Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 9 :
The Fit :
Characteristics and Assessment

Two Situations to Consider

1) Models with a fixed set of distinct fitted values
  - typically seen in models that deal with stratified analyses

2) Models with a potentially different fitted value for each individual.
  - typically seen in models with a measured predictor such as actual age, actual height etc...

1) Models That Determine a Fixed Set of Fitted Values

A Return to 4 Strata: Age and Gender

Lets us consider the study of a disease/exposure relationship with age group (Y O) and gender (F M) as potential modifiers/confounders.

The table from ' table e g a,c(mean d) '

records the 8 estimates of the conditional probabilities: the 'observed' proportions

```
. cs dis exp,by(age gender) or

      age gender |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
             0 0 |    1.179451     .7858152   1.770499        21.12 (Cornfield)
             0 1 |    .7972632     .5030501   1.264035       20.608 (Cornfield)
             1 0 |    .9949764     .6448348   1.535662       20.304 (Cornfield)
             1 1 |    .8470745     .5651307   1.269885       25.568 (Cornfield)
-----------------+-------------------------------------------------
           Crude |    1.840806     1.540884   2.199106
    M-H combined |    .9497717     .7678975   1.174722
-------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(3) =    1.983  Pr>chi2 = 0.5758

                   Test that combined OR = 1:
                               Mantel-Haenszel chi2(1) =      0.23
                                                Pr>chi2 =    0.6350
. table e g a,c(mean d)

----------------------------------------------
          |                a and g
          | -------- y -------    -------- o -------
        e |        f         m           f         m
----------+-----------------------------------------
       ne | .6015037  .2970822    .2942779  .6573427
        e |  .640327  .2520325    .2932331  .6190476
----------------------------------------------
```

The model for p=Pr(D)

would be:

$$\log(\mathrm{p}/(1-\mathrm{p}))=\beta_0+\beta_1 G+\beta_2 A+\beta_3 GA+\beta_4 E+\beta_5 EG+\beta_6 EA+\beta_7 EGA$$

and this model gives fitted values for the log odds

$$\log(\hat{\mathrm{p}}/(1-\hat{\mathrm{p}}))=b_0+b_1 G+b_2 A+b_3 GA+b_4 E+b_5 EG+b_6 EA+b_7 EGA$$

for each individual in the study.

There are, however, only 8 different fitted values. One distinct value for each combination of E, G and A.

Model 1

```
. logit d g a ga e eg ea ega


Logistic regression                                     Number of obs   =       2000
                                                        LR chi2(7)      =     239.85
                                                        Prob > chi2     =     0.0000
Log likelihood = -1259.9624                             Pseudo R2       =     0.0869

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           g |  -1.272966   .2099288    -6.06   0.000    -1.684418   -.8615129
           a |  -1.286432   .2109222    -6.10   0.000    -1.699831   -.8730315
          ga |   2.799137   .2970468     9.42   0.000     2.216936    3.381338
           e |   .1650489   .2078437     0.79   0.427    -.2423173     .572415
          eg |  -.3916184   .3146884    -1.24   0.213    -1.008396    .2251594
          ea |  -.1700852   .3043002    -0.56   0.576    -.7665025    .4263322
         ega |   .2306881   .4374387     0.53   0.598    -.6266761    1.088052
       _cons |   .4117347   .1771099     2.32   0.020     .0646057    .7588638
------------------------------------------------------------------------------

. predict lohat,xb
. predict phat,p
. est stor m1
```

The predict option xb gives the fitted values on the log odds scale: the fitted log odds

The predict option p gives the fitted values on the probability scale: the fitted proportions

```
. tab lohat

     Linear |
 prediction |      Freq.     Percent        Cum.
------------+-----------------------------------
  -1.087801 |        123        6.15        6.15
  -.8797331 |        133        6.65       12.80
  -.8746968 |        367       18.35       31.15
  -.8612309 |        377       18.85       50.00
   .4117347 |        133        6.65       56.65
   .4855078 |        357       17.85       74.50
   .5767836 |        367       18.35       92.85
   .6514745 |        143        7.15      100.00
------------+-----------------------------------
      Total |      2,000      100.00

. tab phat

      Pr(d) |      Freq.     Percent        Cum.
------------+-----------------------------------
   .2520327 |        123        6.15        6.15
   .2932331 |        133        6.65       12.80
   .2942779 |        367       18.35       31.15
   .2970822 |        377       18.85       50.00
   .6015037 |        133        6.65       56.65
   .6190476 |        357       17.85       74.50
    .640327 |        367       18.35       92.85
   .6573427 |        143        7.15      100.00
------------+-----------------------------------
      Total |      2,000      100.00
```

Fitted Proportions

Notice that, for this model here, the fitted proportions are the same as the observed proportions. The model fits the data (the observed proportions) exactly.

There are 8 regression coefficients and these coefficients directly correspond to the 8 conditional probabilities.

Any simpler model “nested” within this model will inevitably yield fitted values that do not reproduce the observed proportions.

Model Assessment

Likelihood ratio tests essentially compare the fitted values obtained from 2 candidate models: one model nested within the other model.

In principle, an assessment the quality of a fit should include comparing fitted values either through analytic testing methods or through graphical methods.

Now consider the following nested model:

Model 2

```
. logit d g a ga e


Logistic regression                                     Number of obs   =       2000
                                                        LR chi2(4)      =     237.86
                                                        Prob > chi2     =     0.0000
Log likelihood = -1260.9553                             Pseudo R2       =     0.0862

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           g |  -1.472424   .1457904   -10.10   0.000    -1.758168    -1.18668
           a |  -1.432539   .1444609    -9.92   0.000    -1.715677   -1.149401
          ga |   2.903936   .2172565    13.37   0.000     2.478121    3.329751
           e |  -.0515205   .1084388    -0.48   0.635    -.2640565    .1610156
       _cons |    .570101   .1223205     4.66   0.000     .3303572    .8098449
------------------------------------------------------------------------------

. predict phat2,p
. est stor m2
```

```
. tab phat2

      Pr(d) |      Freq.     Percent        Cum.
------------+-----------------------------------
   .2781126 |        123        6.15        6.15
   .2861905 |        133        6.65       12.80
   .2885734 |        377       18.85       31.65
   .2968302 |        367       18.35       50.00
   .6265755 |        357       17.85       67.85
   .6268158 |        367       18.35       86.20
   .6385494 |        143        7.15       93.35
   .6387865 |        133        6.65      100.00
------------+-----------------------------------
      Total |      2,000      100.00

. gen diff2=phat-phat2

. tab diff2

      diff2 |      Freq.     Percent        Cum.
------------+-----------------------------------
  -.0372828 |        133        6.65        6.65
  -.0260799 |        123        6.15       12.80
  -.0075278 |        357       17.85       30.65
  -.0025522 |        367       18.35       49.00
   .0070426 |        133        6.65       55.65
   .0085089 |        377       18.85       74.50
   .0135112 |        367       18.35       92.85
   .0187933 |        143        7.15      100.00
------------+-----------------------------------
      Total |      2,000      100.00
```

Residuals

The residuals are the differences between the observed and the fitted.

Here, we consider diff2 and we see very tiny residual values.

Now consider the following model:

Model 3

```
. logit d g a e


Logistic regression                                     Number of obs   =       2000
                                                        LR chi2(3)      =      45.75
                                                        Prob > chi2     =     0.0000
Log likelihood = -1357.0128                             Pseudo R2       =     0.0166

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           g |  -.0042679   .0907935    -0.05   0.963      -.18222    .1736841
           a |   .0165261   .0907765     0.18   0.856    -.1613925    .1944448
           e |   .6101303   .0907885     6.72   0.000      .432188    .7880725
       _cons |  -.4690968   .0910724    -5.15   0.000    -.6475956   -.2905981
------------------------------------------------------------------------------

. predict phat3,p
. est stor m3
```

```
. tab phat3

      Pr(d) |      Freq.     Percent        Cum.
------------+-----------------------------------
   .3838201 |        377       18.85       18.85
     .38483 |        133        6.65       25.50
   .3877361 |        143        7.15       32.65
   .3887497 |        367       18.35       51.00
   .5341382 |        123        6.15       57.15
      .5352 |        367       18.35       75.50
   .5382481 |        357       17.85       93.35
   .5393086 |        133        6.65      100.00
------------+-----------------------------------
      Total |      2,000      100.00

. gen diff3=phat-phat3

. tab diff3

      diff3 |      Freq.     Percent        Cum.
------------+-----------------------------------
  -.2821055 |        123        6.15        6.15
  -.2460755 |        133        6.65       12.80
  -.0944718 |        367       18.35       31.15
  -.0867379 |        377       18.85       50.00
   .0807996 |        357       17.85       67.85
    .105127 |        367       18.35       86.20
   .2166737 |        133        6.65       92.85
   .2696066 |        143        7.15      100.00
------------+-----------------------------------
      Total |      2,000      100.00
```

Fit Assessment

The residuals comparing Model 3 with Model 1 are large.

The residuals comparing Model 2 with Model 1 are much smaller.

Our primary assessment would be based on the epidemiology but we can also observe the comparison in the quality of the fit.

The likelihood ratio tests would be:

```
. lrtest m1 m3

Likelihood-ratio test                                 LR chi2(4)  =    194.10
(Assumption: m3 nested in m1)                         Prob > chi2 =    0.0000

. lrtest m1 m2

Likelihood-ratio test                                 LR chi2(3)  =      1.99
(Assumption: m2 nested in m1)                         Prob > chi2 =    0.5753
```

Coronary Heart Disease and Smoking

Now let us consider a study of CHD and smoking status. Age at entry was 'measured' at entry into the study. Here, only integer ages are available: ranging from 39 to 59. This was a large study and so there were reasonable numbers to study the CHD/Smoking relationship in each of the 21 age 'groups'.

```
. cc chd69 smoke,by(age)

              Age |       OR        [95% Conf. Interval]   M-H Weight
------------------+--------------------------------------------------
               39 |          2       .6966061   6.191832            3 (exact)
               40 |   .3966942       .0677972   1.641754     3.931408 (exact)
               41 |   4.525862       .9284312   43.18059     .9957082 (exact)
               42 |   5.644444       .5428601   280.0709     .4054054 (exact)
               43 |   1.156146       .3089201   4.175369          2.8 (exact)
               44 |   6.354369       1.332136   59.90066     .8765957 (exact)
               45 |      5.875       1.191274   56.22355     .8602151 (exact)
               46 |   3.141553       .7159255   18.92927     1.288235 (exact)
               47 |   .8783784       .1570899   4.919971     2.013605 (exact)
               48 |   3.748387       1.186502   13.90961     1.890244 (exact)
               49 |        1.16       .4101005   3.311061     4.104478 (exact)
               50 |       2.125        .506783   10.35947     1.659259 (exact)
               51 |   3.504902       .9886477   15.57106     1.658537 (exact)
               52 |   1.698113        .483206   6.784542     2.345133 (exact)
               53 |   1.559091       .3915896   6.678235     2.095238 (exact)
               54 |         7.5       1.413465   73.77356     .6728972 (exact)
               55 |   1.058824       .2219518   5.044204        2.125 (exact)
               56 |   1.285714       .3060225   5.372781     2.210526 (exact)
               57 |         .75        .100976   4.903224     1.777778 (exact)
               58 |          .2       .0040691    2.01862     2.232143 (exact)
               59 |   1.309524       .2605058   6.275484     1.787234 (exact)
------------------+--------------------------------------------------
            Crude |   1.877353       1.434086   2.465718               (exact)
    M-H combined |   1.883757       1.444799   2.456078
------------------------------------------------------------------------
```

The right model should help here

We can see that:

1) a number of the OR estimates are much larger than 1

2) these larger OR estimates are at the intermediate ages

Perhaps the pattern among the estimates of the log odds can be seen with the right display.

We can get these estimates from the corresponding logistic regression model:

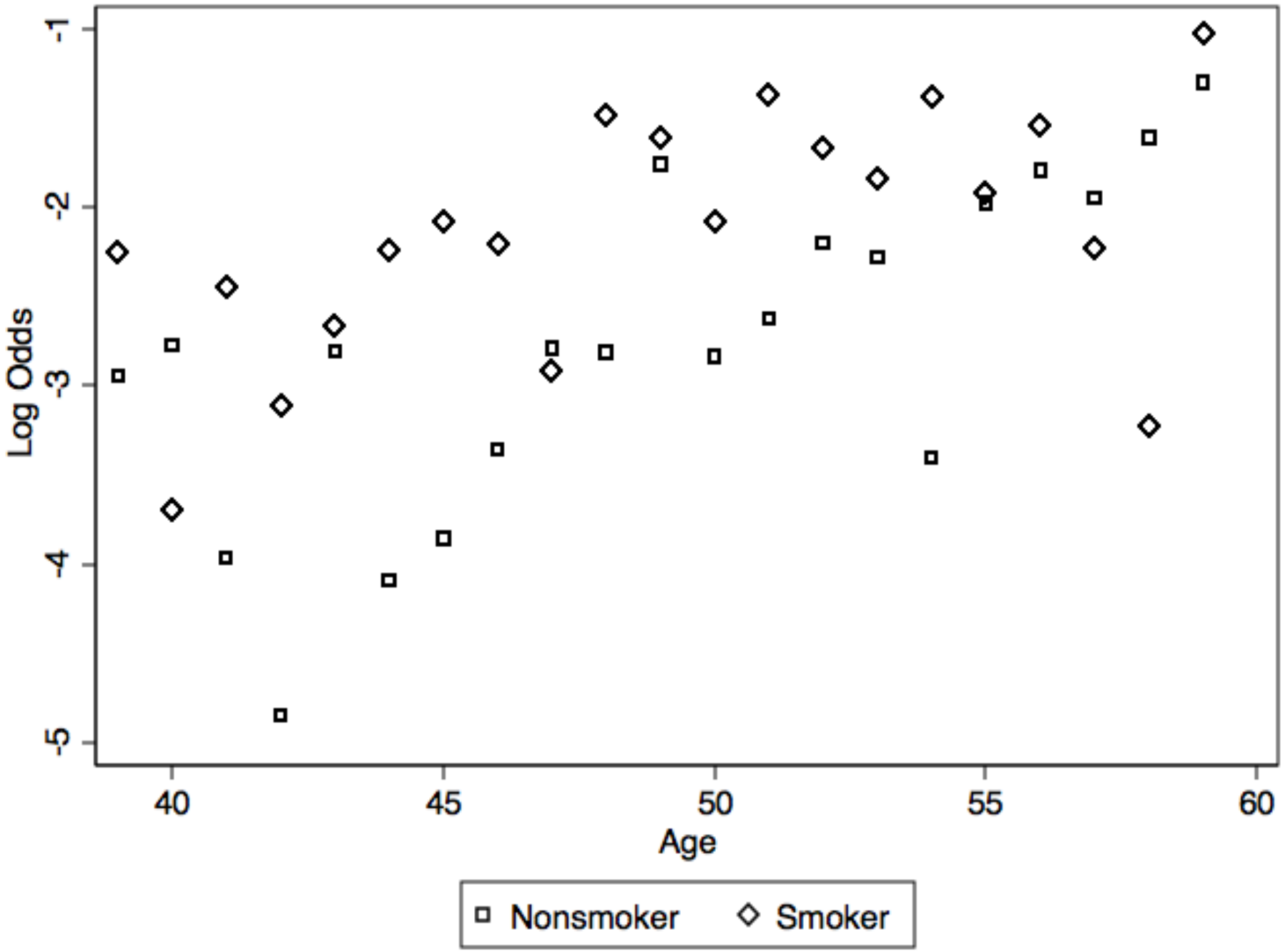
```
. logit chd smoke##age


Logistic regression                                     Number of obs   =       3154
                                                        LR chi2(41)     =     124.40
                                                        Prob > chi2     =     0.0000
Log likelihood = -828.42295                             Pseudo R2       =     0.0698

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.smoke |   .6931472   .4924238     1.41   0.159    -.2719857     1.65828
              |
         age |
          40  |   .1718503   .5181037     0.33   0.740    -.8436144    1.187315
          [the rows recording ages 41 to 58 not included here]
          59  |   1.645156   .6020774     2.73   0.006      .465106    2.825206
              |
    smoke#age |
        1 40  |  -1.617737   .8379347    -1.93   0.054    -3.260059    .0245851
          [the rows recording ages 41 to 58 not included here]
        1 59  |  -.4234836   .8520745    -0.50   0.619    -2.093519    1.246552
              |
       _cons |  -2.944439   .3877834    -7.59   0.000     -3.70448   -2.184398
------------------------------------------------------------------------------

. predict loh,xb
. twoway (scatter loh age if smoke==0,msymbol(sh) color(black) ytitle("Log Odds"))
(scatter loh age if smoke==1,msymbol(Dh) color(black)), legend(label(1 "Nonsmoker")
label(2 "Smoker")) scheme(s1mono)
```

Fitted Values (Observed Log Odds) Versus Age

Here we have used the elaborate logistic regression model to provide us with the observed log odds.

The role here is to try to determine the nature of the relationship between the log odds of CHD and age; separately for the smokers and the nonsmokers.

The first graph next gives some indications but it is not clear.


Log Odds
-1
-2
-3
-4
-5
40
45
50
55
60
Age
Nonsmoker
Smoker

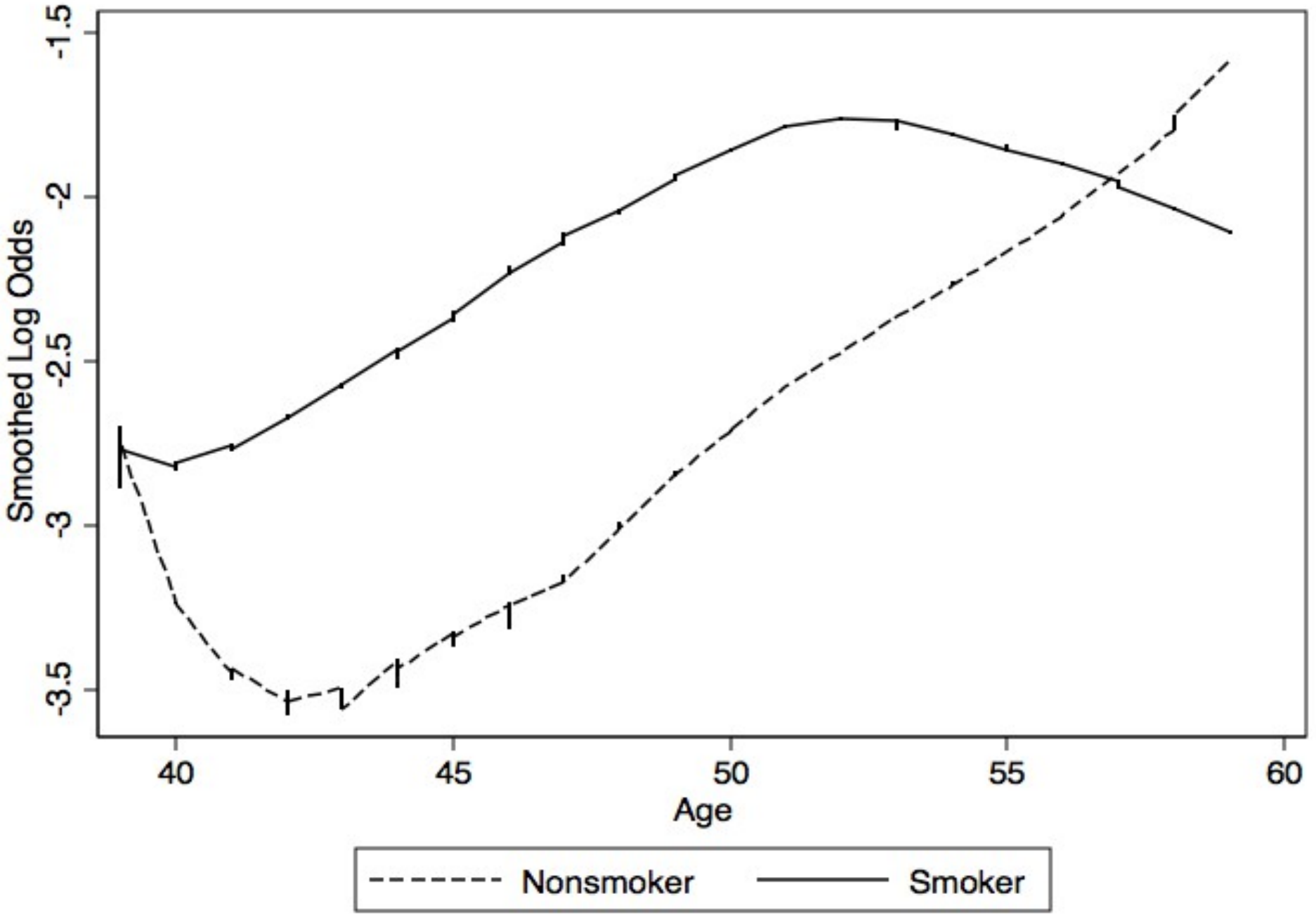
Smoothed Observed Log Odds

Smoothers can often provide the investigator with better cues from such graphs

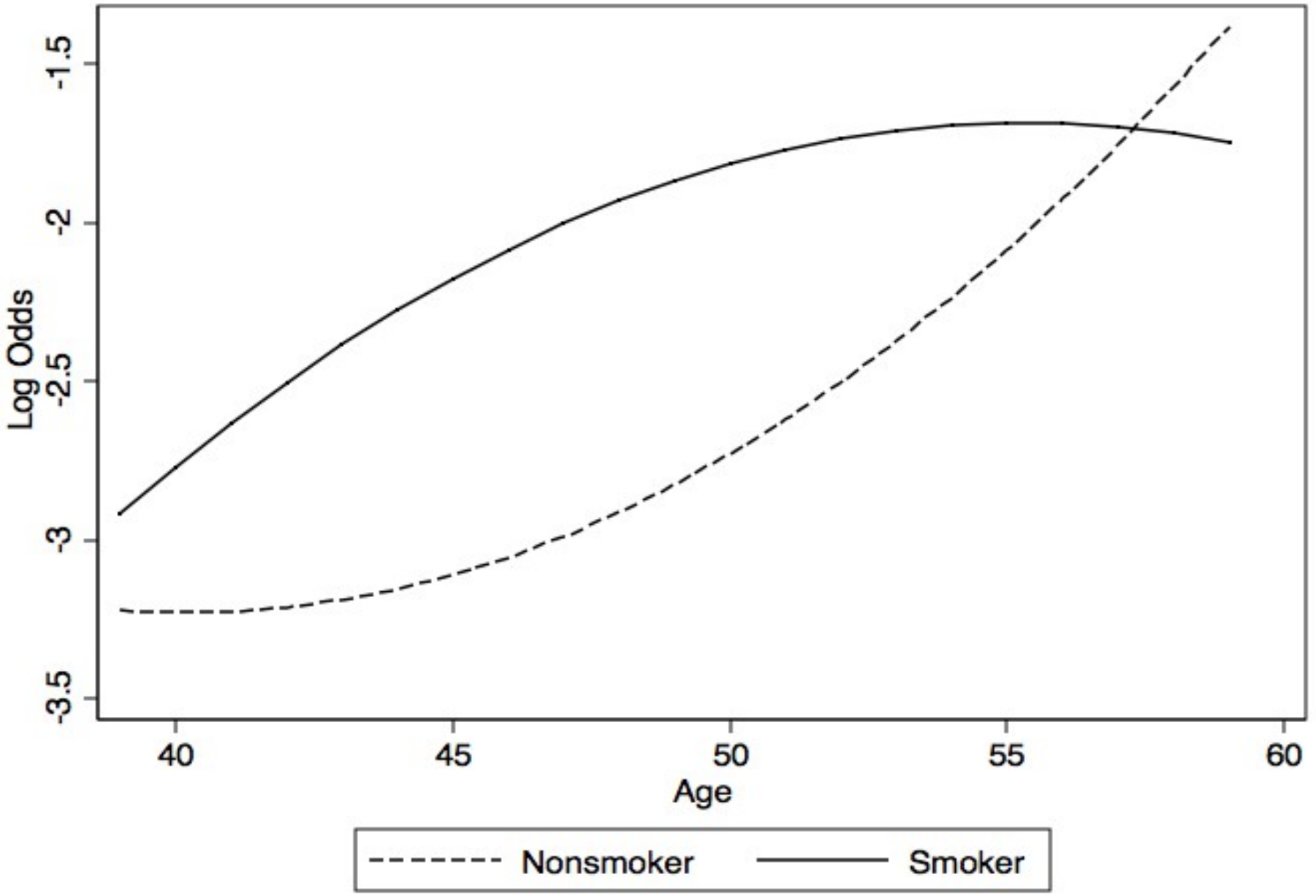
The most commonly seen smoother is called lowess

```
. twoway (lowess loh age if smoke==0,lpattern("-") color(black) ytitle("Smoothed
Log Odds")) (lowess loh age if smoke==1, color(black)), legend(label(1
"Nonsmoker") label(2 "Smoker")) scheme(s1mono)
```

These 2 curves suggest the consideration of parabolae.


Smoothed Log Odds
-1.5
-2
-2.5
-3
-3.5
40
45
50
55
60
Age
Nonsmoker
Smoker

```
. logit chd s a sa a2 sa2

Logistic regression                                     Number of obs   =       3154
                                                        LR chi2(5)      =      72.68
                                                        Prob > chi2     =     0.0000
Log likelihood = -854.28167                             Pseudo R2       =     0.0408

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |  -21.21013   10.07058    -2.11   0.035    -40.94809   -1.472157
           a |  -.4223795   .3125062    -1.35   0.177     -1.03488    .1901213
          sa |   .9384068   .4183366     2.24   0.025     .1184821    1.758332
          a2 |   .0052485   .0031931     1.64   0.100    -.0010099     .011507
         sa2 |   -.0099179   .0042943    -2.31   0.021    -.0183345   -.0015013
       _cons |   5.269134   7.544311     0.70   0.485    -9.517443    20.05571
------------------------------------------------------------------------------

. predict loh2,xb
. twoway (line loh2 age if smoke==0,lpattern("-") color(black) ytitle("Log Odds"))
(line loh2 age if smoke==1,color(black)), legend(label(1 "Nonsmoker") label(2
"Smoker")) scheme(s1mono)
```

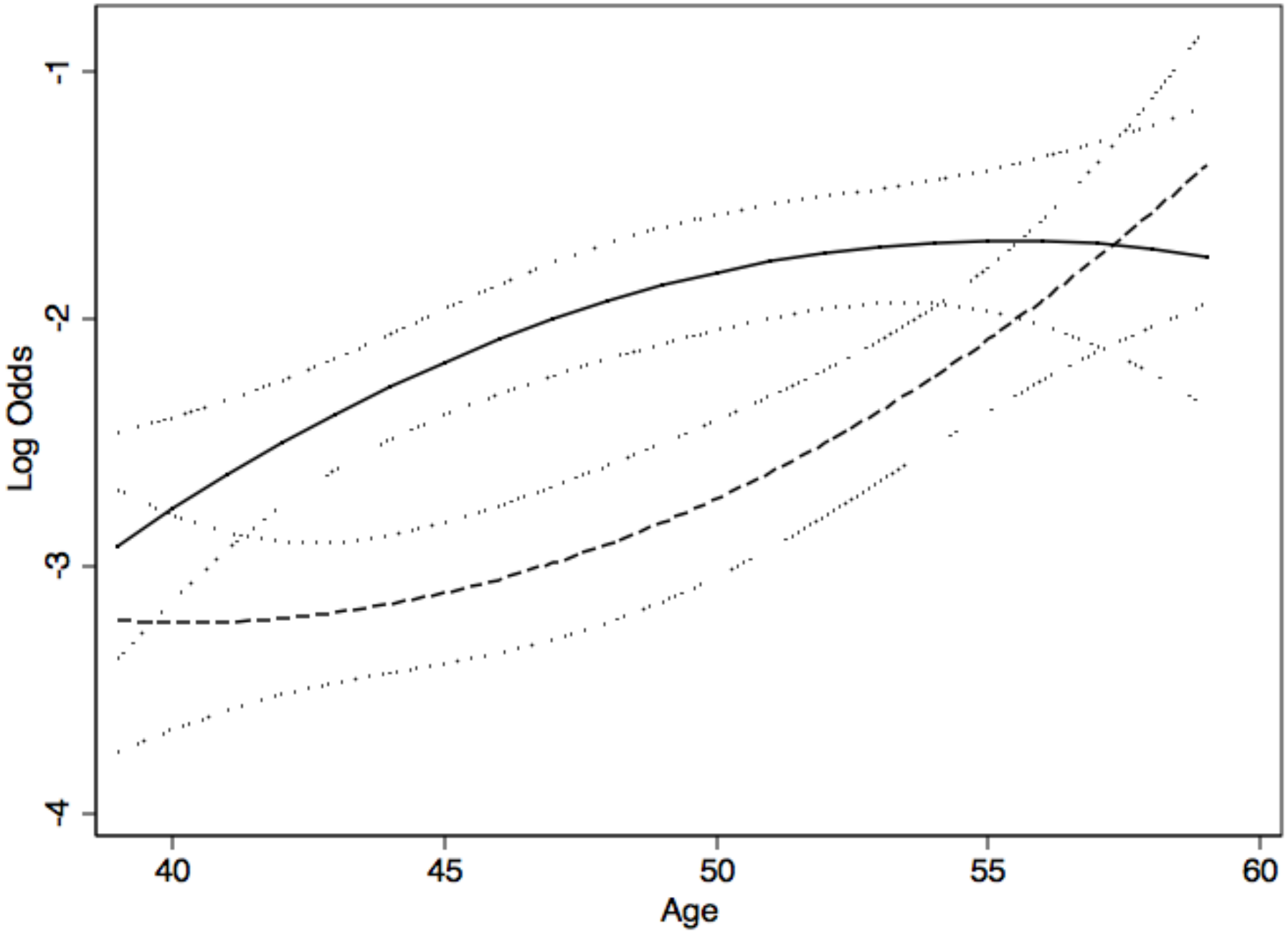

Log Odds
-1.5
-2
-2.5
-3
-3.5
40
45
50
55
60
Age
Nonsmoker
Smoker

Standard Errors for the Fitted Values

As always, the estimates alone are inadequate without standard errors or confidence intervals.

Further, intervals constructed for a log odds can be transformed to intervals for probabilities.

```
. predict seloh2,stdp

. gen cil=loh2-1.96*seloh2

. gen cih=loh2+1.96*seloh2

. twoway (line loh2 age if smoke==0,lpattern("-") color(black) ytitle("Log
Odds")) (line cil age if smoke==0,lpattern(".") color(black))(line cih age if
smoke==0,lpattern(".") color(black))(line loh2 age if smoke==1,color(black))
(line cil age if smoke==1,lpattern(".") color(black))(line cih age if smo
ke==1,lpattern(".") color(black)), legend(off) scheme(s1mono)
```


Log Odds
-1
-2
-3
-4
40
45
50
55
60
Age

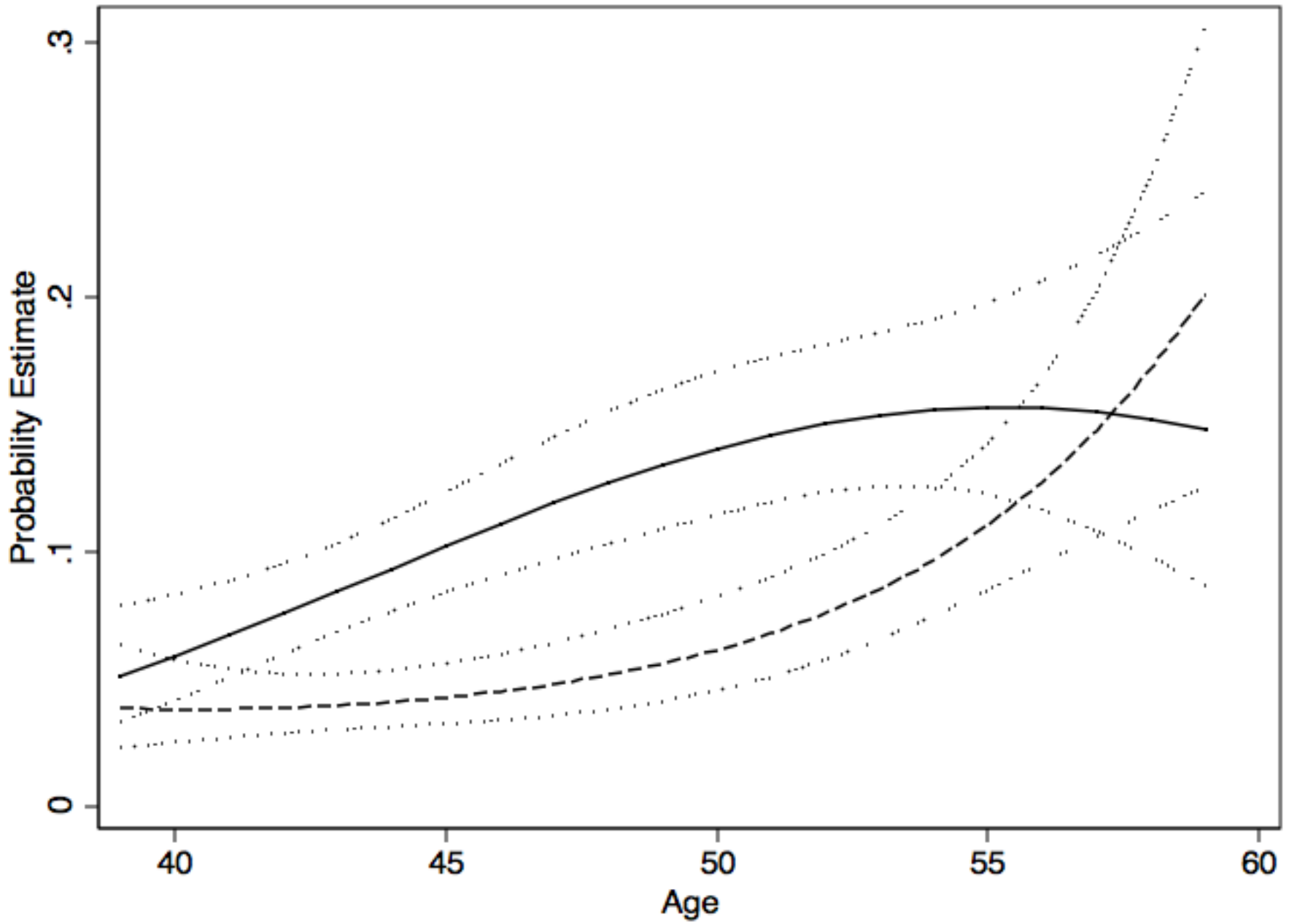
Getting the estimates of the probabilities

```
. gen cilp=1/(1+exp(-cil))

. gen cihp=1/(1+exp(-cih))

. gen pest=1/(1+exp(-loh2))

. twoway (line pest age if smoke==0,lpattern("-") color(black)
ytitle("Probability Estimate")) (line cilp age if smoke==0,lpattern(".")
color(black))(line cihp age if smoke==0,lpattern(".") color(black))(line pest
age if smoke==1,color(black))(line cilp age if smoke==1,lpattern(".")
color(black))(line cihp age if smoke==1,lpattern(".") color(black)), legend(off)
scheme(s1mono)
```

These estimates (fitted proportions) can give us cues to the issues at hand for specified sets of smoking status and age.

i.e. For a person of a given age and of given smoking status, the model provides an estimate of the probability of CHD.

The quality of such estimates and the narrowness of the intervals provide for alternate criteria to assess a model.


Probability Estimate
0
.1
.2
.3
40
45
50
55
60
Age

2) Models That Determine a Potentially Distinct Fitted Value for each Individual in the study

Each individual has a unique set of conditions

All different fitted values

The second situation is where there is one or more measured variables (like “actual” age rather than age group) so that each individual in the study has their own unique set of conditions. A participant's age could be computed from their data of birth compared with the date of entry into a study. There would typically be very few sets of individuals with exactly the same age (in days).

Residuals

Notice, here, that the data, here, cannot be separated into groups where, in each group, all conditions are the same. In a sense, there are no groups and hence no group observed proportions.

In a way, then, the observed proportions are all either 0 or 1 and every model will yield fitted values that are between 0 and 1: all potentially different for each infant. Every residual value is either or .

$\hat{p}$

$0 - \hat{p}$ $1 - \hat{p}$

Fitted Value Assessment

The investigator can view the 2 sets of fitted values:

- for the those with the outcome

- for those without the outcome

Unusual, outlying fitted values or odd clusters of fitted values can give the investigator clues to trouble

Generalized Additive Models

"Generalized Additive Modeling" [Hastie and Tibshirani (1990)] has shown considerable promise in the assessment of models: particularly when the functional form of one or more measured independent variables is under question and may not be a series of lines. In the context of logistic regression, a generalized additive model [GAM] looks like:

$$\log(p/(1-p)) = \sum_i \beta_i x_i + \sum_j s(y_j)$$

: where the s functions are not specified but are constructed using “smoothers” as part of the model algorithm.

The graph of these “smooth” curves can be assessed.

There are tests of significance associated with nonlinearity in this setting as well.

Several software systems [including Splus and R] have implemented GAM. Alas, Stata has not [yet] taken on this implementation.

The purpose(s) for a model

1 : Attempts at Etiology ? -

Understanding the disease-exposure relationship

Identifying modifiers, confounders and other important explanatory variables

2 : Prediction ? Forecast ? The future ? -

Trying to predict a person's outcome based on their explanatory variables

From estimates to prediction

Logistic regression gives estimates of log odds and estimates of probabilities.

How do we get predictions from the estimates?

We establish a threshold.

If a probability estimate is above the threshold, we 'predict' the presence of the outcome [disease, CHD, ...]

If a probability estimate is below the threshold, we 'predict' the absence of the outcome [no disease, no CHD, ...]

Actual outcome compared with prediction

From the data used to build the model and construct the fit, we now have 2 sets of probability estimates: a set of estimates for those with CHD and a set of estimates for those without CHD.

We would hope for high estimates for those with CHD and low estimates for those without CHD.

The prevalence of CHD in this study population provides a guidepost for such deliberations.

Prevalence of CHD

```
. ci chd

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         chd |       3154    .0814838    .0048721         .071931    .0910367
```

Without knowledge of age and smoking status, one could estimate the prevalence of CHD, from the data at hand, to be 0.0814838

From the model, estimates above 0.0814838 could suggest CHD while estimates below 0.0814838 could suggest No CHD.

+ if $\hat{p}$ >0.0814838 and – if $\hat{p}$ < 0.0814838

If we were to classify the study participants in this way and compare with their actual CHD status, we would get:

```
. estat class,cutoff(0.0814838)

Logistic model for chd

              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |       163           1158  |       1321
     -     |        94           1739  |       1833
-----------+--------------------------+-----------
   Total   |       257           2897  |       3154

Classified + if predicted Pr(D) >= .0814838
True D defined as chd != 0
---------------------------------------------------
Sensitivity                     Pr( +| D)   63.42%
Specificity                     Pr( -|~D)   60.03%
Positive predictive value       Pr( D| +)   12.34%
Negative predictive value       Pr(~D| -)   94.87%
---------------------------------------------------
False + rate for true ~D        Pr( +|~D)   39.97%
False - rate for true D         Pr( -| D)   36.58%
False + rate for classified +   Pr(~D| +)   87.66%
False - rate for classified -   Pr( D| -)    5.13%
---------------------------------------------------
Correctly classified                        60.30%
---------------------------------------------------
```

Cutoff or Threshold

Here the cutoff (sometimes called the threshold) determines the classification rule. These classifications can be called predictions.

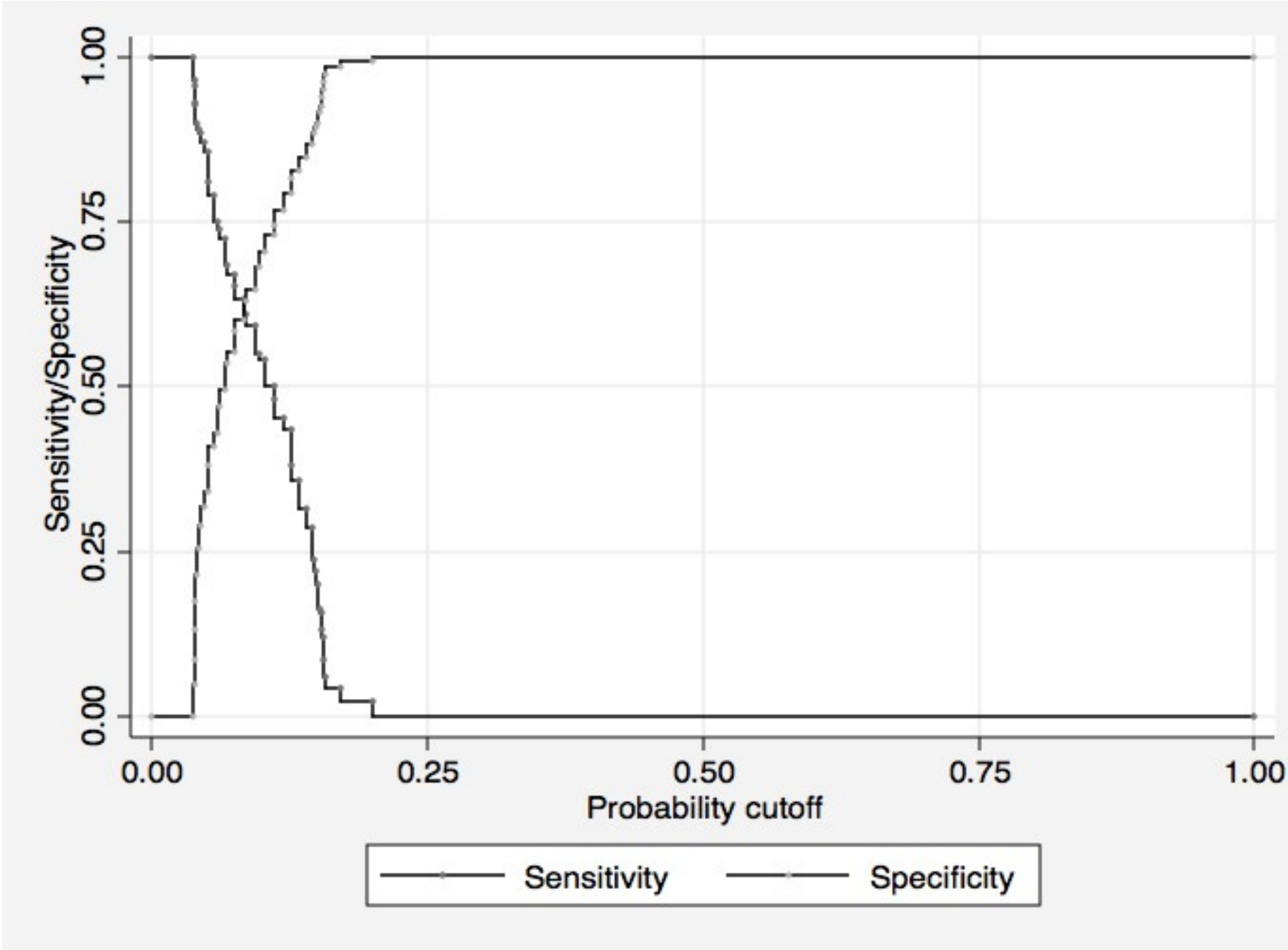
Based on this cutoff, we obtain estimates of sensitivity and specificity:

```
. cii 257 163

                                                          -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        257    .6342412     .030044        .5721294    .6932173

. cii 2897 1739

                                                          -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |       2897    .6002761    .0091008        .5821723    .6181772

.
```

Sensitivity and Specificity as functions of the Cutoff

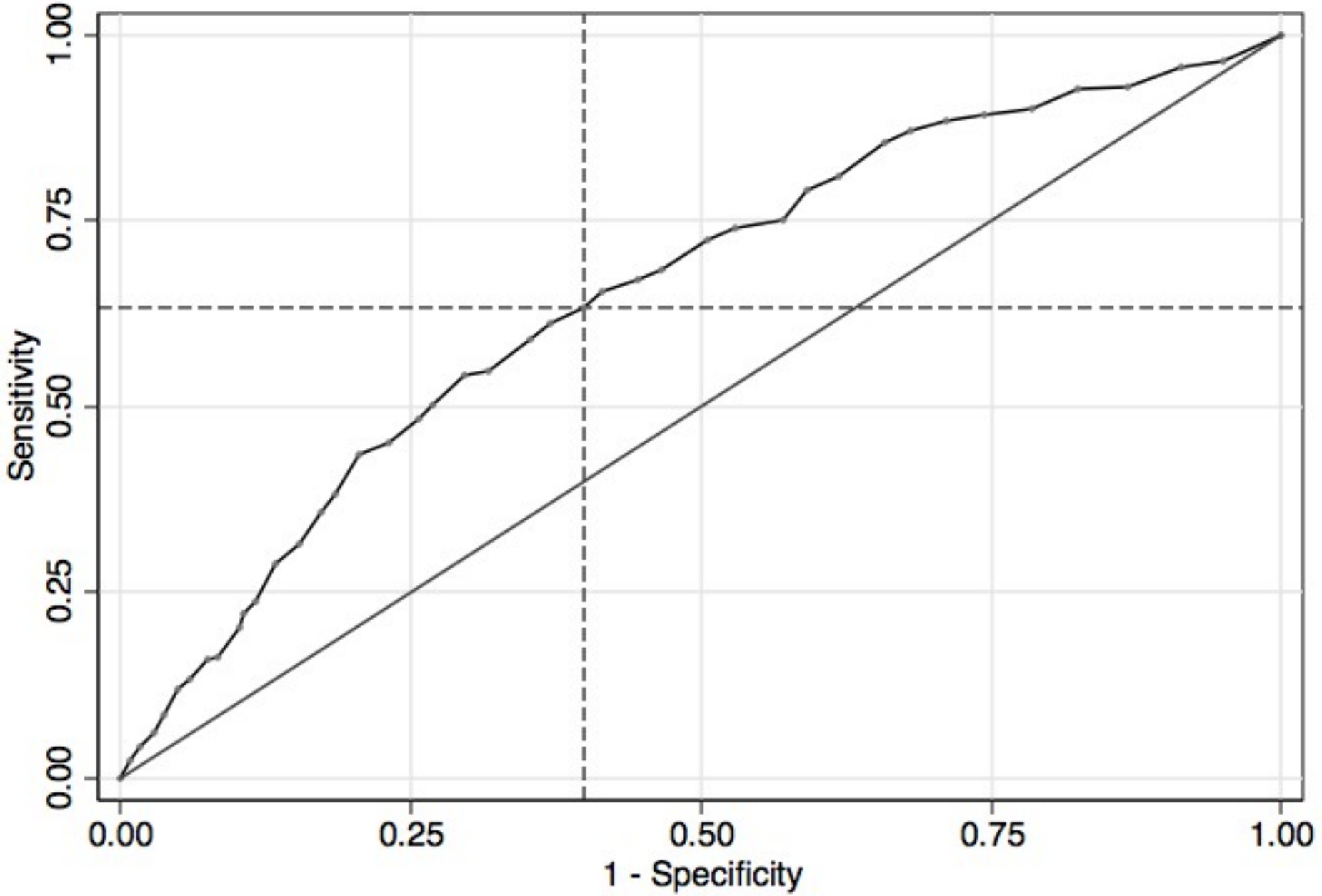
Alternatively, one can think of the sensitivity and specificity as functions of the probability cutoff. One can graph sensitivity and specificity estimates versus the cutoff. When the cutoff is zero, the sensitivity estimate is 1 and the specificity estimate is 0. If the cutoff is one, the sensitivity estimate is 0 and the specificity estimate is 1. As the cutoff rises, the sensitivity estimate declines and the specificity estimate rises. The graphs are “step” functions with a step for every distinct fitted value. Steps down for sensitivity and steps up for specificity.

```
. lsens,connect(stairstep stairstep) msize(tiny tiny) scheme(s2mono)
```


Sensitivity/Specificity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
Probability cutoff
Sensitivity
Specificity

Or one can think in terms of the false positive rate = 1 – specificity. When the cutoff is zero, the sensitivity is 1 and the false positive rate is 1. If the cutoff is one, the sensitivity is 0 and false positive rate is 0. A graph of the sensitivity estimates versus the false positive rate estimates can be useful for assessment of a logistic regression model. The more the plotted values are in the 'upper left ' corner of the display, the better. The “curve” obtained by joining these points based on the ordered fitted values is widely determined. This curve is called a receiver operating characteristic curve (ROC curve).

```
. lroc,msize(tiny) xline(0.3997,lpattern("-")) yline(0.6342,lpattern("-"))
scheme(s1mono)
```


Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity
Area under ROC curve = 0.6546

ROC, Sensitivity and Specificity

The graph shows the ROC for our candidate model.

The dotted lines cross at the point on the curve corresponding to the cutoff of 0.0814838 as seen earlier.

Interpreting the ROC curve

Parts of this curve corresponding to very low sensitivity indicate cutoffs of no use. Similarly, the parts of the curve corresponding to very low specificity indicate cutoffs of no use as well.

One would presumably view the central portion of this curve with some credibility. The cutoff discussed earlier surely falls in this range.

Nevertheless, some investigations focus on the area under the (entire) curve.

The area under the curve (AUC)

aka the c-statistic (history of this label?)

AUC: Standard Errors & Confidence Intervals

AUC estimates are formed from sensitivity (Sn) estimates and specificity (Sp) estimates.

Accordingly, the precision of AUC estimates is materially dependent on the denominators of the Sn and Sp estimates.

```
. quietly: logit chd s a sa a2 sa2

. predict phat2, p

. roctab chd phat2

                    ROC                    -Asymptotic Normal--
         Obs       Area     Std. Err.      [95% Conf. Interval]
       -------------------------------------------------------
        3154     0.6546       0.0178        0.61971     0.68946
```

Comparing AUC

```
. quietly: logit chd smoke##age

. predict phat1,p

. roccomp chd phat1 phat2
```

	Obs	ROC Area	Std. Err.	-Asymptotic Normal- [95% Conf.	Interval]
phat1	3154	0.7014	0.0161	0.66991	0.73286
phat2	3154	0.6546	0.0178	0.61971	0.68946

```
Ho: area(phat1) = area(phat2)
    chi2(1) =    14.85       Prob>chi2 =   0.0001
```

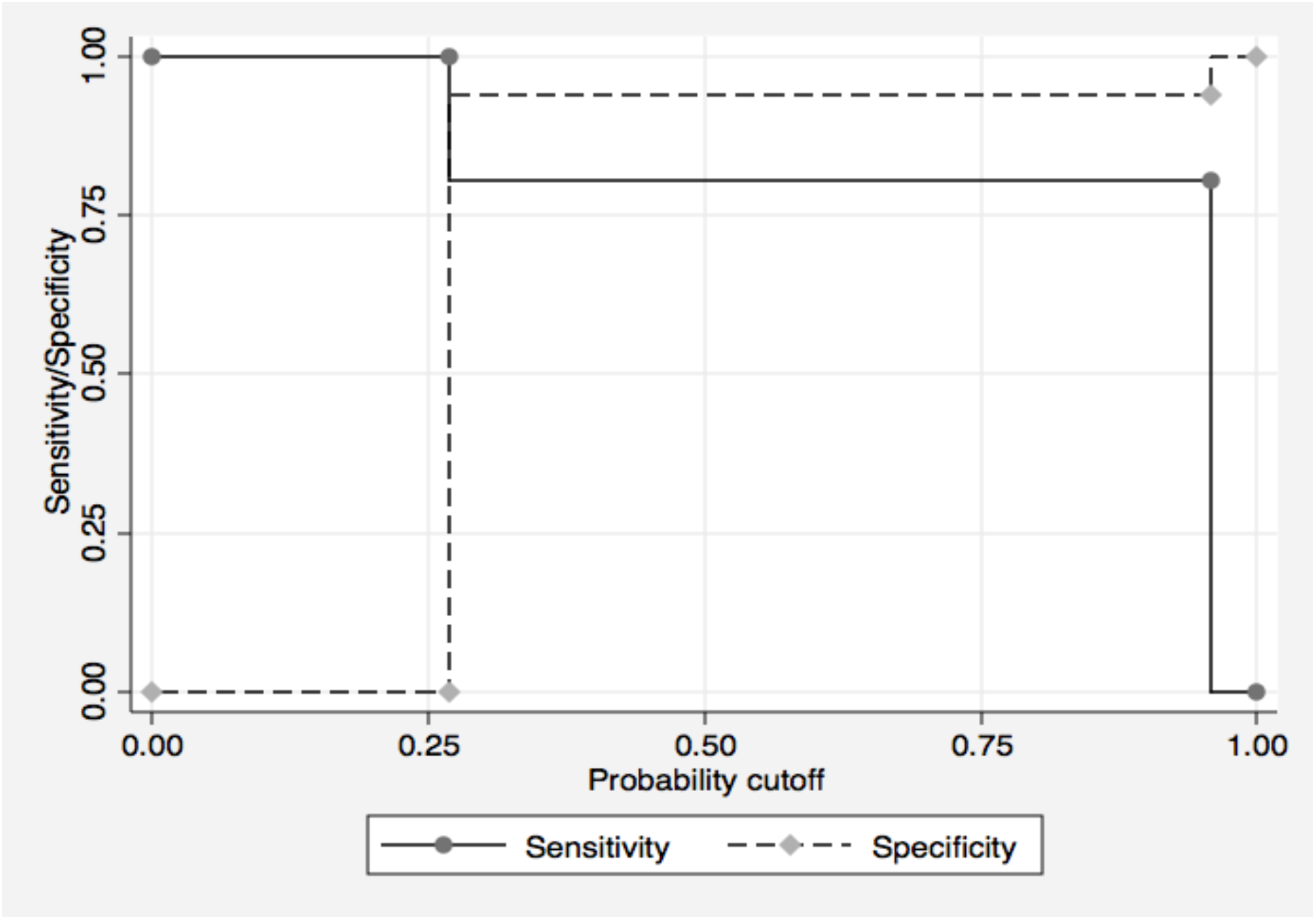
Inside the workings of AUC

Lets take the example from Rabe-Hesketh: Diagnosis of Heart Attacks on the use of serum creatine kinase (CK) levels for the diagnosis of myocardial infarction (MI:heart attack).

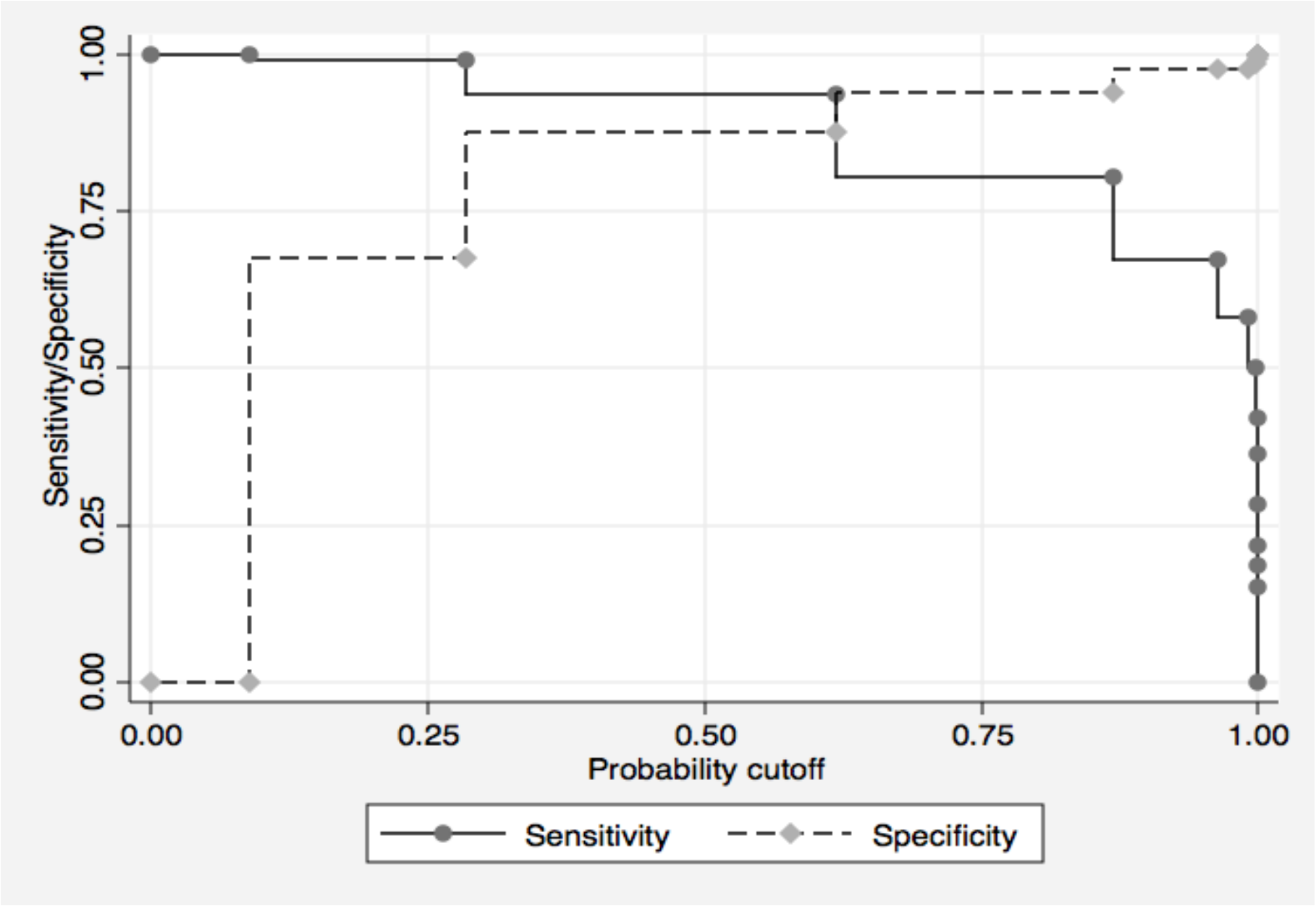
As a start, let us suppose that we wish to assess “a CK of more than 100” as a discriminator.

```
. table ck100, c(mean infct)

----------------------
    ck100 | mean(infct)
----------+-----------
        0 |    .2694611
        1 |    .9585492
----------------------
. quietly: logit infct ck100
. predict phat1,p
. tab phat1

  Pr(infct) |      Freq.     Percent        Cum.
------------+-----------------------------------
    .269461 |        167       46.39       46.39
   .9585493 |        193       53.61      100.00
------------+-----------------------------------
      Total |        360      100.00
```

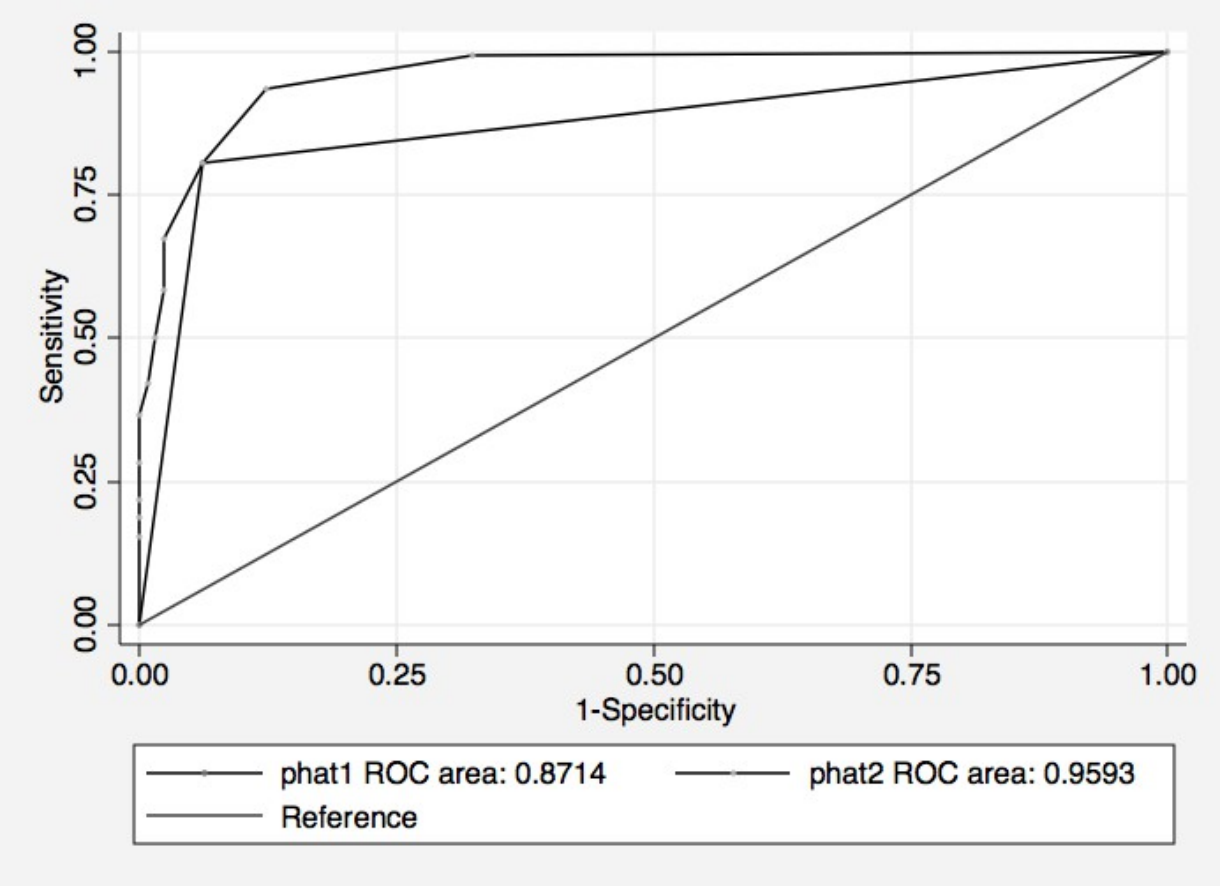
```
. estat class

Logistic model for infct

              -------- True --------
Classified |         D              ~D  |      Total
-----------+--------------------------+-----------
     +     |       185               8  |        193
     -     |        45             122  |        167
-----------+--------------------------+-----------
   Total   |       230             130  |        360

Classified + if predicted Pr(D) >= .5
True D defined as infct != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   80.43%
Specificity                     Pr( -|~D)   93.85%
Positive predictive value       Pr( D| +)   95.85%
Negative predictive value       Pr(~D| -)   73.05%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    6.15%
False - rate for true D         Pr( -| D)   19.57%
False + rate for classified +   Pr(~D| +)    4.15%
False - rate for classified -   Pr( D| -)   26.95%
--------------------------------------------------
Correctly classified                        85.28%
--------------------------------------------------
```


Sensitivity/Specificity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
Probability cutoff
Sensitivity
Specificity

```
. quietly:logit infct ck
. predict phat2,p
. lsens,connect(stairstep stairstep) scheme(s2mono) lpattern( solid dash)
```



```
. roccomp infct phat1 phat2,summary graph plot1opts(msize(tiny))
plot2opts(msize(tiny)) scheme(s2mono)

                                ROC                    -Asymptotic Normal--
                    Obs       Area     Std. Err.      [95% Conf. Interval]
-------------------------------------------------------------------------
phat1               360     0.8714       0.0168        0.83839     0.90442
phat2               360     0.9593       0.0099        0.93991     0.97862
-------------------------------------------------------------------------
Ho: area(phat1) = area(phat2)
    chi2(1) =    59.72       Prob>chi2 =   0.0000
```



Arbitration based on AUC

- The AUC estimate with actual CK levels is 0.9593 suggesting that the actual CK levels (modeled with a linearity assumption) provide “better prediction” than a simple threshold of CK>100.
- This is a very simple example of a comparison of 2 models where a Likelihood Ratio test is not directly available. [One model is not nested within the other model]
- The AUC estimate is not based on the models per se... just their fitted values

AUC as a Probability

You may have noticed that the AUC is a number between 0 and 1. It does, in fact, have an interpretation as an estimate of a probability. Suppose a case and a control are each randomly selected. If the classification is based a rule that classifies the one with the higher fitted value and hence the higher CK value as the case, the one with the lower fitted value and hence the lower CK value as the control, then such a rule will correctly classify with a fixed frequency estimated by the AUC. [This property can be easily verified for the simple 2x2 table. (For this interpretation, the classification rule requires "guessing" if the CK values are the same)]

Recent commentary

NR Cook (2008) 'Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve' Clinical Chemistry 54: 17-23, 2008.

http://www.clinchem.org/cgi/content/full/54/1/17

Cross Validation

We are using our data twice here.

We are using the data to build a model and then using the same data to assess the model.

This process is called internal cross validation.

Really, we should have one data set for the model building and then another data set for the model assessment.

If done with 2 datasets, we call the process external cross validation.

Goodness?

Goodness of Fit (GOF) tests simply will not go away.

They are notoriously low power except with huge datasets.

Criticisms of GOF tests abound.

If a journal requires such a test be included, be aware that the result of this test has little consequence.