Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 5
Logistic Regression with Lines

Actual age

Lets now consider an assessment of a disease-exposure relationship with a potential confounder/modifier that is measured.

For example, lets suppose the age is consider a potential confounder or modifier and that previous research has considered actual age rather than age groups

Now consider the model:

$$\log(p/(1-p)) = \beta_0 + \beta_1 D + \beta_2 A + \beta_3 DA$$

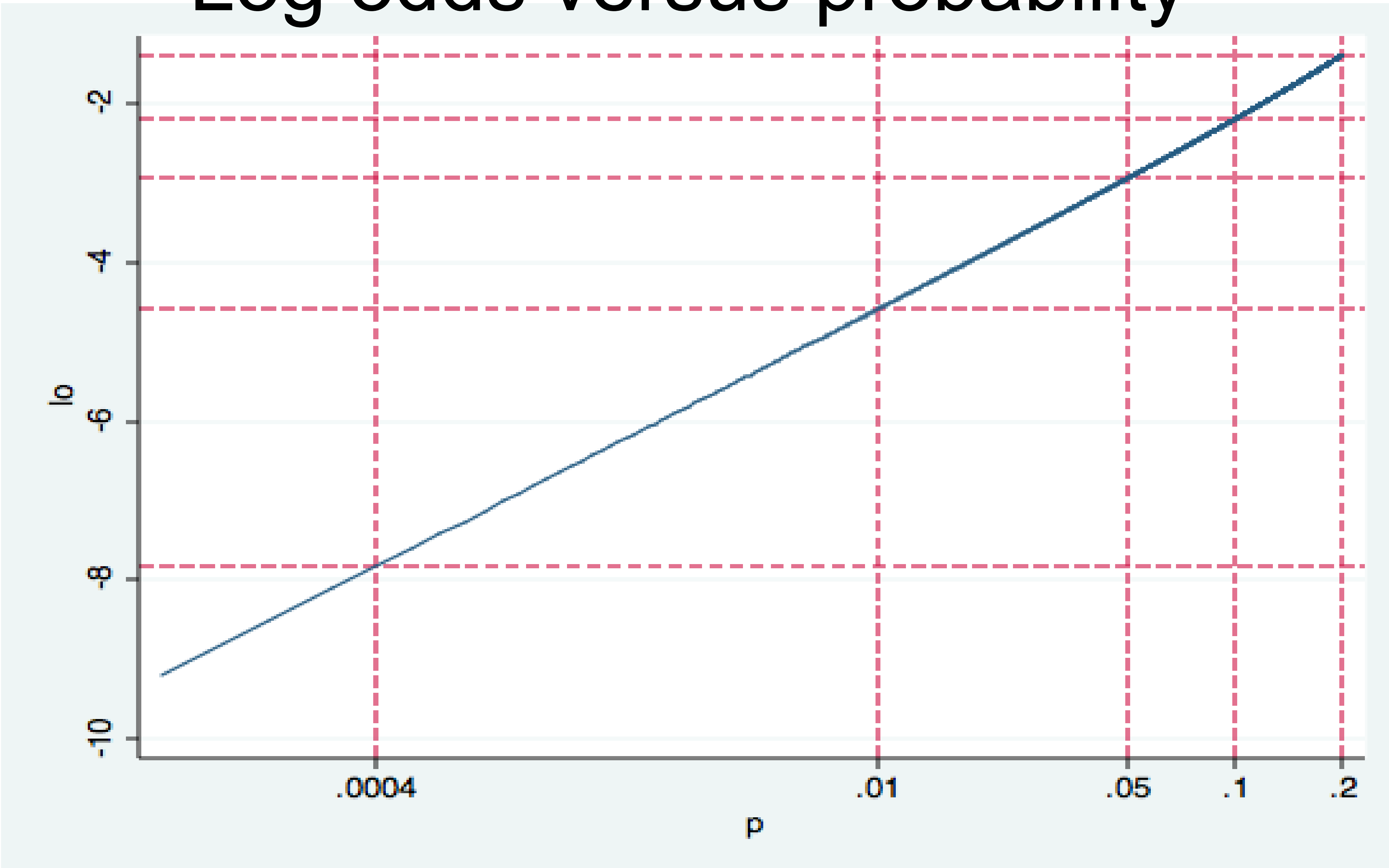
A is now age in years

In previous classes, age was an indicator for age group: A=0 for young and A=1 for old

We cannot list or tabulate all possible scenarios

We think of the log of odds as a function of age

Log odds versus probability



$$\log(p/(1-p)) = \beta_0 + \beta_1 D + \beta_2 A + \beta_3 DA$$

This model can be described in more than one way.

First think of how age can be a modifier

Then the model can be written as:

$$\log(p/(1-p)) = \beta_0 + \beta_2 A + (\beta_1 + \beta_3 A) D$$

so we can see that log(OR) = $\beta_1 + \beta_3 A$

that log(OR) is in fact a line in age. This model describes the way in which age modifies the disease-exposure relationship as a linear function on the log(odds ratio) scale.

Rate of Change

Notice that β_3 is the rate of change of the log of the odds ratio per year of age. The odds ratio continues to be the odds of exposure among those with disease divided by the odds of exposure among those without disease. If this odds ratio depends on age, then we have evidence that age is modifier.

In other words, if $\beta_3 \neq 0$ then age is a modifier.

$$\log(p/(1-p)) = \beta_0 + \beta_1 D + \beta_2 A + \beta_3 DA$$

Lets look at this model in another way:

If D=0, then $\log(p/(1-p)) = \beta_0 + \beta_2 A$

If D=1, then $\log(p/(1-p)) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) A$

So this model specifies 2 lines of possibly different slopes and intercepts: one line for the controls and one line for the cases.

Two Lines

so β_2 is the slope of the line for the controls. It is, for the controls, the rate of change of the log of odds of exposure per year of age

..and $\beta_2 + \beta_3$ is the slope of the line for the cases. It is, for the cases, the rate of change of the log of odds of exposure per year of age.

If $\beta_3 = 0$, the 2 lines are parallel and again we see that age is not a modifier and we can consider:

$$\log(p/(1-p)) = \beta_0 + \beta_1 D + \beta_2 A$$

Notice that now β_1 is the fixed vertical distance between the 2 parallel lines.

And now β_1 is the “assumed common” log(OR).

This is just like the situation where there were 2 age groups only now the “adjustment” is based the assumption that the exposure-age relationship is described by these 2 lines.

Units for an Explanatory Variable

If age in years, A, is changed to age in days, B,

then B = 365.25*A and the regression coefficient β becomes β / 365.25

Duration can be in seconds, minutes, hours, days, weeks, months, years. All a change in scaling.

Temperature could be changed from Fahrenheit (F) to Celsius (C). A change in location and scale. C = 5 / 9 * (F - 32)

Concentration (C) could be changed to log(C). For example acidity (C) might become pH = log(C)

Coefficients and Fitted Values

The Duration and Temperature examples are linear transformations. Some or all of the regression coefficients may change but the fitted values should not change. More on this matter later.

The Concentration example is a nonlinear transformation. The regression coefficients will all change and the fitted values will all change.

Centring

For measured explanatory variables, we may wish to centre. This means that :

X is replaced with X - C in the regression equation

Then, certain regression coefficients become specific to

X = C rather than X = 0.

For example, age = 0 is not possible but if

age = 20 is possible then we can change :

A to A - 20.

We do not centre indicator variables

An example

Consider:

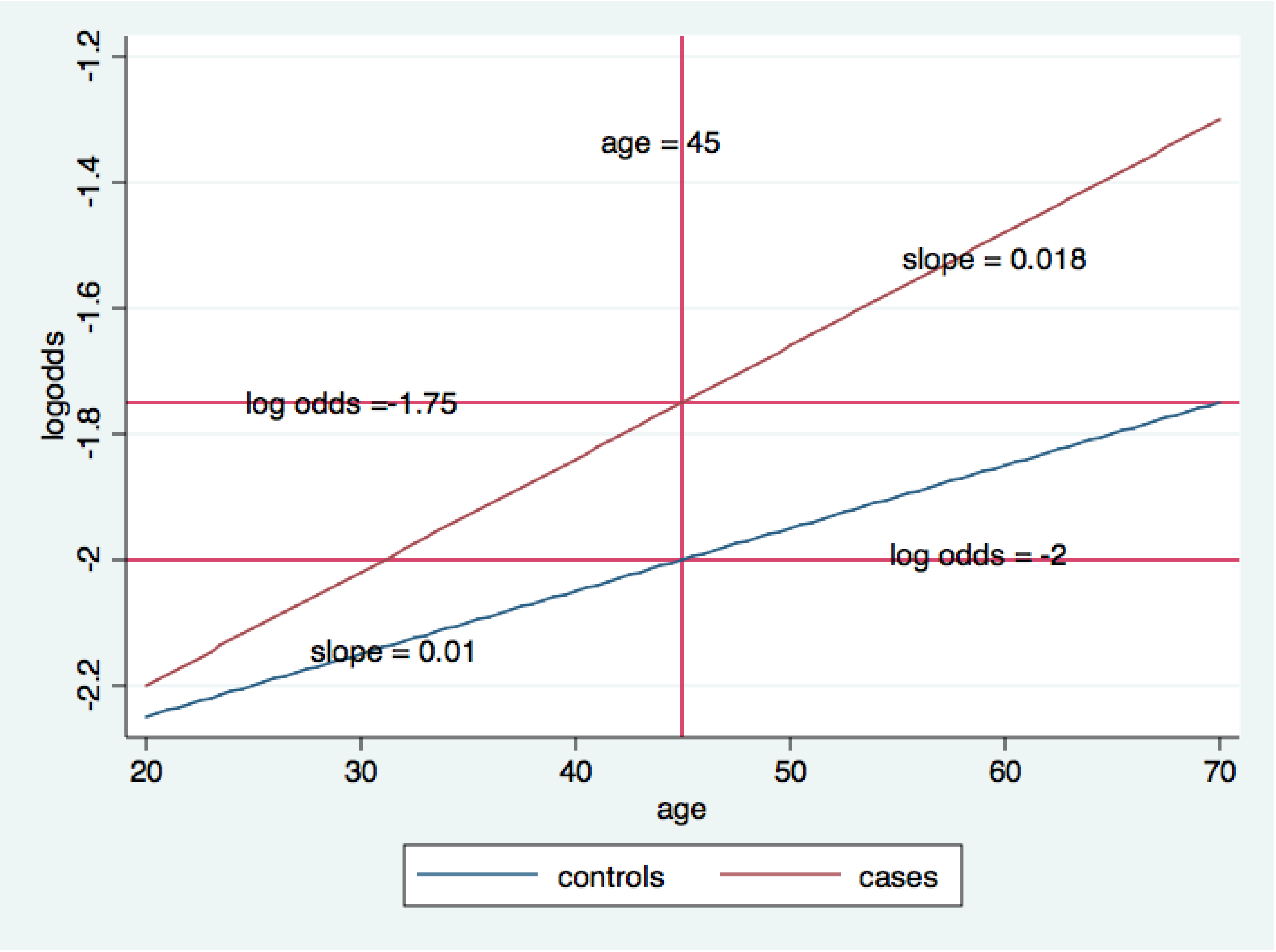
$$\log p/(1-p) = -2 + 0.25\mathrm{D} + (0.01 + 0.008\mathrm{D})(\mathrm{age} - 45)$$

For D=0,

$$\log p/(1-p) = -2 + 0.01(\mathrm{age} - 45)$$

For D=1,

$$\log p/(1-p) = -1.75 + 0.018(\mathrm{age} - 45)$$



age = 45
slope = 0.018
log odds =-1.75
log odds = -2
slope = 0.01
logodds
-1.2
-1.4
-1.6
-1.8
-2
-2.2
20
30
40
50
60
70
age
controls
cases

An example

Consider:

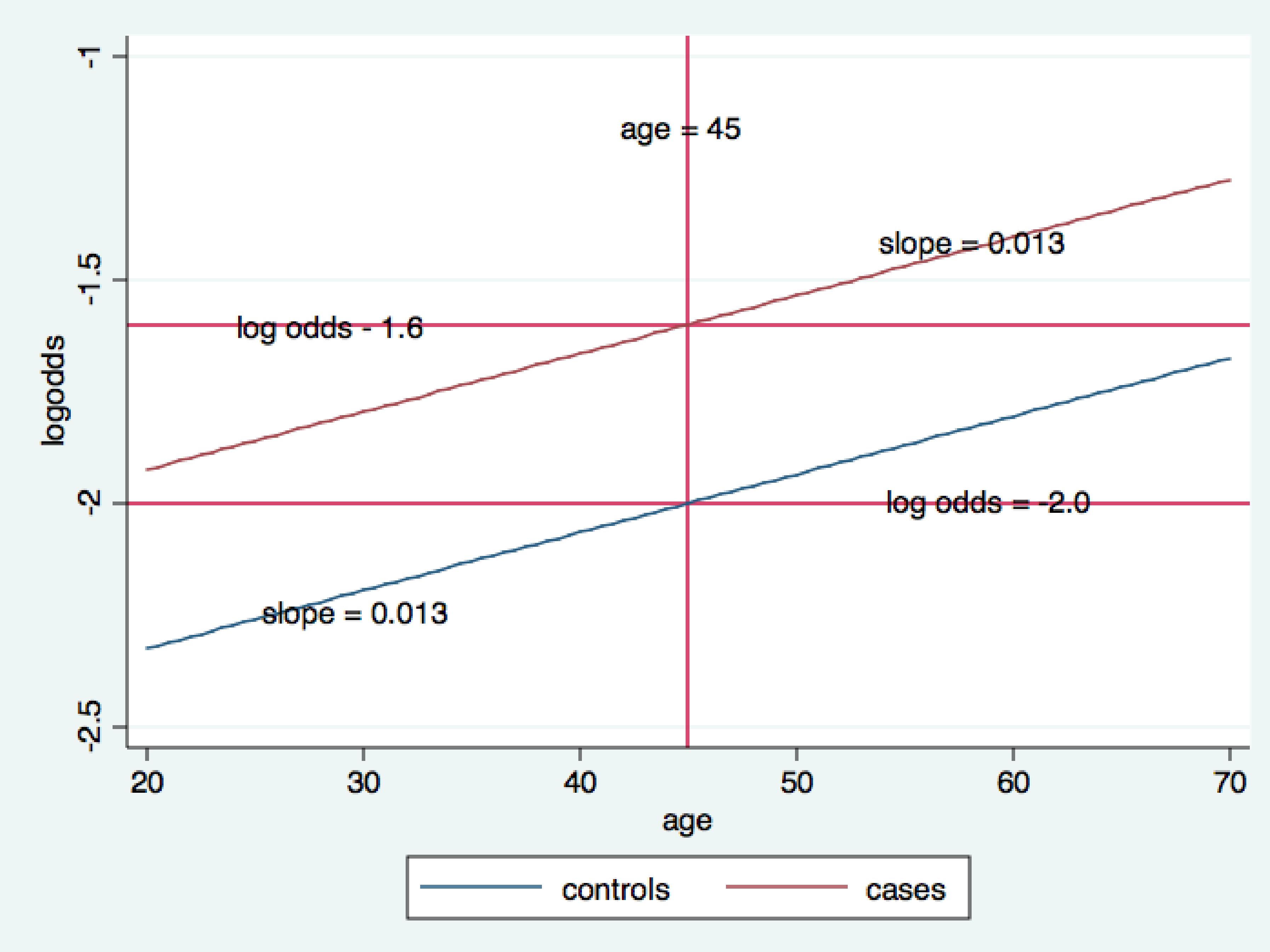
$$\log p/(1-p) = -2 + 0.4\mathrm{D} + (0.013)(\mathrm{age} - 45)$$

For D=0,

$$\log p/(1-p) = -2 + 0.013(\mathrm{age} - 45)$$

For D=1,

$$\log p/(1-p) = -1.6 + 0.013(\mathrm{age} - 45)$$



age = 45
slope = 0.013
log odds - 1.6
log odds = -2.0
slope = 0.013
logodds
-1
-1.5
-2
-2.5
20
30
40
50
60
70
age
controls
cases

Now consider a cohort study and a measured exposure E

Suppose that the exposure is measured in time :
a length of exposure :
Years of exposure to asbestos
Length of time for potential concussion
Dosage of a drug relative to body weight or body size

Gender G is a potential confounder/modifier

Now consider the models :

$$p = P(D)$$

$$\log\left(p/(1-p)\right) = \beta_0 + \beta_1 E + \beta_2 G + \beta_3 \mathrm{GE}$$

$$\log\left(p/(1-p)\right) = \beta_0 + \beta_1 E + \beta_2 G$$

$$\log\left(p/(1-p)\right) = \beta_0 + \beta_1 E$$

The association(s) between disease and exposure are now measured by slope(s)