Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 7 : Nonlinearity

Grouping,

Parabolic Curves,

Piecewise Linear “Curves”,

Transformations and Polynomials

Restricted Cubic Splines

Lines discredited

Sometimes it is clear to the investigator that the relationship between the log odds and age is not linear and so lines will mishandle their role.

Such nonlinearity may come from the literature review or may come from diagnostics or measures of 'goodness of fit' (to be discussed later) based on the data at hand.

Grouping

Even though actual age may be recorded, investigators often consider classifying each person into age groups. Ensuring that the age groups are mutually exclusive and exhaustive, analysis using indicator variables for the age groups may enable the investigator to explore nonlinearity in approximate terms.

If there are k age groups, then there would be (k-1) indicators. Here, the youngest group would typically be the baseline group.

A Return to Stratified Analysis

Returning to age and gender as possible modifiers/confounders, we now see that there would be 2k strata if a joint analysis is required.

i.e. 2k 2x2 tables

This approach (and its logistic regression equivalent) may be viable and useful as a start to analysis but often one wishes for methods that do not require so many variables.

For example, suppose there are four age groups

There would be 8 age-gender specific 2x2 tables, the omnibus test for modification would have 7 degrees of freedom;

4 age specific 2x2 tables age modification test with 3 degrees of freedom;

two gender specific 2x2 tables, as before

and the crude 2x2 table

Models might start with :

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 E + \beta_2 A_2 + \beta_3 A_3 + \beta_4 A_4 + \beta_5 G + \beta_6 GA_2 + \beta_7 GA_3 + \beta_8 GA_4 + \beta_9 EA_2 + \beta_{10} EA_3 + \beta_{11} EA_4 + \beta_{12} EG + \beta_{13} EGA_2 + \beta_{14} EGA_3 + \beta_{15} EGA_4$$

where, for example, we might have :

$$A_1 : \geq 20 \text{ to } < 35,\ A_2 : \geq 35 \text{ to } < 50,$$
$$A_3 : \geq 50 \text{ to } < 65,\ A_4 : \geq 65$$

Next step(s)

We could address whether age modification is modified by gender with :

$$H_0: \ \beta_{13}=\beta_{14}=\beta_{15}=0$$

Then, if there is no evidence against this null hypothesis, one could refit with :

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 E+\beta_2 A_2+\beta_3 A_3+\beta_4 A_4$$
$$+\beta_5 G+\beta_6 GA_2+\beta_7 GA_3+\beta_8 GA_4$$
$$+\beta_9 EA_2+\beta_{10} EA_3+\beta_{11} EA_4+\beta_{12} EG$$

Then ...

one could consider age modification assumed common to gender :

$$H_0: \ \beta_9 = \beta_{10} = \beta_{11} = 0$$

and gender modification assumed common to age : $H_0: \ \beta_{12} = 0$

then some further fitting ?

These models do not tell us about the nature the log odds - age relationship though.

Parabolae

Now contemplate other methods to deal with the nonlinear relationship between the log of odds of disease and age.

Sometimes, the nonlinearity can be managed by part of the arc of a parabola.

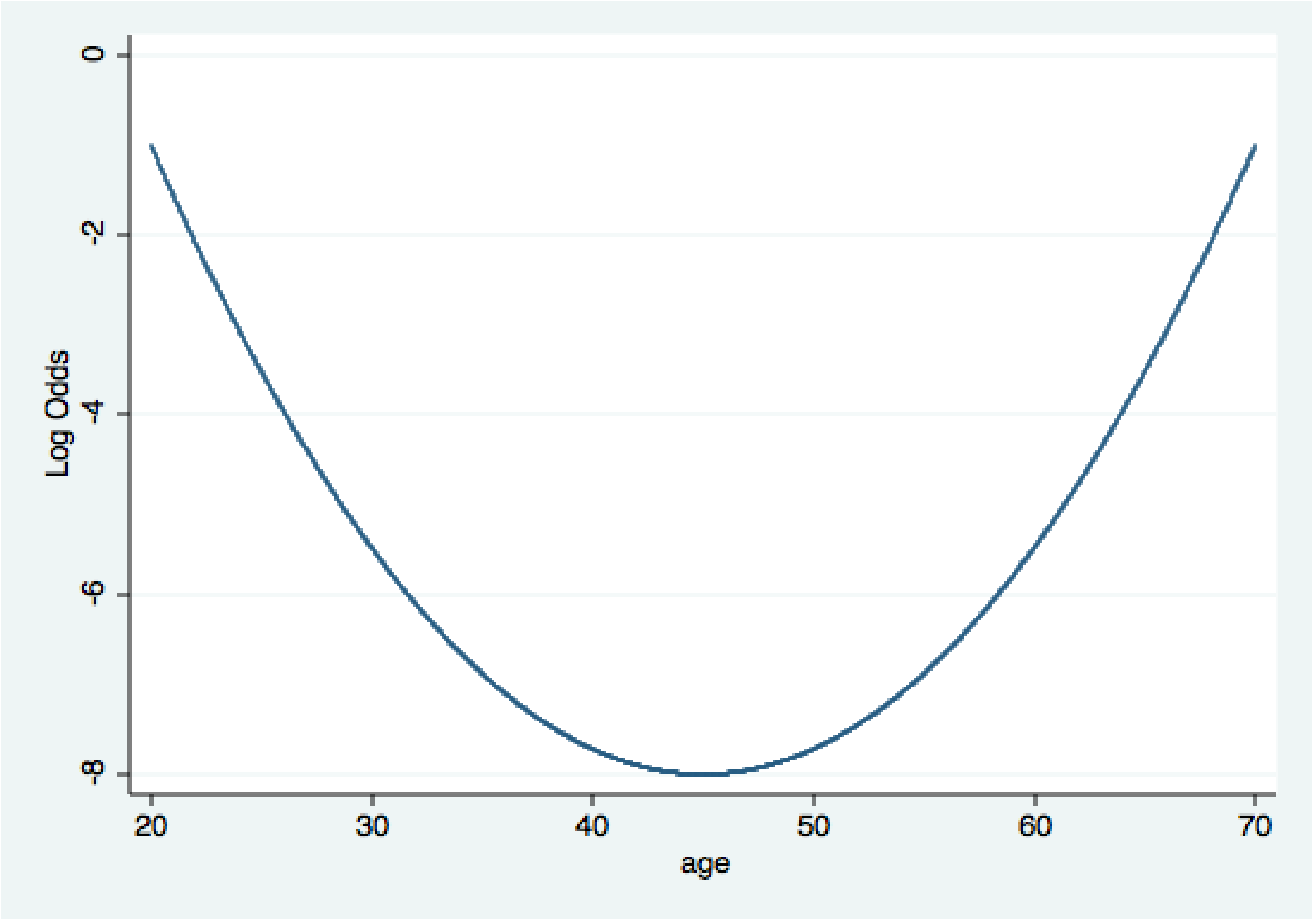
Review

A brief review of such curves is in order. Using 'complete the square', we can write the parabolic function:

$$f(x)=ax^2+bx+c=a\left(x+\frac{b}{2a}\right)^2+c-\frac{b^2}{4a}$$

For example:

$$f(x)=0.0112x^2-1.008x+14.68=\frac{7}{625}(x-45)^2-8$$


Log Odds
0
-2
-4
-6
-8
20
30
40
50
60
70
age

Right side up

So then we can see that: If a>0, then the parabola is U shaped. The bottom of the U is the minimum when

$x = -\frac{b}{2a}$ and the minimum value is $f\left(-\frac{b}{2a}\right) = c - \frac{b^2}{4a}$

e.g. a = 0.0112, b = -1.008 and c = 14.68

so then the minimum value = -8 when x(age) = 45

Upside Down

If a<0, then the parabola is an upside down U and now the maximum is when:

$$x = -\frac{b}{2a} \quad \text{and the maximum value} \quad f\left(-\frac{b}{2a}\right) = c - \frac{b^2}{4a}$$

for example:

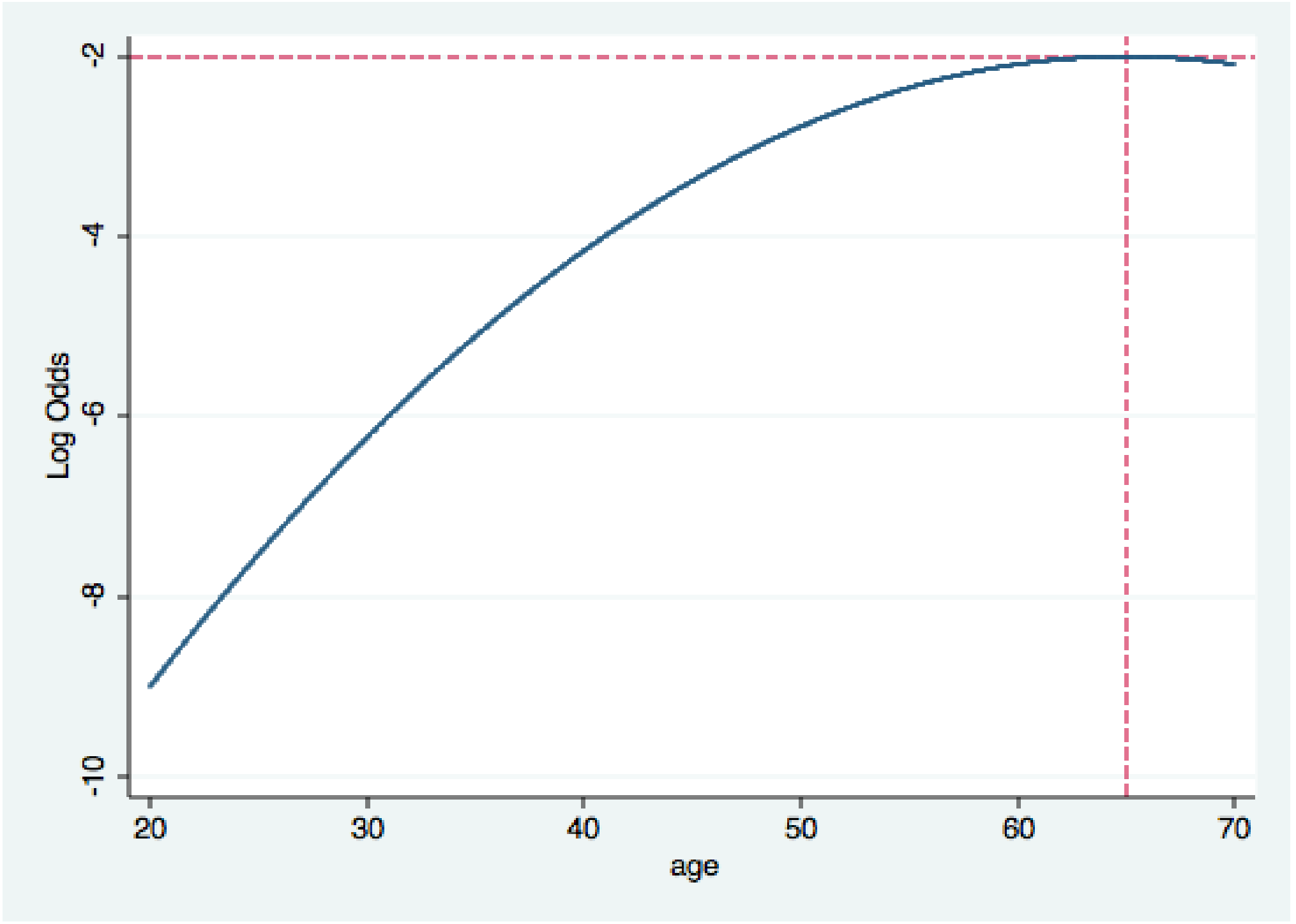
$$f(x) = -0.0035x^2 + 0.4494x - 16.6049 = -\frac{7}{2025}(x-65)^2 - 2$$

The coefficients give the maximum value, the age where maximum occurs and the curvature

a= -0.0035, b=0.4494 and c= -16.6049

so then the maximum value is -2

at x(age)=65


-2
-4
-6
-8
-10
Log Odds
20
30
40
50
60
70
age

Curvature, Minimum/Maximum Location, Minimum/Maximum Value

In the world of parabolic fits, the regression coefficients determine 3 characteristics: the curvature, where minimum/maximum occur and minimum/maximum value.

Let us now add quadratic terms to our initial model for the log odds of disease:

$$\log(p/(1-p))=$$

$$\beta_0+\beta_1 G+\beta_2 A+\beta_3 GA+\beta_4 E+\beta_5 GE+\beta_6 AE+\beta_7 GAE$$
$$+\beta_8 A^2+\beta_9 GA^2+\beta_{10} A^2 E+\beta_{11} GA^2 E$$

$$=(\beta_0+\beta_1 G+\beta_4 E+\beta_5 GE)+(\beta_2+\beta_3 G+\beta_6 E+\beta_7 GE)A$$
$$+(\beta_8+\beta_9 G+\beta_{10} E+\beta_{11} GE)A^2$$

This model determines 4 parabolae:

$$\mathrm{F}\bar{\mathrm{E}} \quad \log(p/(1-p)) = \beta_0 + \beta_2 \mathrm{A} + \beta_8 \mathrm{A}^2$$

$$\mathrm{M}\bar{\mathrm{E}} \quad \log(p/(1-p)) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)\mathrm{A} + (\beta_8 + \beta_9)\mathrm{A}^2$$

$$\mathrm{FE} \quad \log(p/(1-p)) = (\beta_0 + \beta_4) + (\beta_2 + \beta_6)\mathrm{A} + (\beta_8 + \beta_{10})\mathrm{A}^2$$

$$\mathrm{ME} \quad \log(p/(1-p)) = (\beta_0 + \beta_1 + \beta_4 + \beta_5) + (\beta_2 + \beta_3 + \beta_6 + \beta_7)\mathrm{A} + (\beta_8 + \beta_9 + \beta_{10} + \beta_{11})\mathrm{A}^2$$

We can describe the parabolae using the coefficients:

For example, with $F\ \bar{E}$

we see that β_8 is the curvature while $-\frac{\beta_2}{2\beta_8}$ and $\beta_0 - \frac{\beta_2^2}{4\beta_8}$

are the min/max location and value respectively.

The coefficients in front of A^2 provide for assessment of curvature. In the presence of curvature, the coefficients in front of A have specialized interpretations to determine the position and value of the minimum/maximum.

As before, we now have...

.... equations for the log of the Odds Ratio:

$$F \quad \log(\mathrm{OR}) = \beta_4 + \beta_6 A + \beta_{10} A^2$$

$$M \quad \log(\mathrm{OR}) = (\beta_4 + \beta_5) + (\beta_6 + \beta_7) A + (\beta_{10} + \beta_{11}) A^2$$

.... and the log of the ratio of Odds Ratios

$$\log\left(\frac{\mathrm{OR}_M}{\mathrm{OR}_F}\right) = \beta_5 + \beta_7 A + \beta_{11} A^2$$

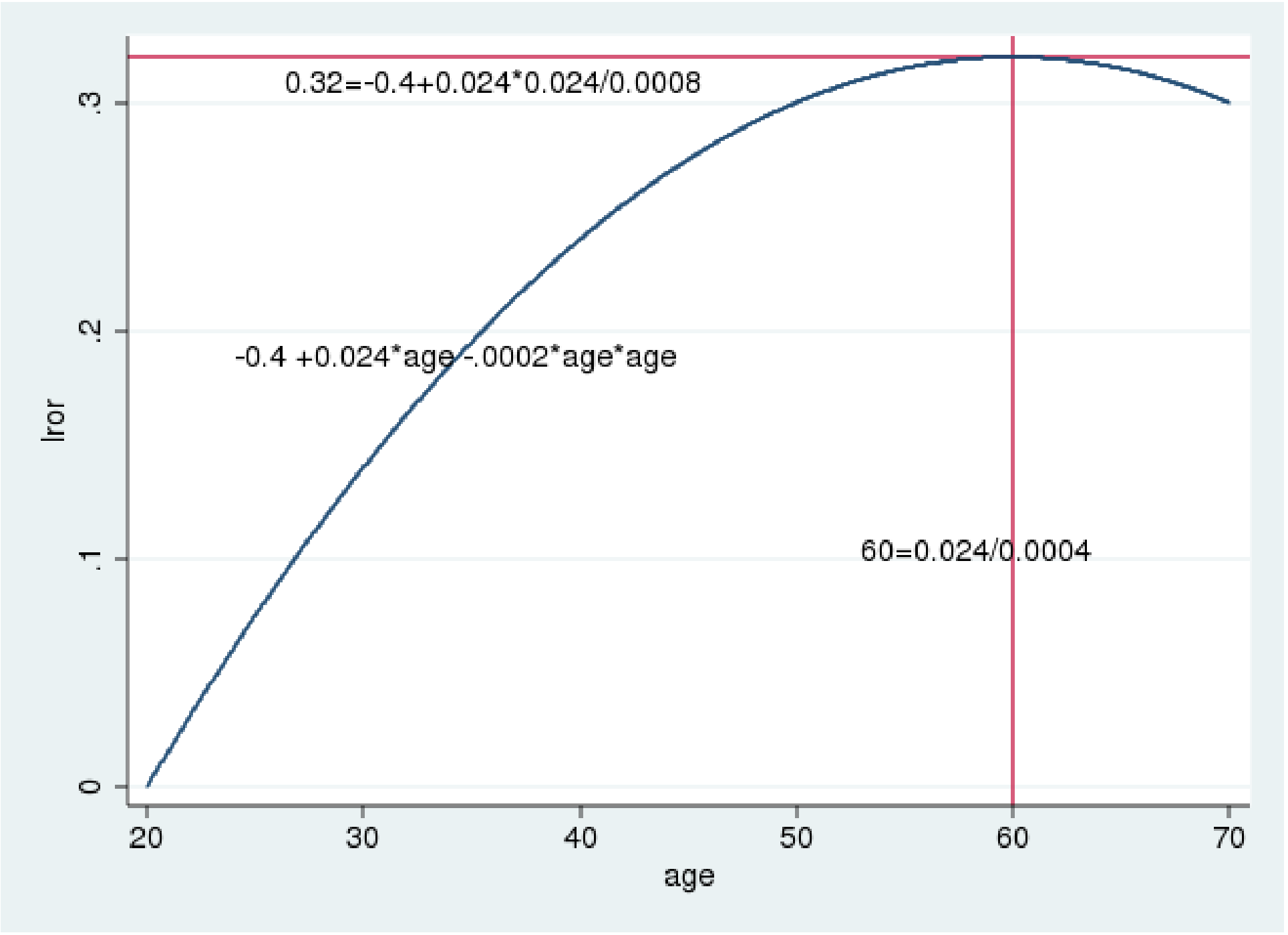
Now assessment would begin with β_{11}

If $\beta_{11} \neq 0$

then, gender modification depends on age and the nature of this modification is nonlinear. The investigator should consider graphs of log of the ratio of odds ratios versus age to study the nature of this modification.

For example, suppose we have

$$\log\left(\frac{\mathrm{OR}_M}{\mathrm{OR}_F}\right) = -0.4 + 0.024\,A - 0.0002\,A^2$$


0.32=-0.4+0.024*0.024/0.0008
-0.4 +0.024*age -.0002*age*age
60=0.024/0.0004
lror
0
.1
.2
.3
20
30
40
50
60
70
age

Interpretation

We see that the log of the ratio of odds ratios rises quite steadily age from age 20 until about age 45, then the curve gradually levels off and is flat at age 60.

There are many other contingencies. For example $\beta_{11}=0$ but $\beta_{10}\neq 0$. Then age modifies and the nonlinear form of the modification does not depend on gender. Again, a graph helps out.

In the absence of modification, age may be confounder but such confounding may only be seen through a correct view of the nonlinear relationship between age and the log of odds of exposure.

For example, one might entertain models like:

$$\log(p/(1-p)) = \beta_0 + \beta_1 G + \beta_2 A + \beta_3 GA + \beta_4 E + \beta_8 A^2 + \beta_9 GA^2$$

$$= \beta_0 + \beta_1 G + \beta_4 E + (\beta_2 + \beta_3 G) A + (\beta_8 + \beta_9 G) A^2$$

to determine whether β_4 changes as various nested models are considered. Joint confounding may be seen through such comparisons as well.

This model provides for 4 parabolae:

For females: 2 curves of the same curvature and the same min/max location differing only by the fixed vertical distance β_4

For males: 2 curves of the same curvature and the same min/max location differing only by the fixed vertical distance β_4

But the curves for the females may be different from the curves for the males

If $\beta_9 \neq 0$ then the curvature for the males will be different from the curvature for the females.

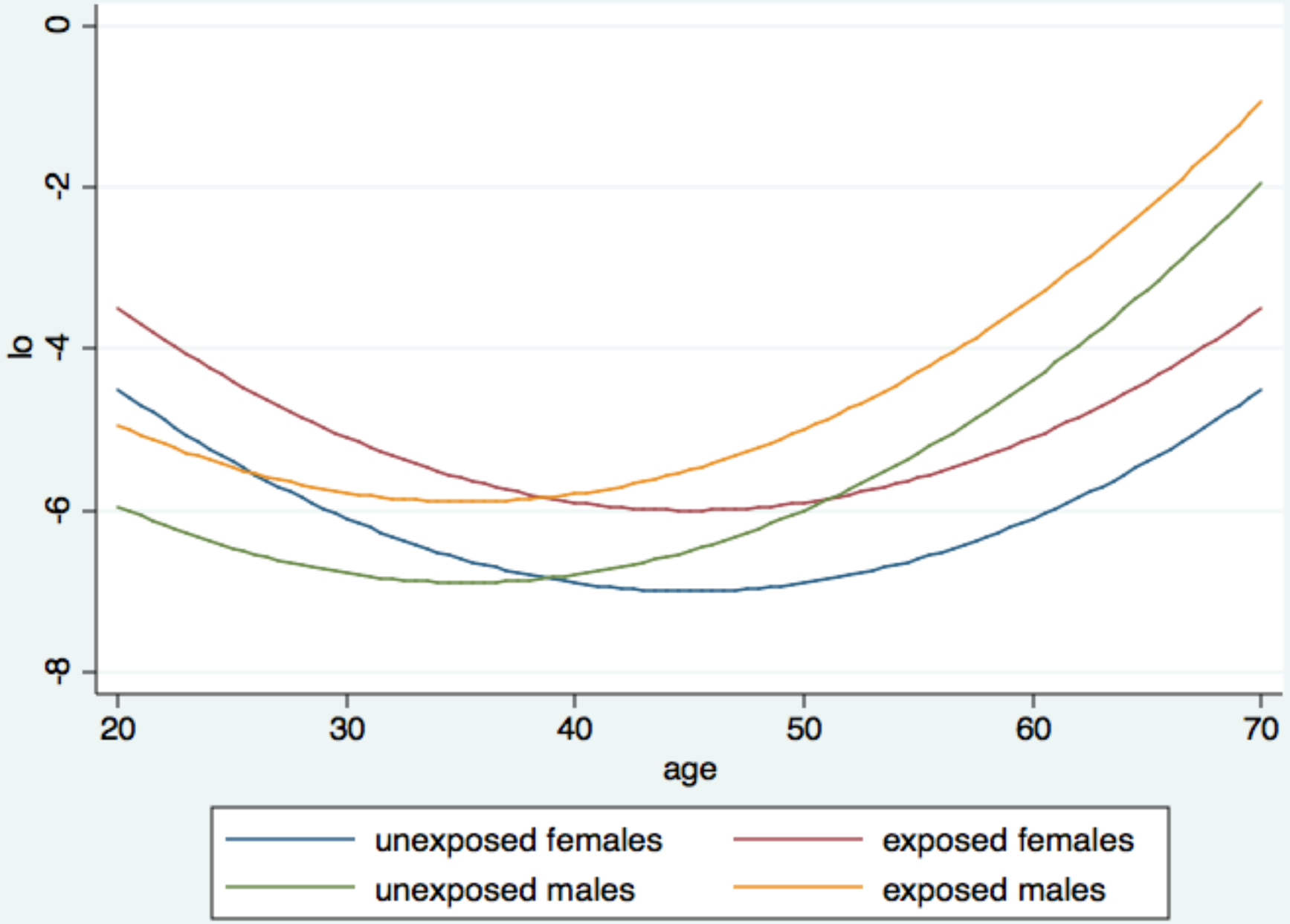
If $\frac{\beta_2}{\beta_8} \neq \frac{\beta_3}{\beta_9}$ then the position of the maximum/minimum for the males is different from the females.

If β_4 for this model is meaningfully different from nested models without nonlinearity terms, then we can see that the model with nonlinearity terms has provided us with insight into the confounding effect of age and gender not attainable from the 'simpler' models.

For example

Consider:

$$\log(p/(1-p)) = -7 + \frac{1}{2}G + \exp + \frac{(2+G)}{25}(A-45) + \frac{2.5+0.05\mathrm{G}}{625}(A-45)^2$$


0
-2
-4
-6
-8
lo
20
30
40
50
60
70
age
unexposed females
exposed females
unexposed males
exposed males

Piecewise Linear

Sometimes, the nonlinearity can be captured using a very simple approach: two lines joined at preselected threshold.

If it is reasonable to consider the relationship to be linear over 2 ranges determined by a threshold (t), one can consider a function like:

$$\begin{aligned} f(x) &= a_1 + b_1 x \quad if\ x \le t \\ &= a_2 + b_2 x \quad if\ x \ge t \end{aligned}$$

The 2 lines are joined at x=t

So we see that: $a_1 + b_1 t = a_2 + b_2 t$

which is the same as: $a_2 = a_1 + b_1 t - b_2 t$

Remember: $f(x) = a_1 + b_1 x \quad if\ x \le t$

For $x \ge t$, we substitute $a_1 + b_1 t - b_2 t$ for a_2

and have $f(x) = a_1 + b_1 t - b_2 t + b_2 x$

$$= a_1 + b_1 t + b_2 (x - t) \quad if\ x \ge t$$

Now let: $\delta(x) = 0 \quad if\ x < t$

$$= 1 \quad if\ x \ge t$$

Using the indicator δ

We consider the equation:

$$\log(p/(1-p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where: $x_1 = x - \delta(x-t)$ and $x_2 = \delta(x-t)$

$x_1 = x$ if $x < t$ and $x_2 = 0$ if $x < t$

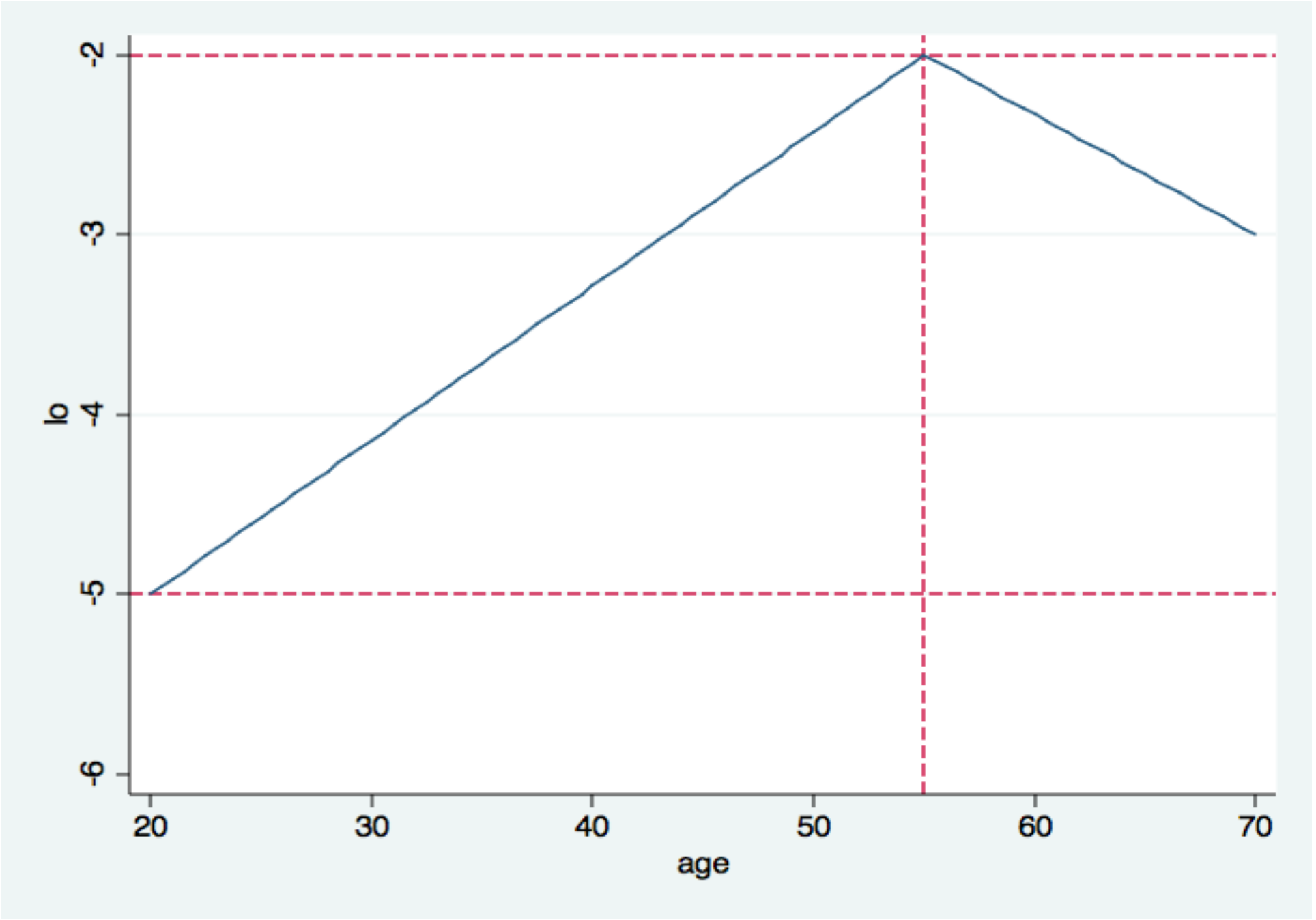
$x_1 = t$ if $x \geq t$ and $x_2 = x - t$ if $x \geq t$

so if $x < t$, $\log(p/(1-p)) = \beta_0 + \beta_1 x$

if $x \geq t$, $\log(p/(1-p)) = \beta_0 + \beta_1 t + \beta_2 (x-t)$

A picture helps here: $\log(p/(1-p)) =$

$$-5 - \frac{60}{35} + \frac{3}{35}(A - \delta(A-55)) - \frac{1}{15}\delta(A-55)$$


-2
-3
-4
-5
-6
lo
20
30
40
50
60
70
age

Two lines joined

We see that the log odds rises with slope 3/35 until age 55 and declines with slope -1/15 after age 55

It is acknowledged that such a picture has a sharp edge at age 55 (not realistic, really)

Further, the selection of the threshold

(t=55, here) may be somewhat arbitrary and hard to justify.

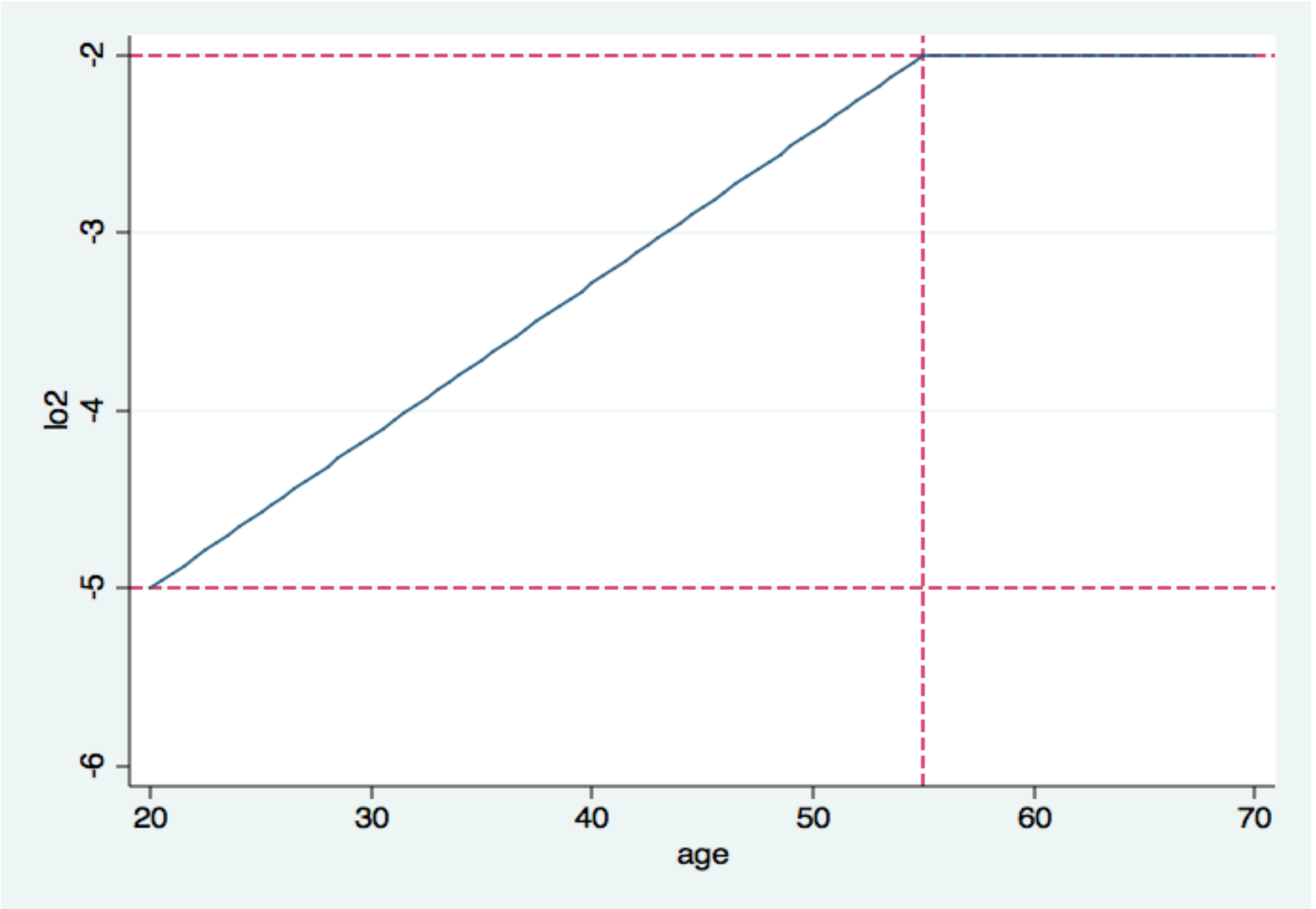
Consideration of $\beta_2 = 0$

Then the line is horizontal after age =55 and we might get:

$$\log(p/(1-p)) = -5 - \frac{60}{35} + \frac{3}{35}(A - \delta(A-55))$$

which is a line of slope 3/35 until age 55 and then it is a horizontal line (slope of 0).

So the log odds rises until age 55 and then it is constant there after.


lo2
-2
-3
-4
-5
-6
20
30
40
50
60
70
age

If age is a potential modifier or a potential confounder...

… we can use all of the techniques developed for the parabolic curves now with the piecewise linear 'curves' by replacing:

$$A \text{ and } A^2 \text{ with } A_1 = A - \delta(A - t) \text{ and } A_2 = \delta(A - t)$$

An illustration

Consider a study of individuals with a gambling addiction. Investigators plan to compare 2 interventions. Let us further suppose that the outcome is whether an individual gambled more than once a week during a 3 month long period (a failure). The intervention begins after the first 3 month period. (a baseline period) Then participants were followed every 3 months for one year.

Perhaps the investigators suspect that the potential benefit of the interventions will be largest in the initial 3 period after the intervention and then afterwards the benefit may be more modest. They plan to model a comparison between the first 3 month period (post baseline) and the baseline period and then allow for a linear trend to compare the last period with the first period

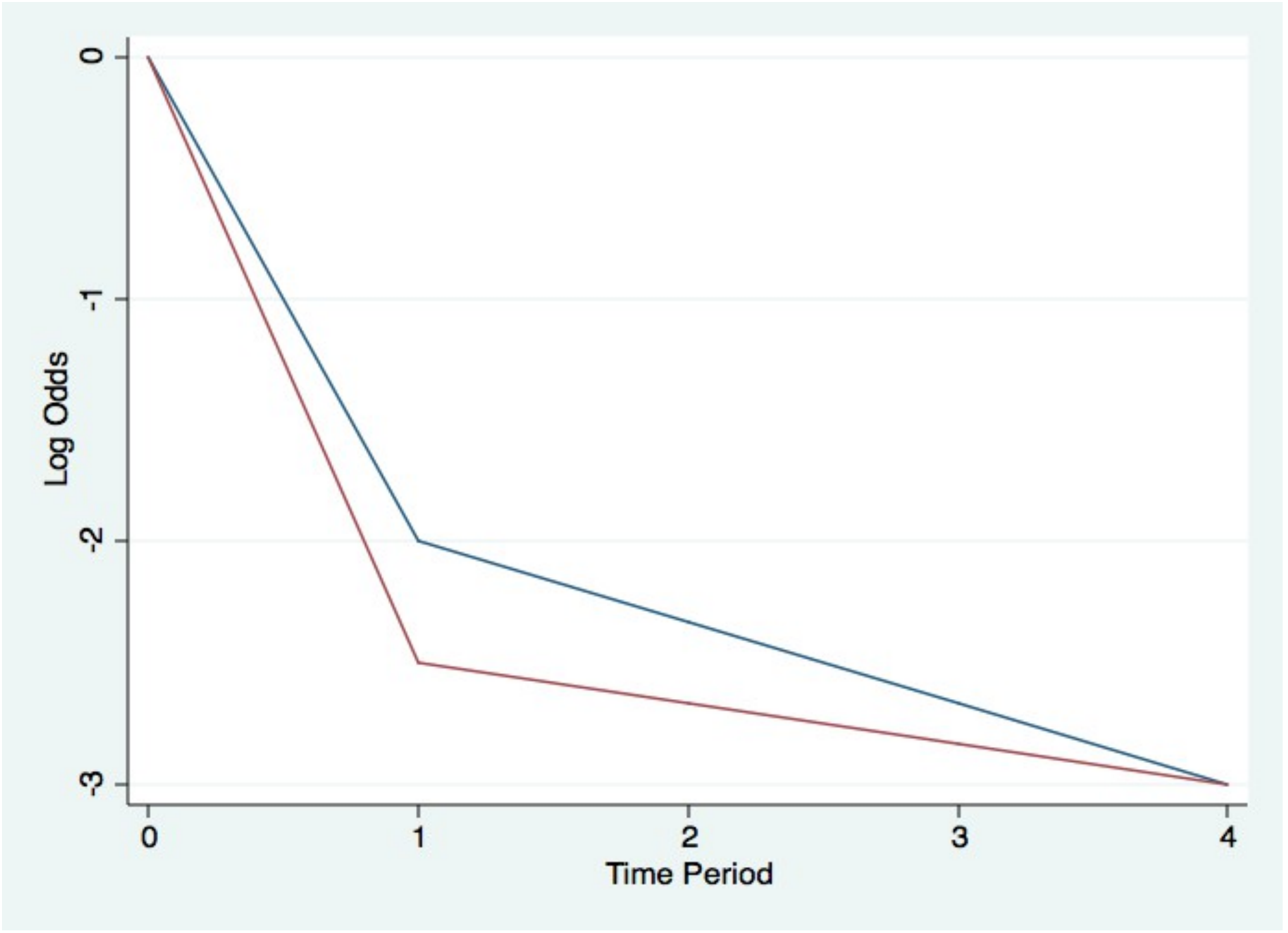
If p=Pr(Failure) and t=1

.... then one might consider:

$$\log(p/(1-p)) = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 E + \beta_4 EA_1 + \beta_5 EA_2$$

where A records the time periods:

0 (baseline), 1, 2, 3, 4


0
-1
-2
-3
Log Odds
0
1
2
3
4
Time Period

Polynomials and More Elaborate Functions

Beyond parabolae, one can consider polynomials:

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \ \ldots \ \ \beta_m x^m$$

One rarely sees polynomials with m>3 though. Such curves often have odd shapes that cannot be the basis for reality.

Transformations

Rather than polynomials, sometimes transformations are warranted:

The most often choices are:

logarithms, square roots, reciprocals and reciprocals of square roots

Thinking logarithm as a power of 0 (not really), the choices are all 'powers' and fractional 'powers':

The above list would be : 0, 1/2, -1, -1/2

More Elaborate Functions

Or, sometimes, functions with powers like above, fractional powers or logarithms:

$$\beta_0 + \beta_1 x^{(p_1)} + \beta_2 x^{(p_2)} + \quad \ldots \quad \beta_m x^{(p_m)}$$

where:

$$\begin{aligned} x^{(p)} &= x^p && \text{if } p \neq 0 \\ &= \log(x) && \text{if } p = 0 \end{aligned}$$

For example

One can also multiply powers by logarithms too

$$\beta_0+\beta_1 x^{-1}+\beta_2 x+\beta_3 x^3+\beta_4 x^3 \log(x)$$

If one plans to consider 'powers' like: $\frac{1}{2}$ or $-\frac{1}{2}$ or logarithm, then one must ensure that x is positive.

There are many other choices in specialized circumstances.

Restricted Cubic Splines

One can have more than two lines joined at places called knots

Piecewise linear functions are examples of a much more general class of nonlinear functions

We can have piecewise quadratics, piecewise cubics

Cubics can be joined at knots so that the joining points are smooth.

Restricted Cubic Splines are piecewise cubics; smooth at joins and linear at the ends

The Stata command 'mkspline' provides for all of these functions