Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 19 :

Hidden Extrapolation

The Inappropriate Use of Explanatory Variables

The Importance of Being Ernest About Graphs

Let us now consider another challenge not yet discussed. A simple generic setting will suffice. An outcome (S), an exposure (E) and subject's age (A).
One might start with a crude analysis ignoring age.

```
. cc out1 expo
                                                                Proportion
                 |   Exposed   Unexposed  |      Total      Exposed
-----------------+------------------------+------------------------
           Cases |        65            9  |         74       0.8784
        Controls |       185          241  |        426       0.4343
-----------------+------------------------+------------------------
           Total |       250          250  |        500       0.5000
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         9.408408       |    4.493071    21.97535 (exact)
 Attr. frac. ex. |         .8937121       |    .7774351    .9544945 (exact)
 Attr. frac. pop |         .7850174       |
                 +-----------------------------------------------
                           chi2(1) =    49.74  Pr>chi2 = 0.0000
. logit out1 expo

Logistic regression                               Number of obs   =        500
                                                  LR chi2(1)      =      55.19
                                                  Prob > chi2     =     0.0000
Log likelihood = -182.01838                       Pseudo R2       =     0.1316

------------------------------------------------------------------------------
        out1 |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expo |   2.241604   .3688426     6.08   0.000     1.518686    2.964522
       _cons |  -3.287572   .3394921    -9.68   0.000    -3.952965    -2.62218
------------------------------------------------------------------------------
```

The Crude Analysis

There is evidence of a disease exposure relationship.

The estimated OR is 9.41

The p-value is less than 0.1 %

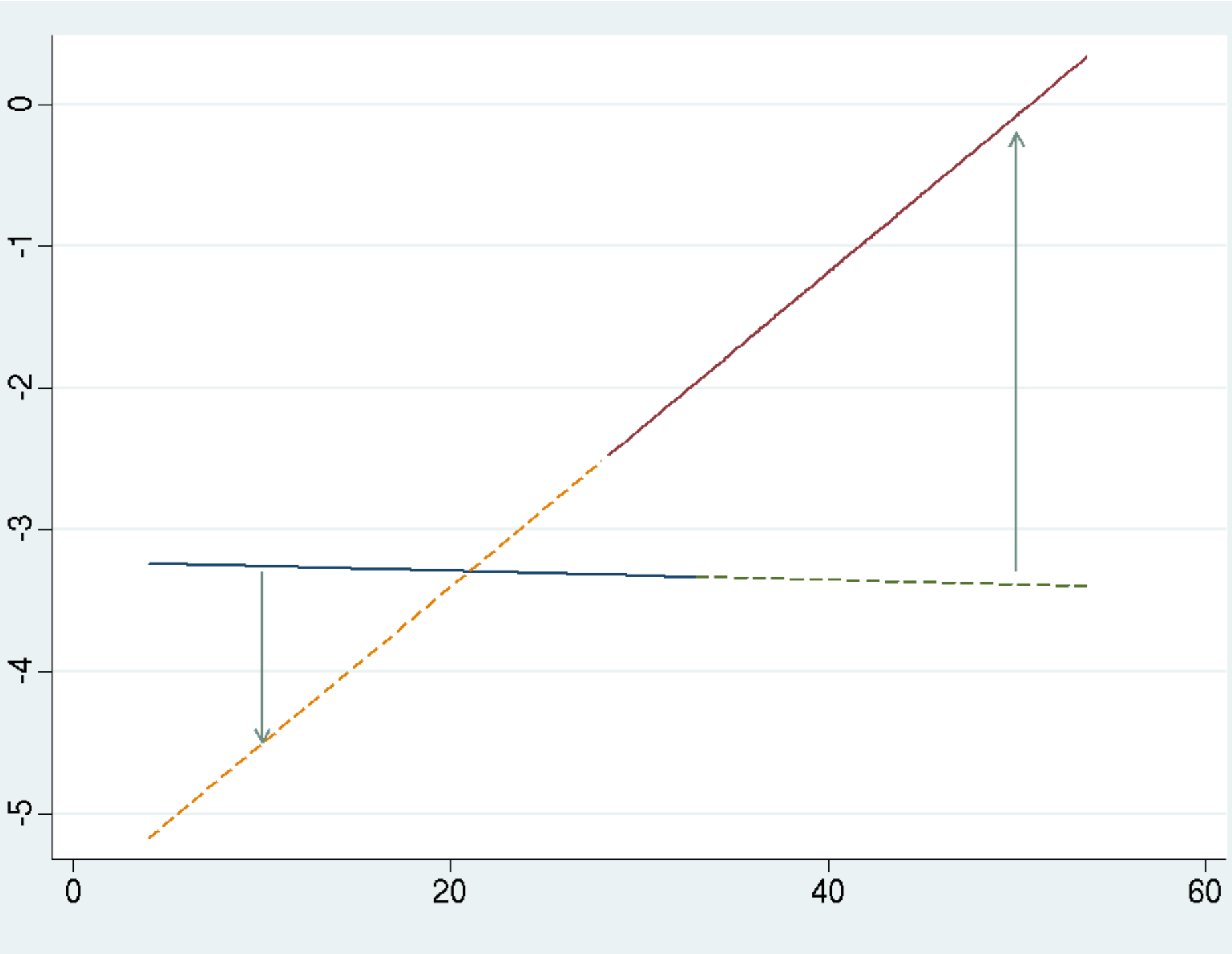
The lower limit for the confidence interval for OR is 4.49 (and well above one)

BUT, what about a 'proper' analysis that begins with assessment of modification/confounding?

```
. gen ae=age*expo

. logit out1 expo age ae

Logistic regression                                     Number of obs   =        500
                                                        LR chi2(3)      =      68.02
                                                        Prob > chi2     =     0.0000
Log likelihood = -175.60304                             Pseudo R2       =     0.1622

------------------------------------------------------------------------------
        out1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expo |  -2.392244    1.90559    -1.26   0.209    -6.127133    1.342644
         age |  -.0032749   .0666619    -0.05   0.961    -.1339298      .12738
          ae |   .1140871   .0739871     1.54   0.123     -.030925    .2590991
       _cons |  -3.223443   1.346249    -2.39   0.017    -5.862043    -.584843
------------------------------------------------------------------------------

. predict yh,xb
. twoway (line yh age if expo==0 & ex==0) (line yh age if expo==1 & ex==0) (line
yh age if expo==0 & ex==1,lpattern(-)) (line yh age if expo==1 &
ex==1,lpattern(-)) (pcarrow var13 var14 var15 var16),legend(off)
```

The slopes are quite different

For illustration, let us suppose that we have judged age to be a modifier [even though the appropriate p-value is 0.123 and is not less than 5%]. A careful assessment of this model reveals a much more serious issue. Lets look at a plot of the fitted values versus age for the exposed and unexposed.

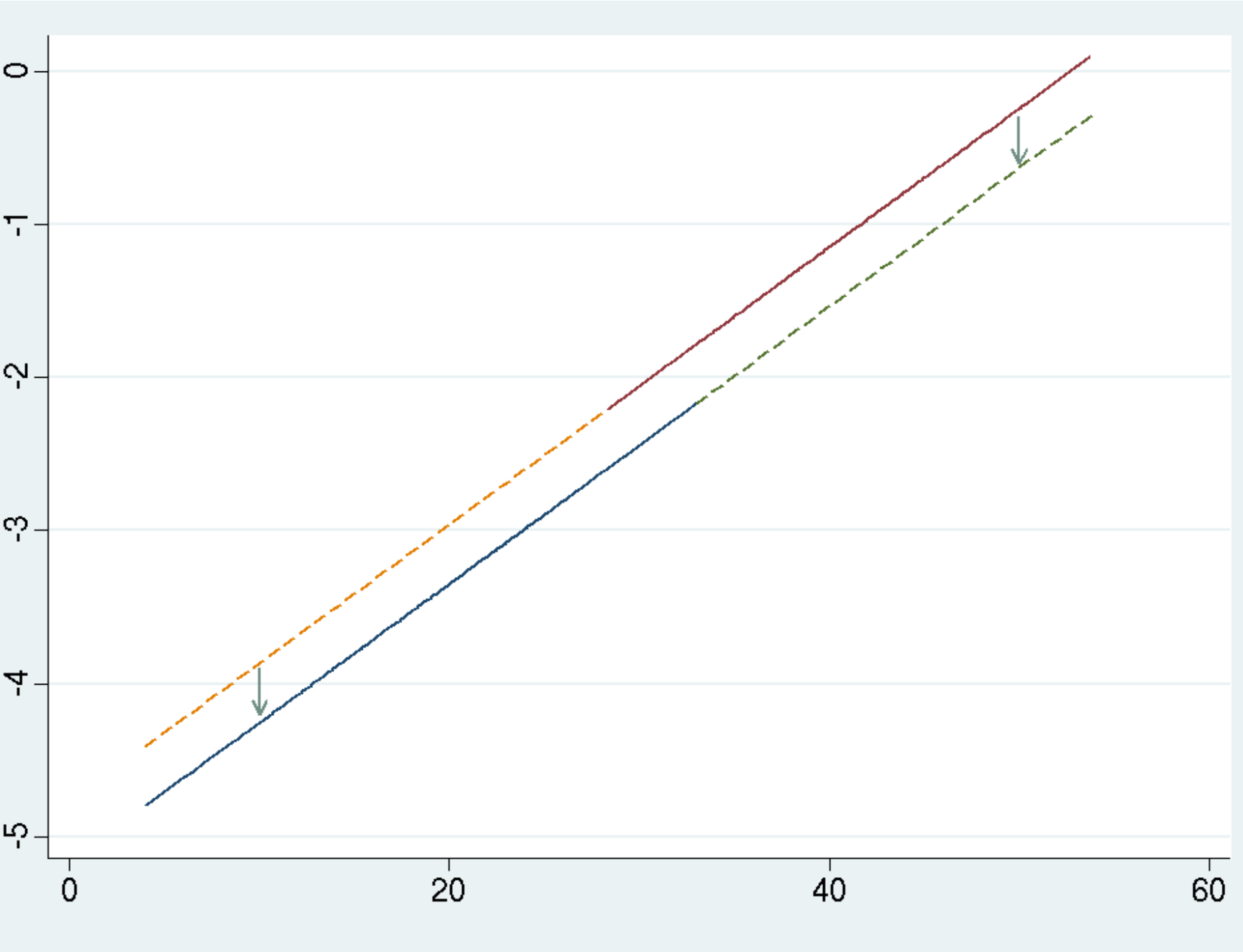

0
-1
-2
-3
-4
-5
0
20
40
60

The blue line is for the unexposed, the red line is for the exposed. Almost all the unexposed are younger than the exposed. There is almost no overlap in age distributions. This could have been seen from a boxplot of the ages by exposure as well. The orange dotted line extends the red line to nonexistent young exposed while the dotted green line extends the blue line to nonexistent old unexposed. Notice that an age specific comparison between exposed and unexposed involves an extrapolation of one of the red or blue lines to individuals that did not exist in the study. This is illustrated for 10 year olds and for 50 year olds with arrows.

What if we had proceeded to assess the model that gives parallel lines

```
. logit out1 expo age

Logistic regression                                     Number of obs   =        500
                                                        LR chi2(2)      =      65.69
                                                        Prob > chi2     =     0.0000
Log likelihood = -176.76674                             Pseudo R2       =     0.1567

------------------------------------------------------------------------------
        out1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expo |   .3883246   .6794215     0.57   0.568    -.9433171    1.719966
         age |   .0905566   .0285993     3.17   0.002     .0345031    .1466101
       _cons |  -5.159796   .7077351    -7.29   0.000    -6.546932   -3.772661
------------------------------------------------------------------------------
. predict yh,xb

twoway (line yh age if expo==0 & ex==0) (line yh age if expo==1 & ex==0) (line yh
age if expo==0 & ex==1,lpattern(-)) (line yh age if expo==1 & ex==1,lpattern(-))
(pcarrow var13 var14 var15 var16),legend(off)
```

Based on this analysis compared with the crude analysis, we might suggest that age is a confounder and there is no outcome/exposure relationship. Even though the crude analysis indicates such a relationship. A careful assessment of these models reveals the same serious issue as the previous. Lets look at a plot of the fitted values versus age for the exposed and unexposed.


0
-1
-2
-3
-4
-5
0
20
40
60

Recall that 0.388 is the vertical distance between the 2 lines but such a vertical distance requires extending the blue line to the older age range to conceptualize this vertical distance at older ages. Such extrapolation is not justified as there were no older exposed subjects. Similiarly, we must extend the red line to the younger age range but they do not exist among the exposed. So here, the interpretation of 'adjustment' is not possible. The number 0.388 refers to no set of subjects at a given age.

So age should not be used for adjustment here. This is not at all clear until one considers the graphs. The Stata analysis output gives no cue to trouble.

The Overlap in Ages

There 250 unexposed persons in the 'study' and 250 exposed persons in the 'study'.

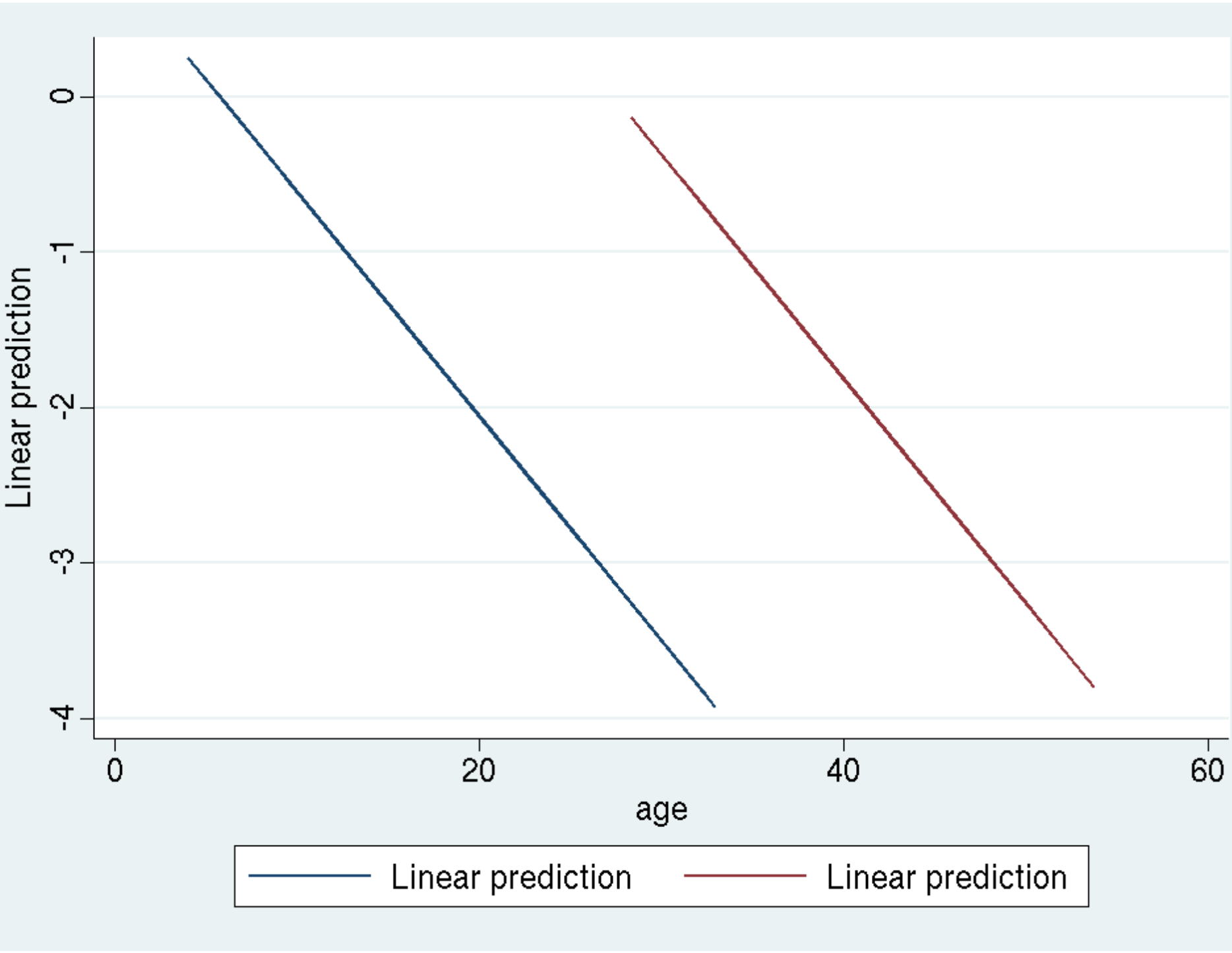
The oldest unexposed person was 33 while the youngest exposed person was 28.

There were only 11 unexposed in the overlap and only 14 exposed in the overlap.

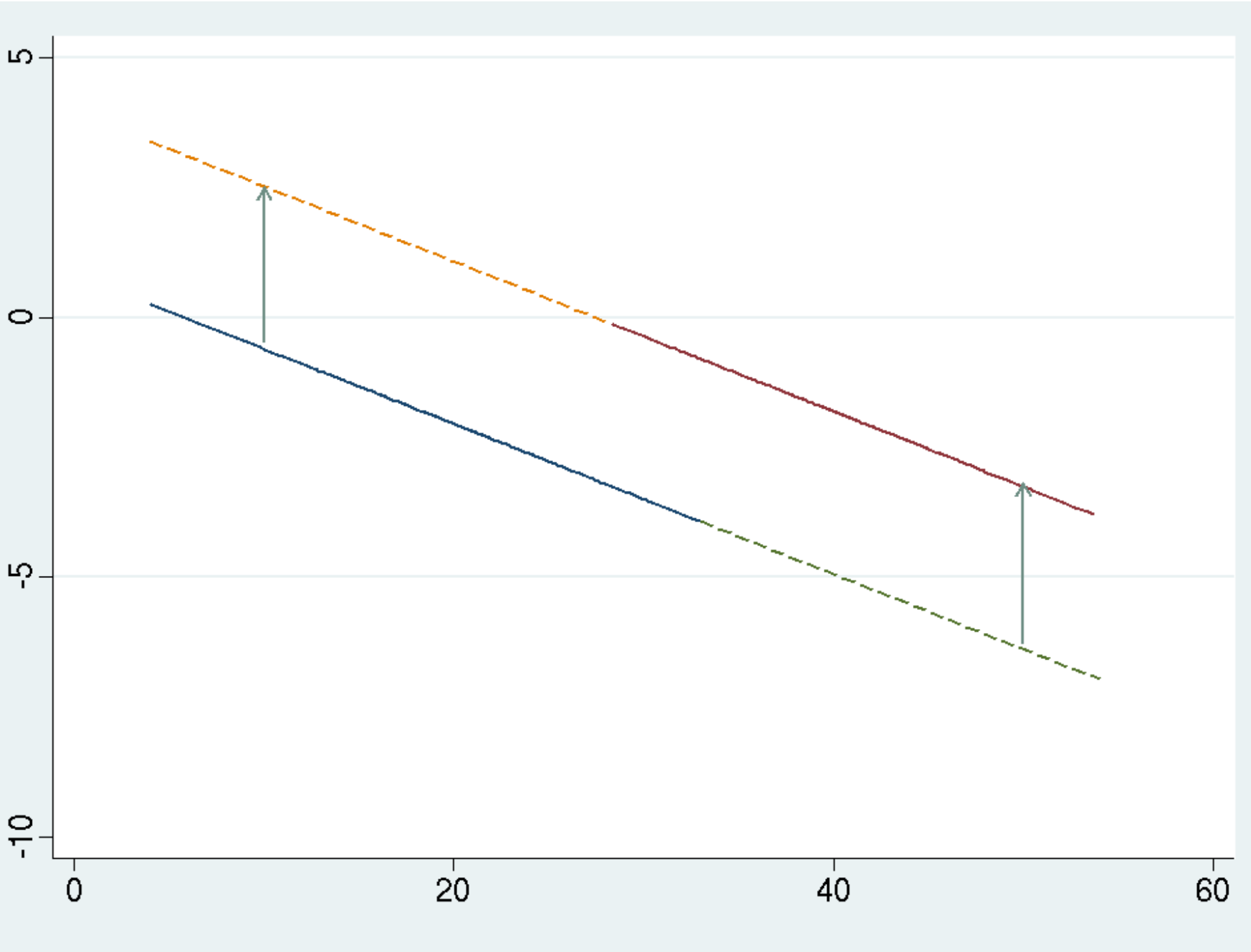
Here is another example:

```
. cc out2 expo
                                                        Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        37           35  |         72      0.5139
        Controls |       213          215  |        428      0.4977
-----------------+------------------------+------------------------
           Total |       250          250  |        500      0.5000
                 |                         |
                 |      Point estimate     |    [95% Conf. Interval]
                 |-------------------------+------------------------
      Odds ratio |          1.067069       |    .6276629    1.816107 (exact)
                 +-------------------------------------------------
                               chi2(1) =     0.06  Pr>chi2 = 0.7989

. logit out2 expo age ae

Logistic regression                               Number of obs   =        500
                                                  LR chi2(3)      =      29.91
                                                  Prob > chi2     =     0.0000
Log likelihood = -191.12345                       Pseudo R2       =     0.0726

------------------------------------------------------------------------------
        out2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expo |   4.389939    1.76595     2.49   0.013     .9287407    7.851138
         age |  -.1247719   .0373285    -3.34   0.001    -.1979345   -.0516093
          ae |  -.0436512   .0564193    -0.77   0.439    -.1542309    .0669285
       _cons |   .4928755   .6784046     0.73   0.468     -.836773    1.822524
------------------------------------------------------------------------------
```

```
. logit out2 expo age

Logistic regression                                     Number of obs   =        500
                                                        LR chi2(2)      =      29.31
                                                        Prob > chi2     =     0.0000
Log likelihood = -191.42448                             Pseudo R2       =     0.0711

------------------------------------------------------------------------------
        out2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expo |   3.128248   .6523577     4.80   0.000      1.84965    4.406845
         age |  -.1444681   .0279461    -5.17   0.000    -.1992415   -.0896947
       _cons |   .8342271   .5163806     1.62   0.106    -.1778603    1.846314
------------------------------------------------------------------------------

. predict yh,xb

. twoway (line yh age if expo==0) (line yh age if expo==1)
```


Linear prediction
0
-1
-2
-3
-4
0
20
40
60
age
Linear prediction
Linear prediction

For this example, the crude shows no outcome/exposure relationship while the adjusted shows a 'strong' outcome/exposure relationship. The graph that accompanies the analysis is a little harder to appreciate. Notice again that there is little overlap in age distributions and there is no meaningful interpretation for the vertical distance between the 2 lines. It is instructive to draw the 2 lines extended to 'Fantasyland' to appreciate the folly in the adjusted here.


5
0
-5
-10
0
20
40
60

It is also true that the use of incorrect adjustment may not be revealed by empirical analysis but rather the identification of an inappropriate adjuster comes from the context, the literature and, indeed, the use of ...er common sense.

Consider a study of 2 ICU interventions on (say) 30 day survival rates. It is common to consider characteristics measured at entry to an ICU [baseline] as candidates for confounding or modification. However, it is clear that one should not consider measures taken just before death [or discharge from ICU if alive] as candidates since such measures would surely swamp any measurable distinction between the interventions. Even without empirical support, most [all?] ICU researchers would agree that adjustment for late measures is foolhardy [and pointless]

Alas, in most real epidemiologic investigations, the identification may not be so black and white but rather involve [many] shades of grey.

One more example should help to illustrate this matter in that the context is not straightforward.

Consider a small part of the data from the Sloane Epidemiology Unit Birth Defects Study.

The outcome is neural tube defect (D=1). The exposure is supplementation with Folic Acid (E=1). The variable under consideration as a confounder or modifier is stillbirth/induced abortion (stillbirth is C=1)

```
. cc d e,by(c)

                c |       OR        [95% Conf. Interval]   M-H Weight
------------------+-------------------------------------------------
                0 |   .7272727        .433258    1.18223     21.56306 (exact)
                1 |     1.0925       .4182351    3.10377     4.624277 (exact)
------------------+-------------------------------------------------
            Crude |   .6528922       .4436503   .9459948               (exact)
    M-H combined |   .7917662       .5240046   1.196352
-------------------------------------------------------------------
Test of homogeneity (M-H)       chi2(1) =     0.62  Pr>chi2 = 0.4313

                     Test that combined OR = 1:
                                 Mantel-Haenszel chi2(1) =      1.21
                                                 Pr>chi2 =    0.2720
```

It might be argued that the crude estimate is 'meaningfully' different from the adjusted estimate [in part, by the magnitude of the relative change and, in part, since the crude analysis yields a p-value less than 5% while the adjusted analysis does not] and, hence, the adjusted analysis is preferred. On the other hand, what business do we have in adjusting for C in the first place? A review of the literature in birth defects reveals that both points of view have been entertained. We do not have the luxury of an empirical assessment here to offer guidance. Here, it might be argued that C 'occurs' after both E and D [even this statement is slippery] and so adjustment for C is, at least, tenuous.