Applied Logistic Regression

Week 5

1. Homework week 4: highlights
2. Statistical adjustment
3. Adjusting odds ratios for confounding
4. Interaction and confounding I
5. Interaction and confounding II
6. Estimating odds ratios in the presence of interaction
7. Homework

Stanley Lemeshow, Professor of Biostatistics

College of Public Health, The Ohio State University

THE OHIO STATE UNIVERSITY

In multivariable logistic regression each estimated coefficient provides an estimate of the log odds ratio adjusting for all other variables included in the model.

What do we mean by "adjusting, statistically, for other variables?"

- To answer this we will first consider the linear regression model known as the ANALYSIS OF COVARIANCE.

Suppose we have a continuous dependent variable *y* (say weight).

- We are interested in exploring whether 2 groups of individuals are comparable with respect to weight.

Let

$$X = \begin{cases} 0 \text{ if individual belongs to group } i \\ 1 \text{ if individual belongs to group } j \end{cases}$$

Unfortunately, it is not enough simply to compare the mean weight of individuals in group *i* to the mean weight of those in group *j*.

- This is because there is another variable (say "age") that is related to the dependent variable.

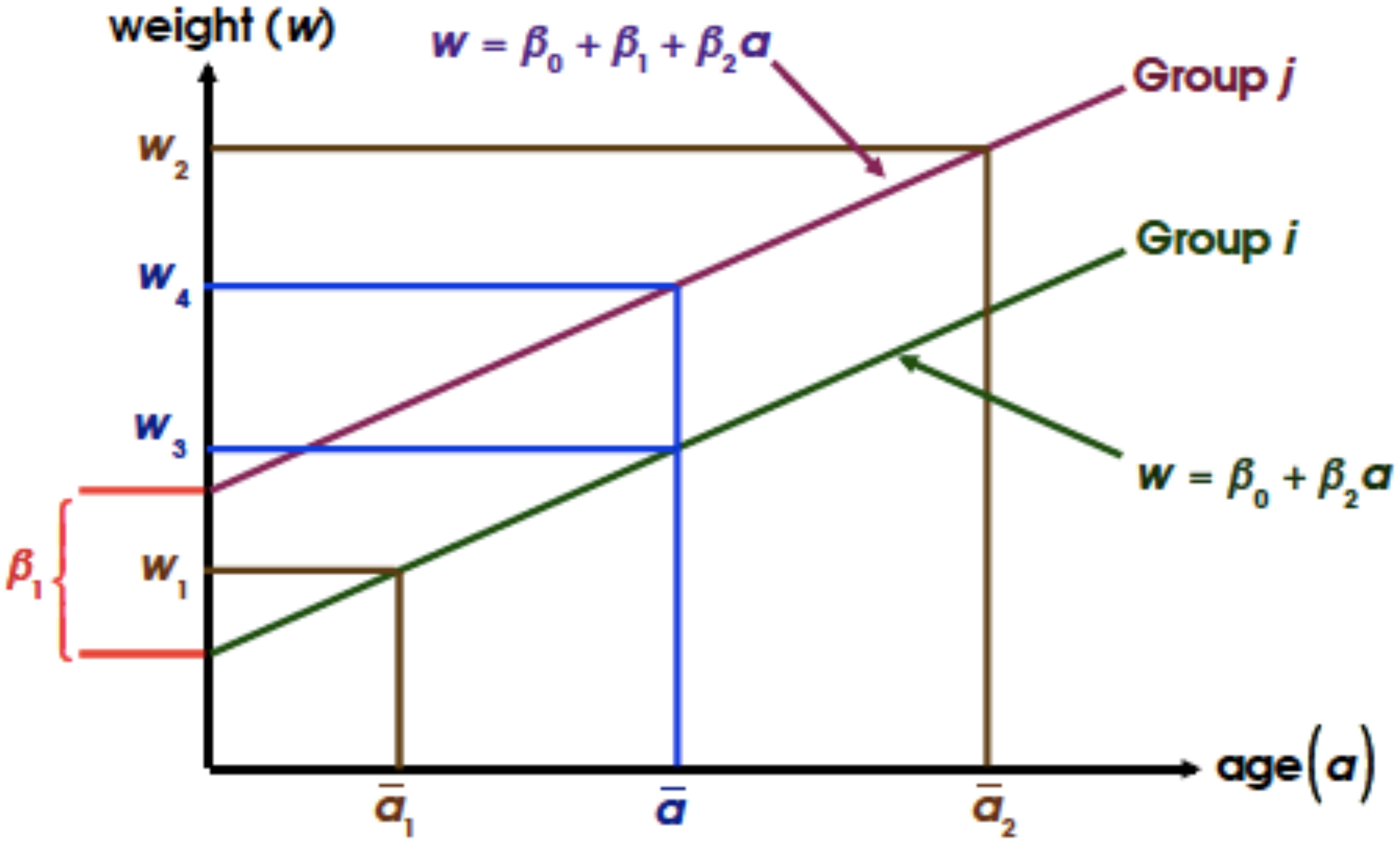
If the two groups differed with respect to this third variable, they would also differ with respect to weight.

- Therefore, our goal is to adjust for age (i.e., make the groups equivalent w.r.t. age) before comparing the groups w.r.t. weight .

That is, if group 1 is much younger than group 2, then group 1 might be much lighter than group 2.

- However, if we controlled for age, there might be no difference whatsoever.

Let us consider the following figure:



For this analysis we assume:

- relationship between weight and age is linear
- each group has the same non-zero slope

If we were doing ANCOVA, both of these assumptions would be tested before making inference about group differences.

In this model:

β_1 = true difference in weight between the two groups

β_2 = rate of change in weight per year of age in each group

Suppose that, for group i, the mean age is $\bar{a}_1$

and for group j, the mean age is $\bar{a}_2$

If that were the case then:

the mean weight for group *i* would be w_1

and the mean weight for group *j* would be w_2

and $\left(w_2 - w_1\right) >> \beta_1$

Therefore our impression of the difference in weight between the groups is being drastically inflated by the inherent difference in age between the groups.

This difference is:

$$w_2 - w_1 = \left[\beta_0 + \beta_1 + \beta_2 \bar{a}_2\right] - \left[\beta_0 + \beta_2 \bar{a}_1\right] = \overset{\text{true difference}}{\beta_1} + \overset{\text{difference in ages between groups}}{\beta_2\left(\bar{a}_2 - \bar{a}_1\right)}$$

Our goal is to eliminate the second term $\beta_2\left(\bar{a}_2 - \bar{a}_1\right)$.

We do this by comparing the two groups at a common value of age. Typically, the value used is $\bar{a}$ = mean age of all study subjects (but it could be any value).

As we see in the figure, using $\bar{a}$ results in a comparison of w_4 to w_3.

But $w_4 - w_3 = \left[\beta_0 + \beta_1 + \beta_2 \bar{a}\right] - \left[\beta_0 + \beta_2 \bar{a}\right] = \beta_1$, the true difference between the groups.

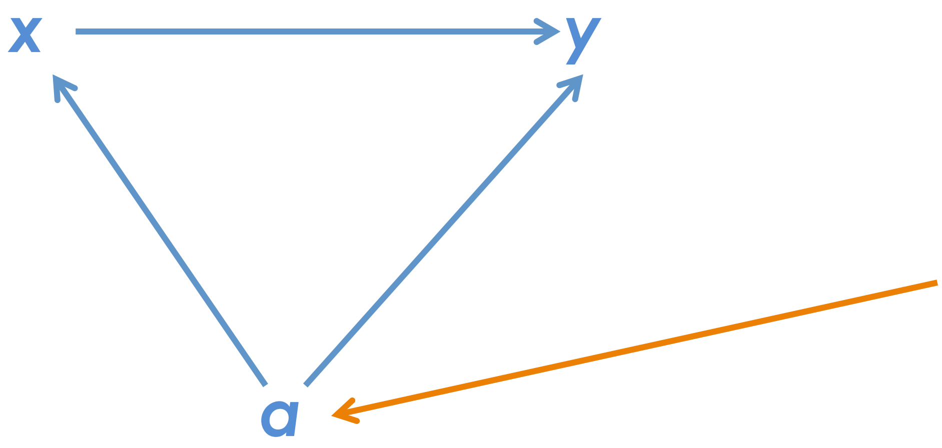
Now let us return to logistic regression

$$\text{Here } y = \begin{cases} 0 \text{ Disease absent} \\ 1 \text{ Disease present} \end{cases}$$

The vertical axis will now denote the logit, $g(x,a)$, where

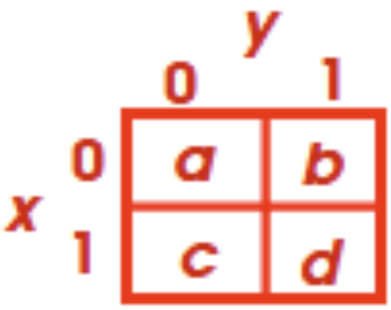
$$g(x,a) = \beta_0 + \beta_1 x + \beta_2 a$$

We assume that "*a*" is related both to the outcome "*y*" and the grouping variable "x" (that may be the presence or absence of a risk factor).



termed a “confounder”

If we simply examined



		y	
		0	1
x	0	a	b
	1	c	d

then $OR = \frac{ad}{bc} \approx e^{\beta_1+\beta_2\left(\bar{a}_2-\bar{a}_1\right)}$

i.e., we would incorrectly estimate the effect of group due to the difference in the distribution of age.

To account or adjust for this difference in age, we include age in the model and calculate the logit difference at a common value of age, $\bar{a}$.

The logit difference is

$$g\left(x=1,\bar{a}\right)-g\left(x=0,\bar{a}\right)=\beta_1$$

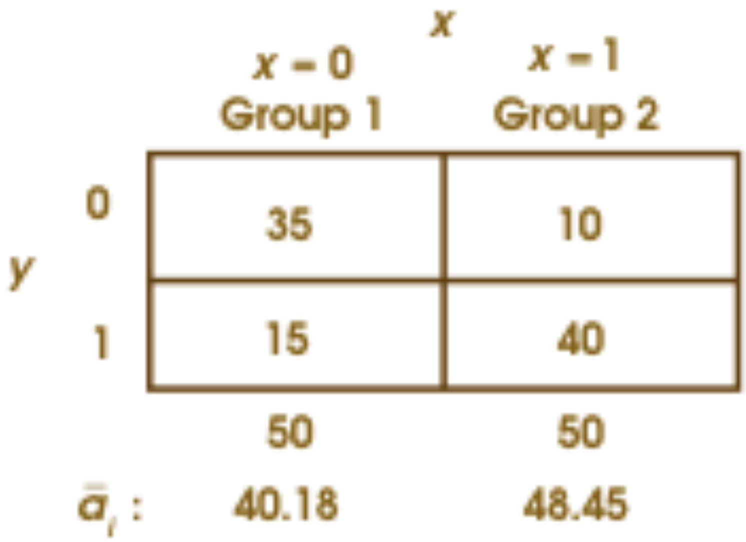
Thus the coefficient associated with group, β_1, is the log odds-ratio that we would expect to obtain from a univariable comparison if the two groups had the same mean age, $\bar{a}$.

Example

$$\text{let } y = \begin{cases} 0 \text{ if the subject had not seen a physician within past 6 months} \\ 1 \text{ if the subject had seen a physician within past 6 months} \end{cases}$$

$$x = \begin{cases} 0 \text{ first group of men} \\ 1 \text{ second group of men} \end{cases}$$

a = age ← covariate



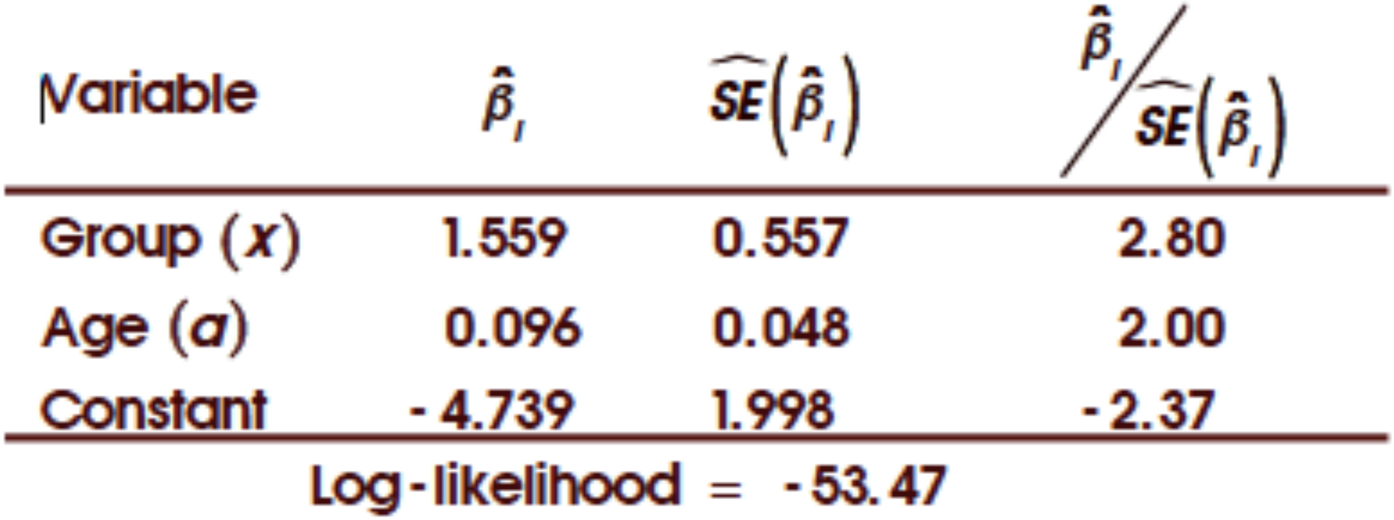
		x	
		$x = 0$ Group 1	$x = 1$ Group 2
y	0	35	10
	1	15	40
		50	50
$\bar{a}_i$:		40.18	48.45

The crude odds ratio is

$$\widehat{OR} = \frac{35 \times 40}{10 \times 15} = 9.33$$

but group 2 is considerably older than group 1 and, if age is related to visiting a physician, it is a possible confounder and should be controlled or adjusted for.

A logistic regression model was fit to these data with the following results:



Variable	$\hat{\beta}_i$	$\widehat{SE}\left(\hat{\beta}_i\right)$	$\hat{\beta}_i \Big/ \widehat{SE}\left(\hat{\beta}_i\right)$
Group (x)	1.559	0.557	2.80
Age (a)	0.096	0.048	2.00
Constant	- 4.739	1.998	- 2.37

Log - likelihood = - 53. 47

$$\text{The crude or unadjusted odds ratio} \approx e^{\hat{\beta}_1+\hat{\beta}_2\left(\bar{a}_2-\bar{a}_1\right)}$$

$$= e^{1.559+0.096\left(48.45-40.18\right)}$$

$$= e^{2.39} = 10.945$$

Note: here we get 10.95 while crude odds ratio = 9.33

- discrepancy due to rounding of coefficients

  - 10.95 is based on difference in average logits
    9.33 is based on average estimated logistic probability for the two groups.

The age adjusted odds ratio is obtained as $e^{\hat{\beta}_1} = e^{1.559} = 4.75$

i.e., the association is not nearly as strong once we adjust for the large difference in age between the two groups.

The same procedure would be followed whether the variables involved are continuous, dichotomous, or polychotomous.

- Adjusted odds ratios are obtained by comparing individuals who differ only in the characteristic of interest and have the values of all other variables constant.

The effectiveness of the adjustment is entirely dependent on the adequacy of the assumptions of the model:

- linearity and constant slope.

WEEK 5: INTERACTION AND CONFOUNDING

When a covariate (z) is associated both with the outcome variable and the primary independent variable (risk factor), then the relationship between the risk factor and the outcome variable is said to be confounded.

By using the model

$$g(x, z) = \beta_0 + \beta_1 x + \beta_2 z$$

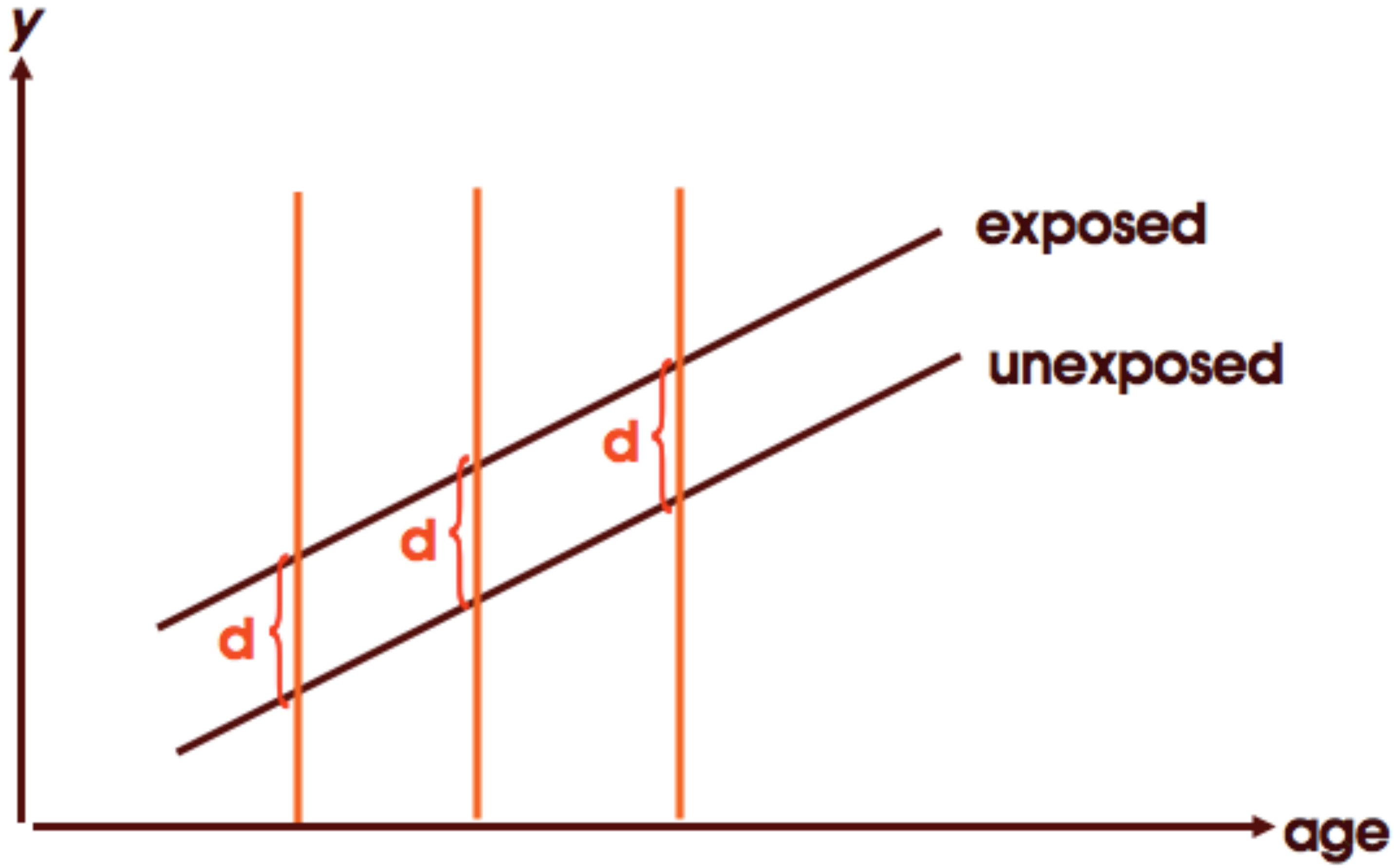
as in the previous example, we can adjust for confounding.

- However, this method may be used only when there is no interaction.

Let us first describe a situation where interaction is absent. Let the risk factor, x, be dichotomous (exposed, unexposed) and let the covariate, z, be continuous (age)

If the association between the covariate (age) and the outcome is the same within each level of the risk factor (i.e., exposure group), then there is no interaction between the covariate and the risk factor.

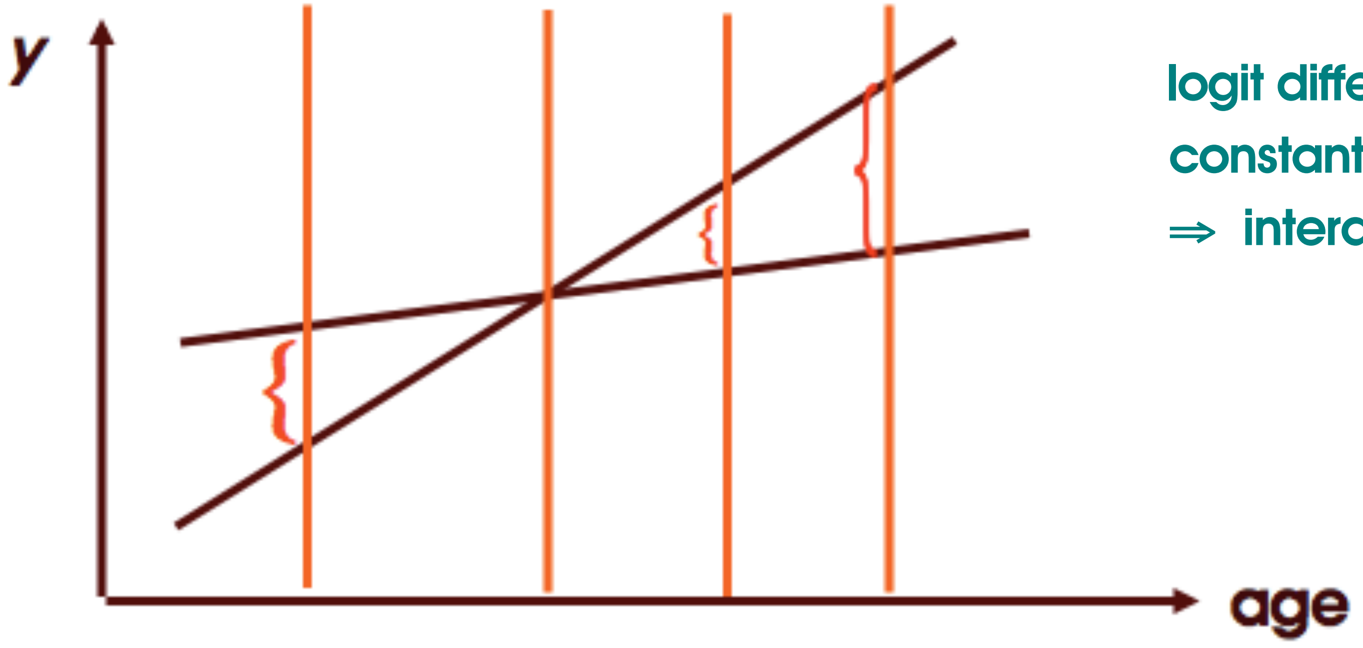
Graphically, the absence of interaction yields a model with two parallel lines - one for each level of the risk factor.



Regardless of which age we study, the relationship between exposure and outcome remains the same.

When interaction is present, the association between the risk factor and the outcome variable is not constant over different levels of the covariate.

i.e., the covariate modifies the effect of the risk factor.



logit difference is not constant over all ages
⇒ interaction

Clearly, the association between exposure and outcome is not constant over all age levels. Any statements made would have to specify the age level they were developed for.

In general

$$g\left(x,a\right)=\beta_0+\beta_1 x+\beta_2 a$$

for group 1: $x=0$

$$g\left(x,a\right)=\beta_0+\beta_2 a$$

for group 2: $x=1$

$$g\left(x,a\right)=\beta_0+\beta_1+\beta_2 a=\beta_0^*+\beta_2 a$$

same slope
$\Rightarrow$ no interaction

If there is interaction we use the model

$$g\left(x,a\right)=\beta_0+\beta_1 x+\beta_2 a+\beta_3 xa$$

for group 1: $x=0$

$$g\left(x,a\right)=\beta_0+\beta_2 a$$

for group 2: $x=1$

$$g\left(x,a\right)=\beta_0+\beta_1+\beta_2 a+\beta_3 a$$

$$=\left[\beta_0+\beta_1\right]+\left[\beta_2+\beta_3\right]a$$

if $\beta_3=0$, then the slopes are the same in each group

$\Rightarrow$ no interaction

if $\beta_3\neq 0$, then the slopes are not equal and interaction exists

In the previous example, age is an effect modifier since it modifies the effect of the risk factor.

When have an effect modifier, we must first specify the level of the covariate before estimating the odds ratio of the exposure and disease.

Example 1

$$\left.\begin{array}{l} x: \text{ SEX} \\ z: \text{ AGE} \\ y: \text{ CHD} \end{array}\right\}$$

$$\underline{\text{Model 1}}: \quad g(x) = -1.046 + 1.535x$$

$$\text{log likelihood} = -61.86 \qquad \widehat{OR} = e^{1.535} = 4.64$$

now, to see if age is a confounder, we build Model 2

$$\underline{\text{Model 2}}: \quad g(x,z) = -7.142 + .979x + .167z$$

$$\text{log likelihood} = -49.59 \qquad \widehat{OR} = e^{.979} = 2.66$$

Significance notwithstanding, $\widehat{\text{OR}}$ changed dramatically by the inclusion of age. This suggests confounding.

now, to see if there is effect modification we build the full model, including the interaction term.

$$\underline{\text{Model 3}}: \quad g(x, z, xz) = -6.103 + .481x + .139z + 0.59xz$$

$$\text{log likelihood} = -\,49.33$$

$$G = -2\left(-49.59 - \left(-49.33\right)\right) = 0.52 < \chi^2_{.95}(1)\left[p = .47\right]$$

Hence, age is <u>not</u> an effect modifier and Model 2 would be the appropriate model.

Example 2

Model	Constant	Sex	Age	Sex × Age	$\widehat{OR}$	log - likelihood	G	p
1	- 0.847	2.505			12.24 ↓	− 52.52		
2	− 6.194	1.734	0.147		5.66	− 46.79	11.46	<.01
3	− 3.105	0.047	0.629	0.206	note Δ in $\widehat{OR}$	− 44.76	4.06	$\underline{\underline{<.05}}$

Here age is an effect modifier

⇒ any estimate of odds ratio for sex should be made with respect to a given age.

To determine whether *z* is a confounder:

- compare estimated coefficient $\left(\text{or } \widehat{OR}\right)$ for the risk factor variable from models containing and not containing the covariate.

- any“biologically important”change would dictate that the covariate is a confounder and should be included in the model

  - this is regardless of the statistical significance of the estimated coefficient for the covariate.

To determine whether *z* is an effect modifier:

- fit a model including the interaction term

- this covariate is an effect modifier only when the interaction term is both biologically meaningful and statistically significant.

***When a covariate is an effect modifier, it cannot be a confounder since the estimate of the effect of the risk factor depends on the specific value of the covariate.

These concepts may be extended to cover any number of variables on any measurement scale(s).

When there is an interaction between a risk factor and a covariate, an estimate of the odds ratio for the risk factor should be made at a specific level of the covariate.

To do this we must take into account the correlation between the two interacting variables.

Consider a model containing: a risk factor: F
a covariate: X
and their interaction: $F \bullet X$

The logit for this model at $F = f$ and $X = x$ is

$$g\left(f, x\right) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 fx$$

The log-odds ratio for $F = f_1$ vs. $F = f_0$ with X held constant at $X = x$ is

$$\ln\left[OR\left(F = f_1, F = f_0, X = x\right)\right] = g\left(f_1, x\right) - g\left(f_0, x\right)$$

$$= \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 x - \left[\beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 x\right]$$

$$= \beta_1\left(f_1 - f_0\right) + \beta_3 x\left(f_1 - f_0\right)$$

and

$$\widehat{OR}\left(F = f_1, F = f_0, X = x\right) = e^{\hat{\beta}_1\left(f_1 - f_0\right) + \hat{\beta}_3 x\left(f_1 - f_0\right)}$$

We estimate the variance as

$$\widehat{Var}\left\{\ln\left[\widehat{OR}\left(F = f_1, F = f_0, X = x\right)\right]\right\} = \left(f_1 - f_0\right)^2 \widehat{Var}\left(\hat{\beta}_1\right)$$

$$+ \left(f_1 - f_0\right)^2 x^2 \widehat{Var}\left(\hat{\beta}_3\right) + 2\left(f_1 - f_0\right) x \widehat{Cov}\left(\hat{\beta}_1, \hat{\beta}_3\right)$$

These values are available in most computer packages.

Then

$$OR \le e^{\hat{\beta}_1(f_1-f_0)+\hat{\beta}_3 x(f_1-f_0)\pm z_{1-\alpha/2}\widehat{SE}\left\{\ln\left[\widehat{OR}(F=f_1,F=f_0,X=x)\right]\right\}}$$

When

$$F = \begin{cases} 0 = f_0 \\ 1 = f_1 \end{cases},$$

a dichotomous risk factor, these expressions simplify to

$$\ln\left[\widehat{OR}(F=1,F=0,X=x)\right] = \hat{\beta}_1(1-0) + \hat{\beta}_3 x(1-0) = \hat{\beta}_1 + \hat{\beta}_3 x$$

and

$$\widehat{Var}\left\{\ln\left[\widehat{OR}(F=1,F=0,X=x)\right]\right\} = \widehat{Var}\left(\hat{\beta}_1\right) + x^2\widehat{Var}\left(\hat{\beta}_3\right) + 2x\widehat{Cov}\left(\hat{\beta}_1,\hat{\beta}_3\right)$$

The confidence interval is established in the usual way.

Example

Low birth weight data

$$y = \text{low birth Weight } (\text{LOW}) \begin{cases} 0 = \text{Birth Weight} \geq 2500\text{g} \\ 1 = \text{Birth Weight} < 2500\text{g} \end{cases}$$

$$x = \text{weight of mother at last menstrual period}$$

$$(\text{LWD}) \begin{cases} 0 = \text{LWT} \geq 110 \text{ lbs} \\ 1 = \text{LWT} < 110 \text{ lbs} \end{cases}$$

$$z = \text{age}$$

We now fit a series of models as follows:

Model	Constant	LWD	AGE	LWD × AGE	Log-likelihood	G	p-value
0	-0.790				-117.34		
1	-1.054	1.054[1]			-113.12	8.44	0.004
2	-0.027	1.010[2]	-0.044		-112.14	1.96	0.160
3	0.774	-1.944	-0.080	0.132	-110.57	3.14	0.080[3]

1. Clearly, model with LWD is significant.

- Crude odds ratio $= \widehat{OR} = e^{1.054} = 2.87$

2. Age is not a strong confounder.

- Adjusted odds-ratio $= \widehat{OR} = e^{1.01} = 2.74$

3. There is some evidence of an interaction, although not a very strong one. It might be safer to present the risk associated with low maternal weight at last menstrual cycle $\left(\text{LWD}\right)$ at various ages because the odds ratio will not be constant over age.

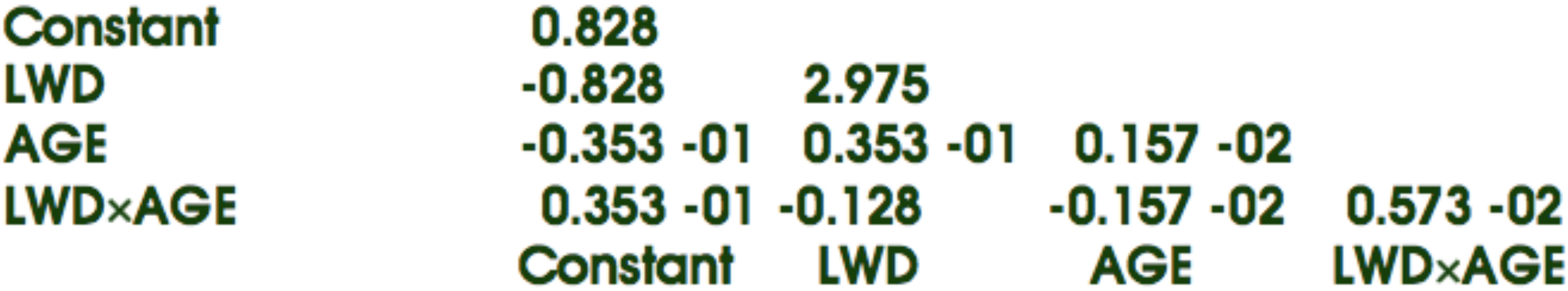
We must use model 3, containing the interaction term.

For a woman of $AGE = a$, we have

$$\ln\left[\widehat{OR}\left(LWD=1,\ LWD=0,\ AGE=a\right)\right]$$

$$= \hat{\beta}_1 + \hat{\beta}_3 a = -1.944 + 0.132a$$

To compute the variance, we obtain the estimated covariance matrix for the estimated parameters from our computer program. This matrix is as follows:



Constant	0.828			
LWD	-0.828	2.975		
AGE	-0.353 -01	0.353 -01	0.157 -02	
LWD×AGE	0.353 -01	-0.128	-0.157 -02	0.573 -02
	Constant	LWD	AGE	LWD×AGE

Using these values we compute:

$$\widehat{Var}\left\{\ln\left[\widehat{OR}\left(LWD=1,\ LWD=0,\ AGE=a\right)\right]\right\}$$

$$=\widehat{Var}\left(\hat{\beta}_1\right)+x^2\widehat{Var}\left(\hat{\beta}_3\right)+2x\widehat{Cov}\left(\hat{\beta}_1,\hat{\beta}_3\right)$$

$$=2.975+a^2\left(.00573\right)+2a\left(-0.128\right)$$

These expressions are then evaluated at a number of values of *a*.

e.g., at age = 30 years,

$$\ln\left[\widehat{OR}\left(LWD=1,\ LWD=0,\ AGE=30\right)\right]$$

$$=-1.944+1.32\left(30\right)=2.016$$

$$\widehat{Var}\left\{\ln\left[\widehat{OR}\left(LWD=1,\ LWD=0,\ AGE=30\right)\right]\right\}$$

$$\widehat{OR}=e^{2.016}=7.5$$

$$=2.975+30^2\left(.00573\right)+2\left(30\right)\left(-0.128\right)=.452$$

so the 95% CI is:

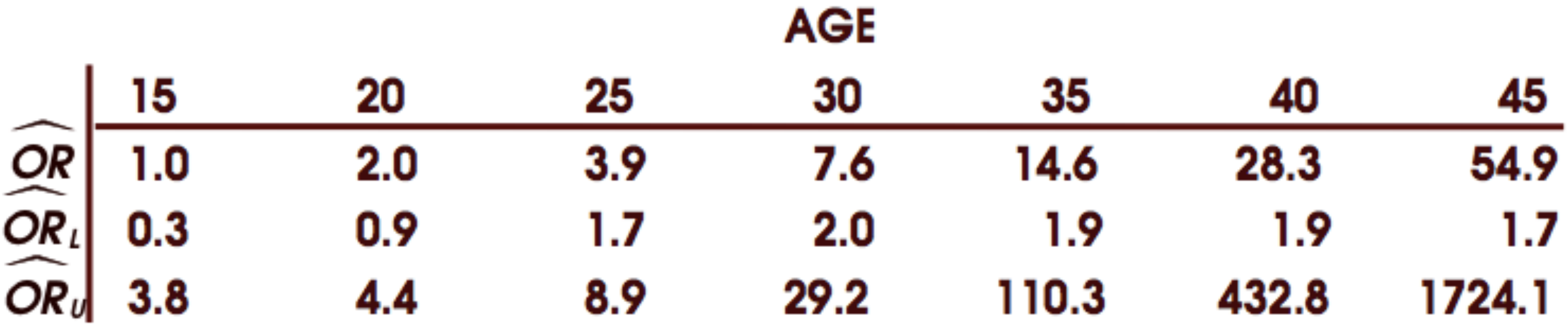
$$e^{2.016-1.96\sqrt{.452}} \le OR \le e^{2.016+1.96\sqrt{.452}}$$

$$e^{.698} \le OR \le e^{3.33}$$

$$2.01 \le OR \le 28.04$$

Use of a spreadsheet can facilitate this process considerably.

Carrying out these computations for a number of choices for *a* results in the following table:



	AGE						
	15	20	25	30	35	40	45
$\widehat{OR}$	1.0	2.0	3.9	7.6	14.6	28.3	54.9
$\widehat{OR}_L$	0.3	0.9	1.7	2.0	1.9	1.9	1.7
$\widehat{OR}_U$	3.8	4.4	8.9	29.2	110.3	432.8	1724.1

Note: $\ln\left(\widehat{OR}\right) = -1.944 + .132a$ ← log odds ratio

increase linearly with age

However, the confidence interval width increased enormously with age indicating that there is considerable uncertainty in these estimates

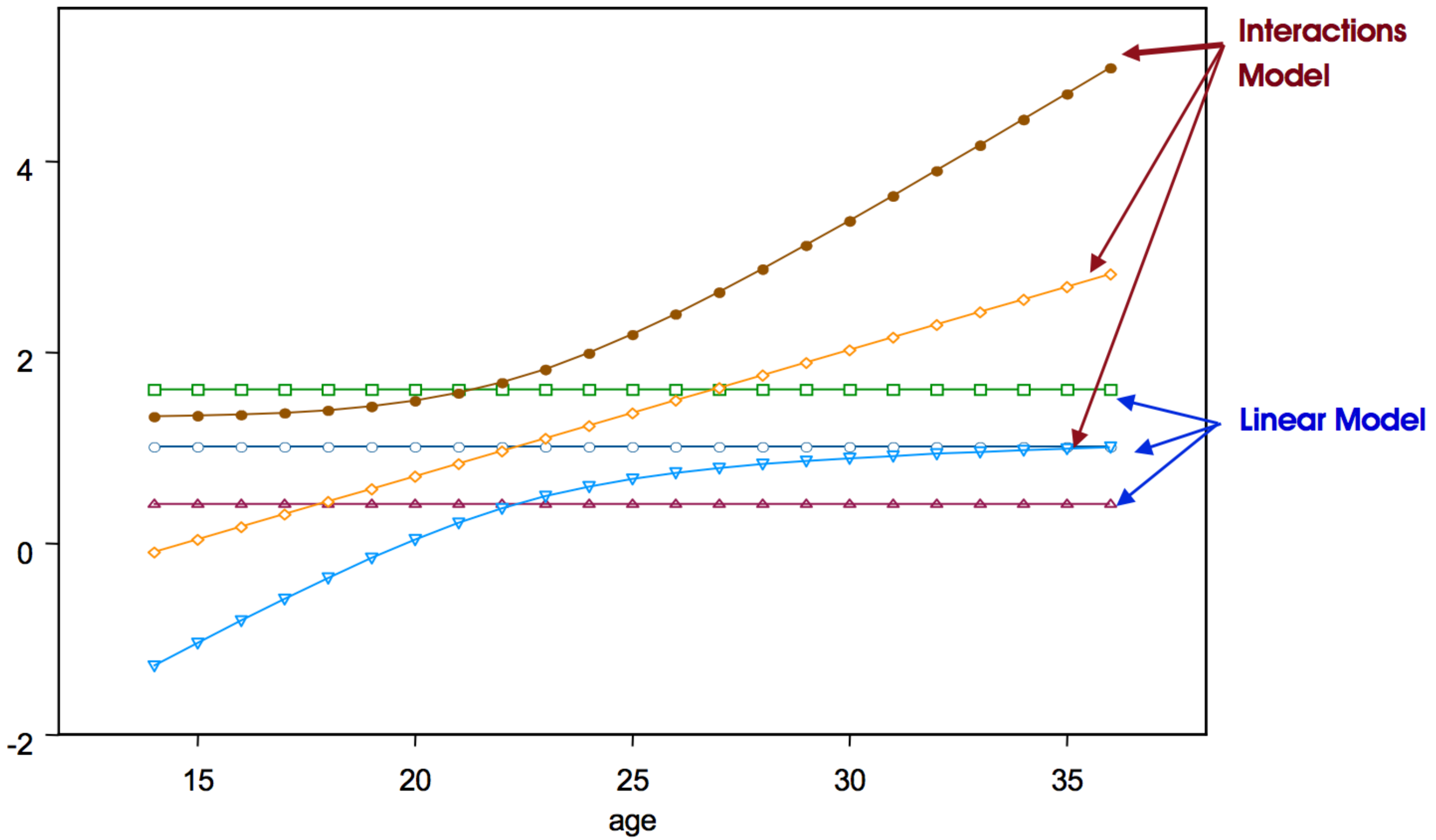
- particularly for women ≥ 30 years of age

```
. gen lwd=0

. replace lwd=1 if lwt<110
(42 real changes made)

. logit low lwd

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(1)      =       8.43
                                                  Prob > chi2     =     0.0037
Log likelihood = -113.12058                       Pseudo R2       =     0.0359

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     lwd |   1.053762   .3615635      2.914   0.004        .3451102   1.762413
   _cons |  -1.053762   .1883882     -5.594   0.000       -1.422996  -.6845277
------------------------------------------------------------------------------

. logistic low lwd

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(1)      =       8.43
                                                  Prob > chi2     =     0.0037
Log likelihood = -113.12058                       Pseudo R2       =     0.0359

------------------------------------------------------------------------------
     low | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     lwd |   2.868421   1.037116      2.914   0.004        1.412146   5.826481
------------------------------------------------------------------------------

. vce

         |      lwd     _cons
---------+------------------
     lwd |  .130728
   _cons |  -.03549    .03549
```

```
. logit low lwd age

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(2)      =      10.39
                                                  Prob > chi2     =     0.0056
Log likelihood = -112.14338                       Pseudo R2       =     0.0443

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     lwd |   1.010122   .3642627      2.773   0.006       .2961806    1.724064
     age |   -.044232   .0322248     -1.373   0.170      -.1073913    .0189274
   _cons |   -.026891   .7621481     -0.035   0.972      -1.520674    1.466892
------------------------------------------------------------------------------

. logistic low lwd age

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(2)      =      10.39
                                                  Prob > chi2     =     0.0056
Log likelihood = -112.14338                       Pseudo R2       =     0.0443

------------------------------------------------------------------------------
     low | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     lwd |   2.745937   1.000242      2.773   0.006       1.344713    5.607271
     age |    .956732   .0308305     -1.373   0.170       .8981741    1.019108
------------------------------------------------------------------------------

. vce

         |      lwd        age      _cons
---------+---------------------------
     lwd |  .132687
     age |  .000726   .001038
   _cons | -.052465   -.02379     .58087
```

```
. gen lwdxage= lwd* age
. logit low lwd age lwdxage

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(3)      =      13.53
                                                  Prob > chi2     =     0.0036
Log likelihood = -110.56997                       Pseudo R2       =     0.0577
------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     lwd |  -1.944089   1.724804     -1.127   0.260      -5.324643    1.436465
     age |  -.0795722   .0396343     -2.008   0.045       -.157254   -.0018904
 lwdxage |   .1321967   .0756982      1.746   0.081      -.0161691    .2805626
   _cons |   .7744952   .9100949      0.851   0.395      -1.009258    2.558248
------------------------------------------------------------------------------

. logistic low lwd age lwdxage

------------------------------------------------------------------------------
     low | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     lwd |   .1431176   .2468497     -1.127   0.260       .0048701    4.205802
     age |   .9235114   .0366027     -2.008   0.045        .854487    .9981114
 lwdxage |   1.141333   .0863969      1.746   0.081       .9839609    1.323874
------------------------------------------------------------------------------

. vce

         |       lwd        age   lwdxage     _cons
---------+----------------------------------------
     lwd |   2.97495
     age |   .035266   .001571
 lwdxage |  -.127603 -.001571    .00573
   _cons |  -.828273 -.035266   .035266   .828273

. lincom _b[ lwd]+30*_b[ lwdxage],or

 ( 1)  lwd + 30.0 lwdxage = 0.0

------------------------------------------------------------------------------
     low | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   7.552007   5.210013      2.931   0.003       1.953582    29.19397
------------------------------------------------------------------------------
```

Graph of the estimated logit and 90% confidence intervals from the linear model and the interactions model



Note that while there is some overlap in the CIE’s the linear model over estimates the effect for young ages and under estimates the effect for older women.