -.0231623      -.0379451      -.0094944
       Direct Effect 1          | -.0682403      -.1265449      -.0106466
       Direct Effect 0     NDE  | -.0704819      -.1297094      -.0110392
       Total Effect             | -.0914026      -.1448045      -.0374986
       % of Total via ACME1     |  .2260153       .1444758       .5579088
       % of Total via ACME0     |  .2502322        .159956       .6176871

       Average Mediation        | -.0220415      -.0366915      -.0088468
       Average Direct Effect    | -.0693611      -.1281134      -.0108416
       % of Tot Eff mediated    |  .2381238       .1522159       .5877979
------------------------------------------------------------------------------------
```

- Model for the mediator required to estimate $M(e)$
- Equality of natural direct effects under two exposure scenarios implies lack of exposure-mediatior interaction, which in turn implies NDE=CDE.
- Average direct effect estimate can be interpreted as a CDE in this case. (slide 35)

`medeff` reference: Hicks R, Tingley D. Causal mediation analysis, 2011; Stata Journal; 11:4,605-619

http://www.stata-journal.com/article.html?article=st0243

Evaluating mediation using logistic regression

Summary:

- Controlled direct effects can be estimated using `margins` as outlined above
  - confounding of mediator-outcome relationship needs to be identified and controlled
- The `medeff` mediation package (or an alternative) should be used to estimate indirect effects and % of total effect mediated
- Presence of causal intermediates (e.g. time dependent confounders) requires special methods (e.g. use of inverse probability weights and marginal structural models)
- More details in Biostat 215 (offered in Spring)

Evaluating interaction using logistic regression

Interaction between two predictors: the magnitude of the association of each predictor with the outcome, as measured by its regression coefficient, varies according to the value of the other predictor.

- As in linear regression, interaction between two predictors is evaluated via incorporation of a constructed variable defined as the product of the two predictors into the model including both predictors
- If the coefficient (odds ratio) of the included product variable is significantly different than zero (one), we conclude that interaction is present
- Interaction assessed via the logistic model is additive on the log-odds scale and multiplicative on the odds scale
  - the logistic model *does not* provide a direct measure of additive interaction on the odds/probability/risk scales
- Presence of interaction does not necessarily imply a causal interpretation
  - interaction terms may be required to improve model fit and/or to adequately address confounding effects of adjustment variables, in the absence of causality considerations

Interpreting two-way interaction between two binary predictors

Example: interaction between binary indicator of impaired fasting glucose (IFG) and BMI in HERS

- Additional analyses of the previous HERS example revealed an interaction between IFG and BMI

- We will explore this further here, focusing only on key variables (to simplify interpretation) and initially considering a binary indicator of high vs. low values of BMI

Note: treating BMI and glucose levels as continuous variables may be more appropriate in practice.

Interaction between two binary predictors

Simple example: interaction between binary indicators of IFG and BMI ≥ 35 in HERS

```
* create binary indicator of BMI greater than or equal to 35
. gen bmihi = bmi
. recode bmihi min/34.9999=0 35/max=1
* logistic model including interaction using i. and ## syntax:
. logistic hypt i.ifg##i.bmihi
. Logistic regression                          Number of obs     =      1,055
                                              LR chi2(3)        =      16.06
                                              Prob > chi2       =     0.0011
Log likelihood = -664.41499                   Pseudo R2         =     0.0119
------------------------------------------------------------------------------
        hypt | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.ifg |   2.603175   .7056691     3.53   0.000     1.530239    4.428404
     1.bmihi |   1.671196    .393736     2.18   0.029     1.053133    2.651988
   ifg#bmihi |
         1 1 |   .2163763   .1301662    -2.54   0.011     .0665503    .7035091
       _cons |   .4552846    .032823   -10.91   0.000     .3952911    .5243833
_______________________________________
Note: _cons estimates baseline odds
```

- Confidence interval for `i.ifg#bmihi` excludes null value (1), implying that we have evidence that an interaction between `ifg` and `bmihi` exists.
- Interpretation of Z (Wald) test statistic for testing the null hypothesis of no interaction (H: $\beta_3 = 0$): $Z = -1.530736/0.6015732 = -2.54$ (coeff. on next page)
- Referring to the standard normal distribution, the probability of a more extreme value of Z under the (two-sided) null hypothesis is 0.011.

Interaction between two binary predictors

Example: interaction between IFG & BMI in HERS (*continued*)

```
. * Coefficient form of the model:
. logistic hypt i.ifg##i.bmihi, coef
Logistic regression                             Number of obs     =      1,055
                                                LR chi2(3)        =      16.06
                                                Prob > chi2       =     0.0011
Log likelihood = -664.41499                     Pseudo R2         =     0.0119

------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.ifg |   .9567317   .2710802     3.53   0.000     .4254242    1.488039
     1.bmihi |   .5135393   .2356014     2.18   0.029     .0517691    .9753095
             |
   ifg#bmihi |
        1 1  |  -1.530736   .6015732    -2.54   0.011    -2.709798   -.3516745
             |
       _cons |  -.7868327   .0720933   -10.91   0.000    -.9281329   -.6455324
------------------------------------------------------------------------------
```

Interaction between two binary predictors

- Logistic model equation (log-odds form):

$$\log\left[\frac{P\left(\texttt{1.ifg},\texttt{1.bmihi},\texttt{1.ifg\#1.bmihi}\right)}{1-P\left(\texttt{1.ifg},\texttt{1.bmihi},\texttt{1.ifg\#1.bmihi}\right)}\right]=\beta_0+\beta_1\texttt{1.ifg}+\beta_2\texttt{1.bmihi}+\beta_3\texttt{1.ifg\#1.bmihi}$$

- Calculation of component log-odds ratios:

$$\log\left[\frac{P(1,0,0)}{1-P(1,0,0)}\right]-\log\left[\frac{P(0,0,0)}{1-P(0,0,0)}\right]=\beta_1=0.9567317$$

$$\log\left[\frac{P(0,1,0)}{1-P(0,1,0)}\right]-\log\left[\frac{P(0,0,0)}{1-P(0,0,0)}\right]=\beta_2=0.5135393$$

$$\log\left[\frac{P(1,1,1)}{1-P(1,1,1)}\right]-\log\left[\frac{P(0,1,0)}{1-P(0,1,0)}\right]=\beta_1+\beta_3=0.9567317-1.530736=-0.574$$

$$\log\left[\frac{P(1,1,1)}{1-P(1,1,1)}\right]-\log\left[\frac{P(1,0,0)}{1-P(1,0,0)}\right]=\beta_2+\beta_3=0.5135393-1.530736=-1.017$$

$$\log\left[\frac{P(1,1,1)}{1-P(1,1,1)}\right]-\log\left[\frac{P(0,0,0)}{1-P(0,0,0)}\right]=\beta_1+\beta_2+\beta_3=-0.060$$

- Interaction effect is *additive* on the log-odds scale

Example: log ORs for interaction between IFG & BMI in HERS (continued)

Constructing `lincom` statements:

IFG	BMIHI	_cons β_0	1.ifg β_1	1.bmihi β_2	1.ifg#1.bmihi β_3	
Yes	No	1	1	0	0	β_1
No	No	1	0	0	0	
Diff		0	1	0	0	

IFG	BMIHI	_cons β_0	1.ifg β_1	1.bmihi β_2	1.ifg#1.bmihi β_3	
No	Yes	1	0	1	0	β_2
No	No	1	0	0	0	
	Diff	0	0	1	0	

IFG	BMIHI	_cons β_0	1.ifg β_1	1.bmihi β_2	1.ifg#1.bmihi β_3	
Yes	Yes	1	1	1	1	$\beta_1+\beta_3$
No	Yes	1	0	1	0	
Diff		0	1	0	1	

IFG	BMIHI	_cons β_0	1.ifg β_1	1.bmihi β_2	1.ifg#1.bmihi β_3	
Yes	Yes	1	1	1	1	$\beta_2+\beta_3$
Yes	No	1	1	0	0	
	Diff	0	0	1	1	

Example: log ORs for interaction between IFG & BMI in HERS (continued)

Constructing `lincom` statements (*continued*):

<table>
<tr><td>IFG</td><td>BMIHI</td><td>_cons β_0</td><td>1.ifg β_1</td><td>1.bmihi β_2</td><td>1.ifg#1.bmihi β_3</td><td rowspan="4">$\beta_1+\beta_2+\beta_3$</td></tr>
<tr><td>Yes</td><td>Yes</td><td>1</td><td>1</td><td>1</td><td>1</td></tr>
<tr><td>No</td><td>No</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td></td><td>Diff</td><td>0</td><td>1</td><td>1</td><td>1</td></tr>
</table>

Example: ORs for interaction between IFG & BMI in HERS (*continued*)

```
. lincom 1.ifg
```
$\exp(\beta_1)$ *Note*: `lincom` after `logistic` defaults to ORs

hypt	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	2.603175	.7056691	3.53	0.000	1.530239	4.428404

```
. lincom 1.bmihi
```
$\exp(\beta_2)$

hypt	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	1.671196	.393736	2.18	0.029	1.053133	2.651988

```
. lincom 1.ifg + 1.ifg#1.bmihi
```
$\exp(\beta_1 + \beta_3)$

hypt	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.5632653	.3024928	-1.07	0.285	.1966023	1.613754

```
. lincom 1.bmihi + 1.ifg#1.bmihi
```
$\exp(\beta_2 + \beta_3)$

hypt	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.3616071	.2001561	-1.84	0.066	.1222029	1.070021

```
. lincom 1.ifg + 1.bmihi + 1.ifg#1.bmihi
```
$\exp(\beta 1 + \beta 2 + \beta 3)$

hypt	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.9413265	.4643066	-0.12	0.902	.3580038	2.475101

Example: log ORs for interaction between IFG & BMI in HERS (*continued*)

Note: `lincom` after `logit` returns coefficients (*ORs usually preferred*)

```
. lincom 1.ifg     β1
------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .9567317   .2710802     3.53   0.000     .4254242    1.488039
------------------------------------------------------------------------------
. lincom 1.bmihi   β2
------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .5135393   .2356014     2.18   0.029     .0517691    .9753095
------------------------------------------------------------------------------
. lincom 1.ifg + 1.ifg#1.bmihi    β1 + β3
------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.5740045   .5370343    -1.07   0.285    -1.626572    .4785633
------------------------------------------------------------------------------
. lincom 1.bmihi + 1.ifg#1.bmihi β2 + β3
------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -1.017197   .5535181    -1.84   0.066    -2.102072    .0676786
------------------------------------------------------------------------------
. lincom 1.ifg + 1.bmihi + 1.ifg#1.bmihi β1 + β2  + β3
------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0604652   .4932471    -0.12   0.902    -1.027212    .9062813
------------------------------------------------------------------------------
```

Interaction between two binary predictors

- Logistic model equation (odds form):

$$\frac{P\left(\texttt{1.ifg}, \texttt{1.bmihi}, \texttt{1.ifg\#1.bmihi}\right)}{1 - P\left(\texttt{1.ifg}, \texttt{1.bmihi}, \texttt{1.ifg\#1.bmihi}\right)} = \exp\left(\beta_0 + \beta_1 \texttt{1.ifg} + \beta_2 \texttt{1.bmihi} + \beta_3 \texttt{1.ifg\#1.bmihi}\right)$$

$$= e_0^{\beta} \times e^{\beta_1 \texttt{1.ifg}} \times e^{\beta_1 \texttt{1.bmihi}} \times e^{\beta_1 \texttt{1.ifg\#1.bmihi}}$$

- Calculation of component odds ratios

$$\frac{P\left(1,0,0\right)}{1 - P\left(1,0,0\right)} \Big/ \frac{P\left(0,0,0\right)}{1 - P\left(0,0,0\right)} = e^{\beta_1}$$

$$\frac{P\left(0,1,0\right)}{1 - P\left(0,1,0\right)} \Big/ \frac{P\left(0,0,0\right)}{1 - P\left(0,0,0\right)} = e^{\beta_2}$$

$$\frac{P\left(1,1,1\right)}{1 - P\left(1,1,1\right)} \Big/ \frac{P\left(0,1,0\right)}{1 - P\left(0,1,0\right)} = e^{\beta_1 + \beta_3}$$

$$\frac{P\left(1,1,1\right)}{1 - P\left(1,1,1\right)} \Big/ \frac{P\left(1,0,0\right)}{1 - P\left(1,0,0\right)} = e^{\beta_2 + \beta_3}$$

$$\frac{P\left(1,1,1\right)}{1 - P\left(1,1,1\right)} \Big/ \frac{P\left(0,0,0\right)}{1 - P\left(0,0,0\right)} = e^{\beta_1 + \beta_2 + \beta_3}$$

- Interaction effect is *multiplicative* on the odds scale

Interaction between two binary predictors

- `nlcom` can be used to compute interaction effects on the odds scale:
  - *example:* $\exp(\beta_1 + \beta_3)$

```
. nlcom exp(_b[1.ifg] + _b[1.ifg#1.bmihi])

       _nl_1:  exp(_b[1.ifg] + _b[1.ifg#1.bmihi])

------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .5632653   .3024928     1.86   0.063    -.0296096     1.15614
------------------------------------------------------------------------------
```

- *Note*: `lincom` result is preferred because the resulting confidence intervals are more accurate (note also that null hypotheses differ)

```
. lincom 1.ifg + 1.ifg#1.bmihi

 ( 1)  [hypt]1.ifg + [hypt]1.ifg#1.bmihi = 0

------------------------------------------------------------------------------
        hypt | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .5632653   .3024928    -1.07   0.285     .1966023    1.613754
------------------------------------------------------------------------------
```

- *Recommendation:* Use `lincom` when feasible

Reporting presence of interaction

- Detected interactions can be presented in tabular or figure form
  - results should specify interaction scale (multiplicative or additive)
    - presence/absence of additive interaction should be evaluated if causal interpretation is appropriate (discussed more below)
  - include estimates and associated 95% CIs; guidelines for significance same as for linear models

- *Qualitative interaction*:
  - direction of effect of one predictor differs between levels of the other
  - presence implies interaction occurs on *both* additive and multiplicative scales

Evaluating presence of additive interaction

- Additive interaction may be more biologically meaningful than multiplicative interaction in many contexts (Rothman, *Epidemiology: An Introduction*, 2002)
  - logistic regression *does not* provide a direct measure of additive interaction
    - recall that interaction is multiplicative on the odds scale
  - presence of multiplicative interaction doesn't imply presence of additive interaction
- Degree of additive interaction for 2 binary predictors can be estimated empirically by:

  $(P_{11} - P_{00}) - [(P_{10} - P_{00}) + (P_{01} - P_{00})] = P_{11} - P_{10} - P_{01} + P_{00}$

  (i.e. this $\neq 0$ if the combined effect of both factors does not equal the sum of their separate effects)

  - In the following linear model for binary predictors, this quantity is β_3 (i.e. the coef. for interaction)

$$P(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- Dividing by the background risk (P_{00}), this can expressed in terms of relative risks:
  - $\mathrm{RERI}_{RR} = RR_{11} - RR_{10} - RR_{01} + 1$ (RERI = rel. excess risk *due to* interaction)

    $\mathrm{RERI}_{RR} = 0$ implies absence of additive interaction

    $\mathrm{RERI}_{RR} \neq 0$ implies presence of additive interaction

Evaluating presence of additive interaction for logistic regression

- The presence of additive interaction on the probability scale can be evaluated based on odds ratio estimates from a fitted logistic model (binary predictors) as follows

$$\log\left[\frac{P(x_1,x_2)}{1-P(x_1,x_2)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

  - $\mathrm{RERI}_{OR} = OR_{11} - OR_{10} - OR_{01} + 1$

    $= \exp(\beta_1+\beta_2+\beta_3) - \exp(\beta_1) - \exp(\beta_2) + 1$

- For rare outcomes (≤10% in strata defined by predictor values), approximates RERI_{RR}
  - often useful in case control studies
- RERI_{RR} can be estimated directly for prospective/cross-sectional studies using a log relative risk model (example below; more detail in lecture 11)
- Methods for continuous predictors covered in the reference:

Ref: Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. Int J Epidemiol. 2012;41:514-20.

Example: logistic regression assessment of additive interaction between IFG & BMI in HERS

```
* Estimate RERI_OR using coefficients from previous logistic model

. nlcom exp(_b[1.ifg]+_b[1.bmihi]+_b[1.ifg#1.bmi]) - exp(_b[1.ifg]) - exp(_b[1.bmihi]) + 1

       _nl_1:  exp(_b[1.ifg]+_b[1.bmihi]+_b[1.ifg#1.bmi]) - exp(_b[1.ifg]) - exp(_b[1.bmihi]) + 1

------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |  -2.333044    .933796    -2.50   0.012     -4.16325   -.5028372
------------------------------------------------------------------------------
```

- indicates negative interaction on the additive scale
- likely not a reliable estimate of RERI_{RR} since outcome is not rare
- SE & CIs approximate - a bootstrap 95% CI would be better in smaller samples

Example: log relative risk regression assessment of additive interaction

```
* Estimate RERI_RR using a log relative risk model (including interaction)
. binreg hypt i.ifg##i.bmihi, rr

Generalized linear models                         No. of obs      =      1,055
Optimization     : MQL Fisher scoring             Residual df     =      1,051
                   (IRLS EIM)                     Scale parameter =          1
Deviance         =  1328.829981                   (1/df) Deviance =   1.264348
Pearson          =  1054.999988                   (1/df) Pearson  =   1.003806
------------------------------------------------------------------------------
             |                 EIM
        hypt | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.ifg |   1.733656   .2244061     4.25   0.000     1.345187    2.234309
             |
       bmihi |
   BMI >= 35 |   1.381173   .1887708     2.36   0.018       1.0566     1.80545
             |
   ifg#bmihi |
1#BMI >= 35  |   .4004749   .1549208    -2.37   0.018     .1876264    .8547842
             |
       _cons |   .3128492   .0154982   -23.46   0.000     .2839013    .3447487
---------------------------------------

. nlcom exp(_b[1.ifg]+_b[1.bmihi]+_b[1.ifg#1.bmi]) - exp(_b[1.ifg]) - exp(_b[1.bmihi]) + 1

       _nl_1:  exp(_b[1.ifg]+_b[1.bmihi]+_b[1.ifg#1.bmi]) - exp(_b[1.ifg]) - exp(_b[1.bmihi]) + 1

------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |    -1.1559   .4388867    -2.63   0.008    -2.016103   -.2956982
------------------------------------------------------------------------------
```

- indicates negative interaction on the additive scale
- doesn't work for case-control studies (approaches for relative risk regression covered in lecture 11)

Example: direct estimate of additive interaction between IFG & BMI in HERS

```
* Use the risk difference model for binary outcomes

. binreg hypt i.ifg i.bmihi i.ifg##i.bmihi, rd

Generalized linear models                      No. of obs      =      1,055
Optimization     : MQL Fisher scoring          Residual df     =      1,051
                   (IRLS EIM)                  Scale parameter =          1
Deviance         =  1328.829981                (1/df) Deviance =   1.264348
Pearson          =         1055                (1/df) Pearson  =   1.003806

Variance function: V(u) = u*(1-u)              [Bernoulli]
Link function    : g(u) = u                    [Identity]
                                               BIC             =  -5987.492
------------------------------------------------------------------------------
             |                 EIM
        hypt | Risk Diff.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.ifg |   .2295237   .0666862     3.44   0.001     .0988211    .3602263
             |
       bmihi |
    BMI >= 35 |   .1192496   .0571812     2.09   0.037     .0071764    .2313228
             |
   ifg#bmihi |
1#BMI >= 35  |  -.3616225   .1340767    -2.70   0.007    -.6244079   -.0988371
             |
       _cons |   .3128492   .0154982    20.19   0.000     .2824732    .3432251
------------------------------------------------------------------------------
```

- *Note*: Model can't be fitted in many situations (discussed in lecture 11)
  - multiple & continuous predictors
  - case-control studies

Interpreting interaction - causal effect modification

- Presence of a statistical interaction does not imply a causal interpretation
  - interactions of targeted interventions/exposures with additional variables can be estimated from logistic models (e.g. subgroups in randomized studies)
    - additional predictors can be interpreted as *effect modifiers*
- For a potential (counterfactual) binary outcome $Y(e)$ dependent on a binary exposure E=e, causal effect modification by a binary predictor X can be defined as follows:
  - *Effect modification* present if causal effect of E differs between levels of X
    - Additive:

$$\mathrm{E}\left[y(e=1)|\, X=1\right] - \mathrm{E}\left[y(e=0)|\, X=1\right] \neq \mathrm{E}\left[y(e=1)|\, X=0\right] - \mathrm{E}\left[y(e=0)|\, X=0\right]$$

    - Multiplicative: $$\frac{\mathrm{E}\left[y(e=1)|\, X=1\right]}{\mathrm{E}\left[y(e=0)|\, X=1\right]} \neq \frac{\mathrm{E}\left[y(e=1)|\, X=0\right]}{\mathrm{E}\left[y(e=0)|\, X=0\right]}$$

  - (Recall that for binary outcomes: $\mathrm{E}\left[Y(e=1)|\, X=1\right] = P\left[Y(e=1)|\, X=1\right]$)

- Example with `margins` provided below

Example: Multiplicative effect modification of the effect of exercise on hypertension at fixed BMI levels in HERS (including additional adjustment variables)

- Estimating marginal RDs of exercise (> 3 times/wk) at each level of BMI (low/high)

```
. * Run margins at each fixed level of bmihi
. quietly logistic hypt i.exer3##i.bmihi i.edu i.csmker agec
. quietly margins exer3, at(bmihi = 0) post
. lincom _b[1.exer3] - (_b[0.exer3]
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0749345   .0304127    -2.46   0.014    -.1345422   -.0153268
------------------------------------------------------------------------------


. quietly logistic hypt i.exer3##i.bmihi i.edu i.csmker agec
. quietly margins exer3, at(bmihi = 1) post
. lincom _b[1.exer3] - (_b[0.exer3]
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.2363081   .1141314    -2.07   0.038    -.4600016   -.0126146
------------------------------------------------------------------------------
```

- lack of overlap of 95% CIs implies presence of additive interaction (not a formal test) - in this case there is some overlap apparent
- Note that bootstrap CIs may be preferable (especially in small samples)

Causal interaction

- *Causal interaction* is evaluable only in the context of dual (joint) interventions
- For a potential (counterfactual) binary outcome $Y(e_1, e_2)$ dependent on a joint intervention $E_1 = e_1$ & $E_2 = e_2$, causal interaction can be defined as follows:
  - *Interaction between* E_1 & E_2 present if causal effect of E_1 differs between levels of E_2
    - Additive:

$$\mathrm{E}\left[Y(e_1 = 1, e_2 = 1)\right] - \mathrm{E}\left[Y(e_1 = 0, e_2 = 1)\right] \neq \mathrm{E}\left[Y(e_1 = 1, e_2 = 0)\right] - \mathrm{E}\left[Y(e_1 = 1, e_2 = 0)\right]$$

    - Multiplicative:

$$\frac{\mathrm{E}\left[Y(e_1 = 1, e_2 = 1)\right]}{\mathrm{E}\left[Y(e_1 = 0, e_2 = 1)\right]} \neq \frac{\mathrm{E}\left[Y(e_1 = 1, e_2 = 0)\right]}{\mathrm{E}\left[Y(e_1 = 0, e_2 = 0)\right]}$$

Reference on causal effect modification & interaction :

VanderWeele TJ. On the distinction between interaction and effect modification. *Epidemiology*. 2009;20:863-71.

Evaluating interaction using logistic regression

Summary:

- Interaction models for logistic regression are fitted as in linear models
  - interaction is multiplicative on the odds scale (additive in log-odds)
  - interactions may be included to improve model fit, or to address hypotheses (e.g. about targeted relationships or effects of interventions/exposures)

- Logistic model coefficients can be used to provide direct estimates of the effects of multiplicative interaction
  - interactions may be included to improve model fit
  - interaction summaries include presenting estimates and CIs of effects of one involved predictor over levels of the other
  - the possible presence of additive interaction can be evaluated using RERI_{OR}
  - if outcomes are not rare, fitting a relative risk regression model and using RERI_{RR} is preferable

- Interactions involving a target intervention/exposure can be interpreted as *effect modification* using marginal estimates

- Strategies for evaluating interaction otherwise unchanged from recommendations made for linear models

Likelihood ratio testing

Example: contribution of type A/B behavior pattern to predicting CHD in WCGS

Goal: use the likelihood ratio test to assess the contribution of behavior pattern to a model already controlling for `age`, `chol`, `sbp`, `bmi` and `smoke`.

- The `lrtest` procedure in Stata is used to compare likelihoods from nested models
- First, the full model including predictor(s) to be evaluated is fitted.
- Next, the likelihood from this model is saved in a named temporary location (e.g. "*A*") for further reference
- The nested model excluding the predictor(s) to be evaluated is fitted.
- The likelihood ratio test is then performed using the `lrtest` command and referencing the location (i.e. "*A*") of the likelihood from the full model.
- *Note:* The nested and full models must be base on the same set of observations for the LR test to make sense

Likelihood ratio testing

Example: contribution of type A/B behavior pattern to predicting CHD in WCGS study

```
* First, fit full model including behpat:
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke i.behpat
Logistic regression                               Number of obs   =       3141
                                                  LR chi2(8)      =     184.57
                                                  Prob > chi2     =     0.0000
Log likelihood =    -794.81                       Pseudo R2       =     0.1040
------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |    1.83375   .2198681     5.06   0.000     1.449707    2.319529
     chol_50 |   1.704097   .1301391     6.98   0.000     1.467201    1.979243
      sbp_50 |   2.463505   .5086518     4.37   0.000     1.643621    3.692369
      bmi_10 |   1.739414    .462034     2.08   0.037     1.033479    2.927551
       smoke |   1.830672   .2583097     4.29   0.000      1.38837    2.413882
      behpat |
          2  |   1.068257   .2363271     0.30   0.765     .6924157    1.648103
          3  |   .5141593   .1245593    -2.75   0.006     .3198064    .8266243
          4  |    .572071   .1826116    -1.75   0.080     .3060107    1.069457
------------------------------------------------------------------------------
* Save the log-likelihood in the temporary variable "A":
. estimates store A
* Note : the variables age_10, chol_50, sbp_50, bmi_10 represent centered and scaled
versions of the originals.
```

Likelihood ratio testing

Example: contribution of type A/B behavior pattern to predicting CHD in WCGS study (*continued*)

```
* Next, the desired "nested" model excluding predictor(s) of interest is fit, and
compared to the previous model using the lrtest command:

. logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke

Logistic regression                               Number of obs   =       3141
                                                  LR chi2(5)      =     159.80
                                                  Prob > chi2     =     0.0000
Log likelihood = -807.19249                       Pseudo R2       =     0.0901

------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   1.904989   .2268333     5.41   0.000     1.508471    2.405735
     chol_50 |   1.710974   .1297977     7.08   0.000     1.474584    1.985259
      sbp_50 |   2.623972   .5367142     4.72   0.000     1.757326    3.918016
      bmi_10 |   1.775994   .4680612     2.18   0.029     1.059518    2.976972
       smoke |   1.886037   .2643914     4.53   0.000     1.432933    2.482417
------------------------------------------------------------------------------
* perform likelihood ratio test comparing fitted model to full model (A):
. lrtest A

Likelihood-ratio test                                  LR chi2(3)  =     24.76
(Assumption: . nested in A)                            Prob > chi2 =    0.0000
```

Summary: procedure for likelihood ratio testing in nested models:

1. Fit the full model including all the variable(s) of interest
2. Issue the `estimates` command, using the `store` option to save results of the current model in a location labeled by the results with a name (e.g. `estimates store A`)
3. Fit the nested model excluding the variable(s) of interest
4. Issue the `lrtest` command, identifying the results of the full model fit in step 1 by passing the name specified in step 2 (e.g. `lrtest A`)

The `lrtest` procedure computes the likelihood ratio comparing the two models, and refers the results to the chi-squared distribution with degrees of freedom equal to the difference in number of degrees of freedom in the two fitted models. The results of the test evaluate the null hypothesis that the true values of coefficients from the additional variables in the full model are all zero.

Summary: procedure for likelihood ratio testing in nested models:

- the likelihood/log-likelihood always increases as variables are added to an existing model
- like the *F* test for linear regression models, the likelihood ratio (*LR*) test is used to compare two nested models
- the *LR* statistic comparing two nested models is computed as twice the difference between the log-likelihoods:

$$2 \times \left[(\text{log} - \text{likelihood from more complex model}) - (\text{log} - \text{likelihood from simpler model}) \right] =$$

$$2 \times \log\left(\frac{\text{likelihood from more complex model}}{\text{likelihood from simpler model}} \right) = LR$$

- the *LR* statistic is approximately chi-squared distributed, with degrees of freedom equal to the difference in number of predictors between nested models
- likelihood ratio tests for evaluating contribution of predictors are more reliable than Wald tests, especially for smaller samples

Note: for the *LR* test to make sense, models being compared should be based on exactly the same set of observations.