Logistic Regression
Model assessment & Prediction

March 10, 2015

Biostatistics 208

Model assessment techniques for logistic regression

Logistic regression involves a number of implicit assumptions that need to be evaluated for valid inferences. These include:

- the logistic model is the correct representation of the relationship between the outcome probability and the predictors
- the observed outcomes are independent
- the outcome and predictors are measured without error
- no important predictors are omitted

If any of these assumptions fail to be met, biased estimates and inferences can result. Data analyses are not complete until these issues are evaluated.

Even in situations where the form of the model is correct, it may fail to fit the observed data adequately for a number of reasons

- failure to control for important interactions
- failure to account for nonlinearity
- excessive collinearity between predictors

Linearity assessment

Covered in lab and previous lectures. Issues and techniques for logistic models are similar as those used for linear models (except that component + residual plots aren't as well-developed / easily interpretable as the linear regression counterparts).

Example: using splines to check linearity of age in an adjusted model from the WCGS:

```
. logistic chd69 agesp* chol sbp bmi smoke dibpat
Logistic regression                               Number of obs   =       3142
                                                  LR chi2(9)      =     196.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -791.44399                       Pseudo R2       =     0.1103
------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agesp1 |   .7352456   .1094136    -2.07   0.039     .5492412    .9842416
      agesp2 |   36.58274   60.52281     2.18   0.030     1.429028    936.5085
      agesp3 |   .0002443   .0010267    -1.98   0.048     6.46e-08    .9242007
      agesp4 |   216.1135   717.6407     1.62   0.105     .3222004    144956.5
        chol |   1.010755   .0015186     7.12   0.000     1.007783    1.013736
         sbp |   1.018255   .0042098     4.38   0.000     1.010037     1.02654
         bmi |   1.057025   .0280302     2.09   0.036      1.00349    1.113416
       smoke |   1.818675    .256924     4.23   0.000     1.378814    2.398856
      dibpat |   2.001371   .2892674     4.80   0.000     1.507647    2.656781
       _cons |   12.03652   72.33961     0.41   0.679     .0000922     1571206
------------------------------------------------------------------------------

. testparm agesp2-agesp4

 ( 1)  [chd69]agesp2 = 0
 ( 2)  [chd69]agesp3 = 0
 ( 3)  [chd69]agesp4 = 0

           chi2(  3) =    7.09
         Prob > chi2 =    0.0691
```

Collinearity

- Presence of highly collinear predictors can lead to estimated regression coefficients that are unreliable and imprecise

- Solution:

  - avoid entering collinear predictors in the same model, especially if these involve a predictor of primary interest

  - use alternative measures which reduce collinearity (e.g. "centering" of predictors)

Collinearity

Example: the extreme case of entering the same predictor (BMI and its centered version) twice

```
. logistic chd69 bmi bmicen
note: bmicen omitted because of collinearity

Logistic regression                                     Number of obs   =       3141
                                                        LR chi2(1)      =      11.21
                                                        Prob > chi2     =     0.0008
Log likelihood = -881.48822                             Pseudo R2       =     0.0063
------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   1.085959   .026313     3.40   0.001     1.035592    1.138776
      bmicen |  (omitted)
------------------------------------------------------------------------------
```

- Stata will omit predictors that are perfectly collinear

Collinearity

Example: WCGS including interaction between age and arcus

```
. logistic chd69 i.arcus##c.age

Logistic regression                               Number of obs   =       3152
                                                  LR chi2(3)      =      53.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -858.93362                       Pseudo R2       =     0.0301

------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.arcus |   15.70823   17.90923     2.42   0.016     1.681347    146.7564
         age |   1.093788    .016287     6.02   0.000     1.062328     1.12618
             |
 arcus#c.age |
           1 |   .9513913   .0222084    -2.13   0.033     .9088444    .9959301
             |
       _cons |   .0011271   .0008093    -9.45   0.000     .0002759     .004604
------------------------------------------------------------------------------
```

- Note large OR for `arcus` with large standard error
- Coefficient for `arcus` is the effect of `arcus` among those with `age`=0
- *Not a problem* if the interaction is interpreted for a sensible age value

Collinearity

Diagnosing collinearity for logistic models:

- Looking at the correlation matrix of predictors is a start
  - pairwise correlations ≥ 0.8 a concern
- Recall from lecture 7 that multi-collinearity is reflected in *variance inflation factors* (*VIF*).
  - `"estat vif"` command works only for linear models (`regress`)
  - Can use `regress` to provide a crude assessment:

```
. * pre-centered results
quietly regress chd69 arcus i.arcus##c.age
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
       arcus |     71.04    0.014076
         age |      1.58    0.634016
 arcus#c.age |
           1 |     73.85    0.013542
-------------+----------------------
    Mean VIF |     48.82
```

- watch for *VIF*>5

Collinearity

Example: Above model, now including interaction between *centered* age and arcus

```
. * refit with centered age
. logistic chd69 i.arcus##c.agecen
Logistic regression                               Number of obs   =       3152
                                                  LR chi2(3)      =      53.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -858.93362                       Pseudo R2       =     0.0301
------------------------------------------------------------------------------
        chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
      1.arcus |   1.565375    .226971     3.09   0.002     1.178145    2.079879
       agecen |   1.093788    .016287     6.02   0.000     1.062328     1.12618
              |
arcus#c.agecen |
            1 |   .9513913   .0222084    -2.13   0.033     .9088444    .9959301
              |
        _cons |   .0714084   .0062745   -30.04   0.000     .0601113    .0848286
------------------------------------------------------------------------------
```

- As before, results indicate an interaction between arcus and age. However, precision is improved for estimating arcus effect, and now the model coefficients are more interpretable.
- Note: highly collinear terms should be avoided when the primary predictor is involved - less of an issue for adjustment variables

Regression diagnostics for logistic models

- Residuals and influence statistics for logistic models can be used as in linear regression to identify:
  - outliers
  - observations with high leverage / influence on model fit
- The following fit statistics are provided by as options to the predict command:
  - residuals, defined for outcome y and predicted outcome probability P as:

$$\frac{y - P}{\sqrt{P \times (1 - P)}}$$

  - "dfbeta" influence statistics (calculated for each observation; interpreted as in linear models)
  - leverage statistics

Regression diagnostics for logistic models

Example: model for the relationship between behavior pattern and other predictors (WCGS)

```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp
Logistic regression                               Number of obs   =       3142
                                                  LR chi2(8)      =     200.57
                                                  Prob > chi2     =     0.0000
Log likelihood = -789.31393                       Pseudo R2       =     0.1127
------------------------------------------------------------------------------
        chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   1.806691   .2168309     4.93   0.000     1.427994    2.285816
     chol_50 |   1.812835   .1389486     7.76   0.000      1.55997    2.106689
      sbp_50 |   2.743347   .5665264     4.89   0.000     1.830206    4.112082
      bmi_10 |   2.746212   .8264588     3.36   0.001     1.522542    4.953349
       smoke |   1.812646   .2549975     4.23   0.000     1.375843    2.388127
      dibpat |   2.058848   .2982595     4.98   0.000     1.549934    2.734863
     bmichol |   .5005402   .1220925    -2.84   0.005      .310322    .8073565
      bmisbp |   .2476601   .1555466    -2.22   0.026      .072318    .8481365
       _cons |   .0329928   .0049553   -22.71   0.000     .0245795    .0442858
------------------------------------------------------------------------------
. predict p                          predicted probabilities (xb option gives log odds)
(option pr assumed; Pr(chd69))
(12 missing values generated)
. predict r, residuals               residuals
(12 missing values generated)
. predict h, hat                     leverage statistics
(12 missing values generated)
. predict db, dbeta                  coefficient influence statistics (DFbetas)
(12 missing values generated)
```

Regression diagnostics for logistic models

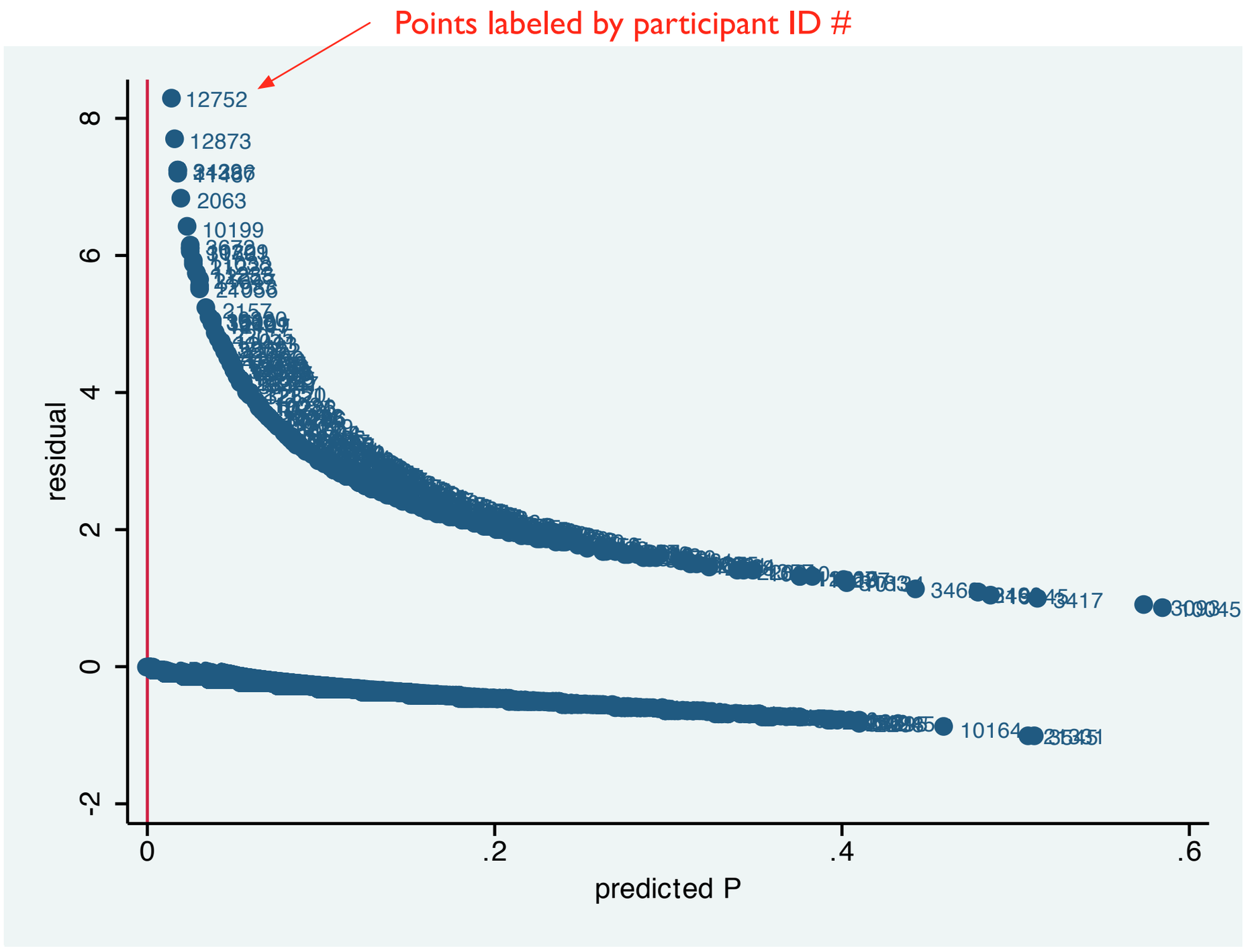
Residuals from first 10 participants:

```
. list id chd69 age arcus dibpat chol r if _n<=10
```

	id	chd69	age	arcus	dibpat	chol	r
1.	2001	no	49	absent	A1,A2	225	-.2166576
2.	2002	no	42	present	A1,A2	177	-.27567
3.	2003	no	42	absent	B3,B4	181	-.0919408
4.	2004	no	41	absent	B3,B4	132	-.0958496
5.	2005	yes	59	present	B3,B4	255	2.084351
6.	2006	no	44	absent	B3,B4	182	-.1901838
7.	2007	no	44	absent	B3,B4	155	-.0900676
8.	2008	no	40	absent	A1,A2	140	-.100299
9.	2009	no	43	absent	B3,B4	149	-.1843805
10.	2010	no	42	present	A1,A2	325	-.4215957

Regression diagnostics for logistic models

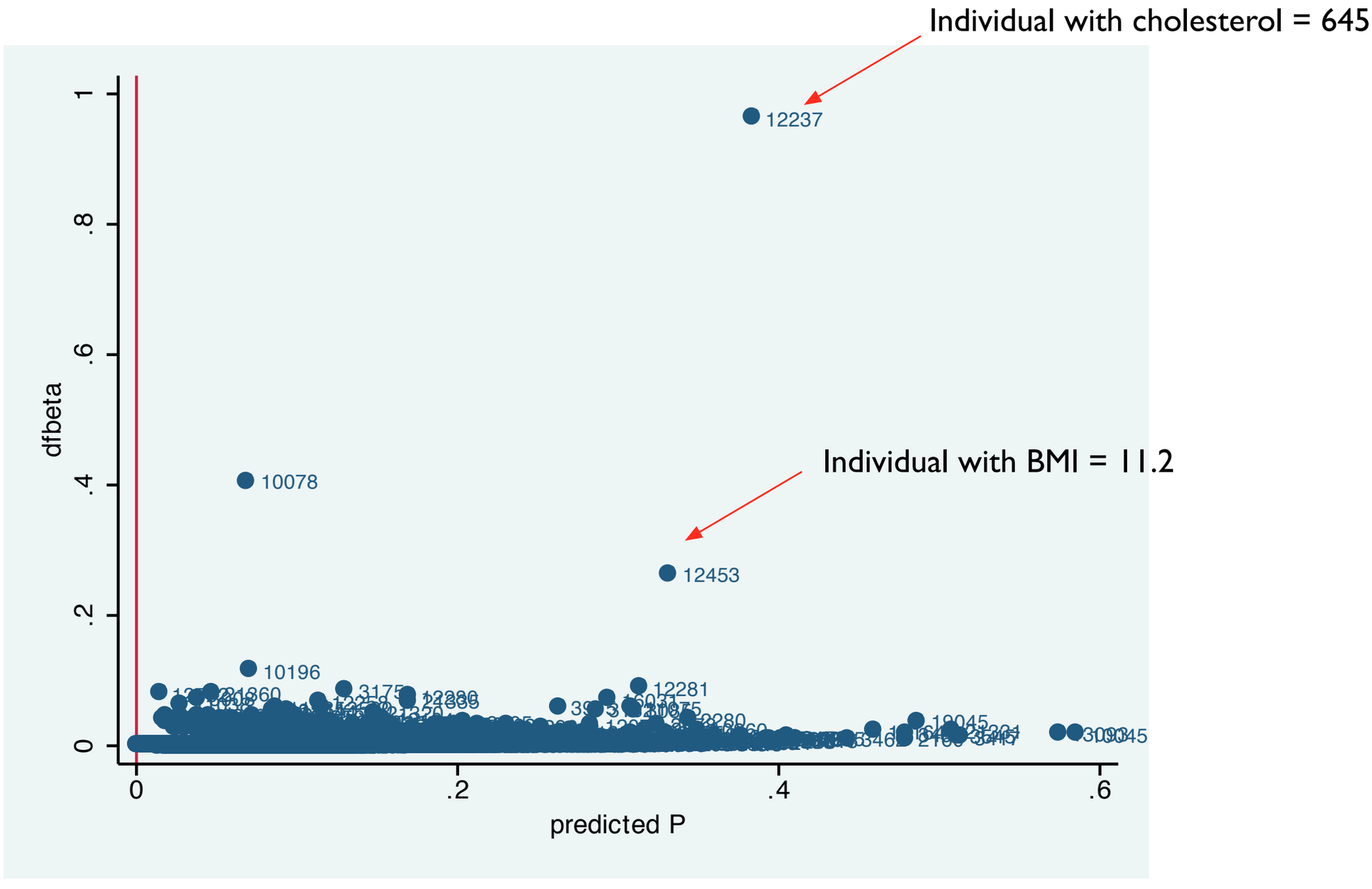
Residuals versus fitted values (probability scale):



Residuals are harder to interpret than with linear models

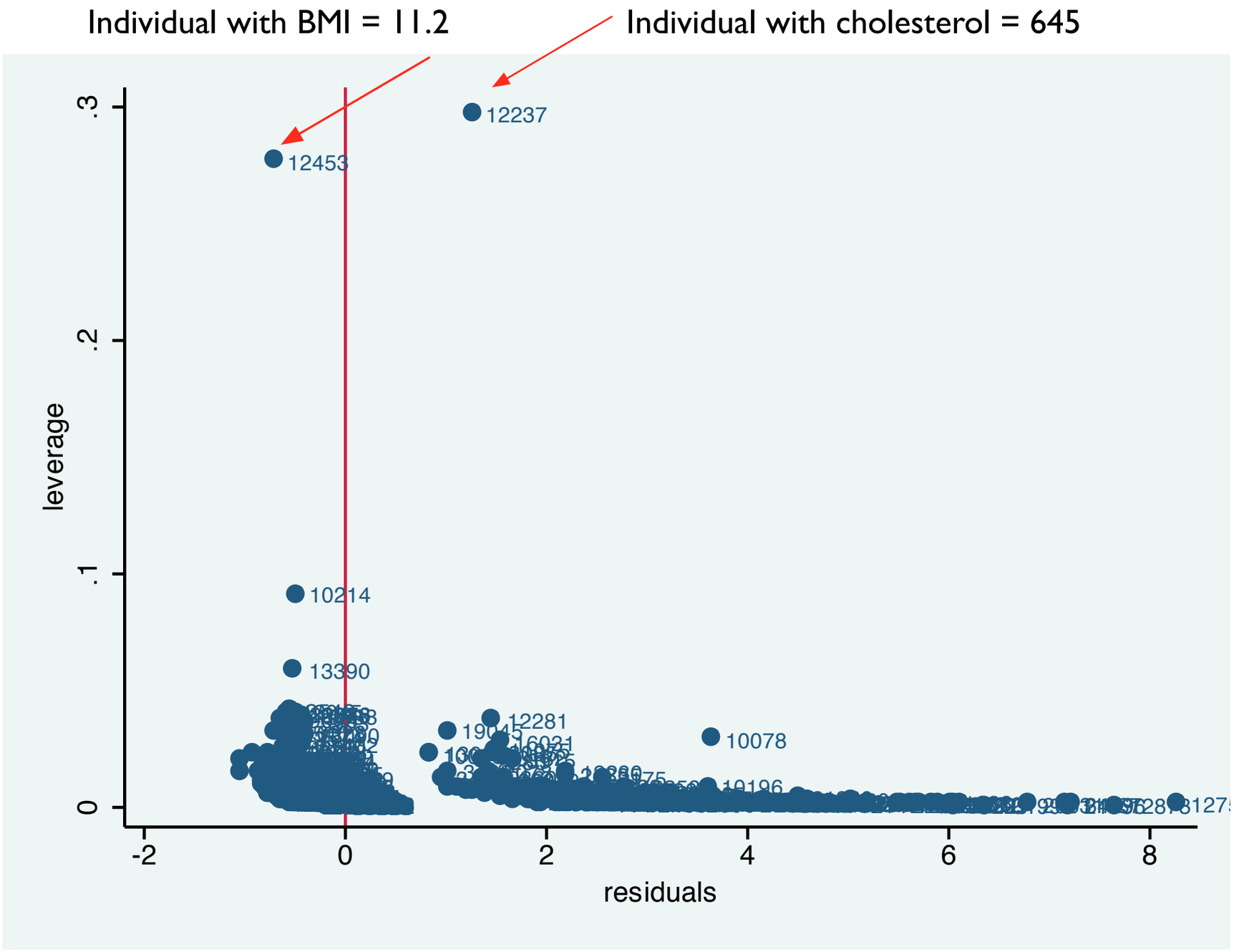
Regression diagnostics for logistic models

dfbeta's versus fitted values (probability scale):



Regression diagnostics for logistic models

leverage versus fitted values (probability scale):



Regression diagnostics for logistic models

List of individuals with highest influence/leverage:

```
. list id chd69 age chol sbp bmi smoke dibpat if id==10078 | id==12237 |id==12453 |
id==12752 | id==12873
      +-----------------------------------------------------------------+
      |    id   chd69   age   chol   sbp         bmi        smoke   dibpat |
      |-----------------------------------------------------------------|
1094. | 10078     yes    43    188   166    38.94737   nonsmoker    B3,B4 |
1828. | 12237     yes    43    645   154    30.12857   nonsmoker    A1,A2 |
1968. | 12453      no    47    188   196    11.19061      smoker    A1,A2 |
2158. | 12752     yes    41    184   134    19.66339      smoker    B3,B4 |
2182. | 12873     yes    48    155   134     24.3898   nonsmoker    B3,B4
      +-----------------------------------------------------------------+
```

implausibly large?

implausibly small?

Comparison of estimated coefficients from model excluding five high leverage/influence observations:

```
             |      Coef.     Coef.
   predictor |    all obs   excluding five obs.
+----------------------------------------------
     age_10 |   .5914968   .6098452
    chol_50 |   .5948921   .5952878
     sbp_50 |   1.009179   1.054829
     bmi_10 |   1.010222   1.075941
      smoke |   .5947879   .6166823
     dibpat |   .7221468   .7668107
    bmichol |  -.6920674  -.8431974
     bmisbp |  -1.395698  -2.181028
      _cons |  -3.411467  -3.464393
+----------------------------------------------
```

- Decision about whether or not to retain influential observations depends on validity and degree of effect on conclusions

Regression diagnostics for logistic models

Fitted model excluding two individuals with outlying values of cholesterol & BMI:

```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp if (id!=12237 &
id!=12453)

Logistic regression                                     Number of obs   =       3140
                                                        LR chi2(8)      =     198.96
                                                        Prob > chi2     =     0.0000
Log likelihood = -787.53196                             Pseudo R2       =     0.1121

------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   1.807172   .2171423     4.92   0.000     1.427981    2.287053
     chol_50 |   1.771936    .138068     7.34   0.000     1.520978    2.064301
      sbp_50 |   2.883486   .6100443     5.01   0.000      1.90473     4.36518
      bmi_10 |   2.864043   .8610826     3.50   0.000     1.588779    5.162924
       smoke |   1.835662   .2588757     4.31   0.000      1.39236    2.420105
      dibpat |   2.064876   .2991801     5.00   0.000     1.554401    2.742992
     bmichol |    .418712    .116608    -3.13   0.002     .2425842    .7227172
      bmisbp |   .1700948   .1225737    -2.46   0.014     .0414284    .6983674
       _cons |    .032955   .0049595   -22.68   0.000     .0245369    .0442611
------------------------------------------------------------------------------
```

- Two individuals excluded above have relatively large influence/leverage and represent possible measurement errors

Regression diagnostics for logistic models

Evaluation of predictor overlap for binary primary predictor

Example: behavior pattern in WCGS

- Use of propensity scores (see lecture 6 & Sect. 9.4) from model for the log odds of reporting type A behavior (using the predictors in the model fitted in the last slide)

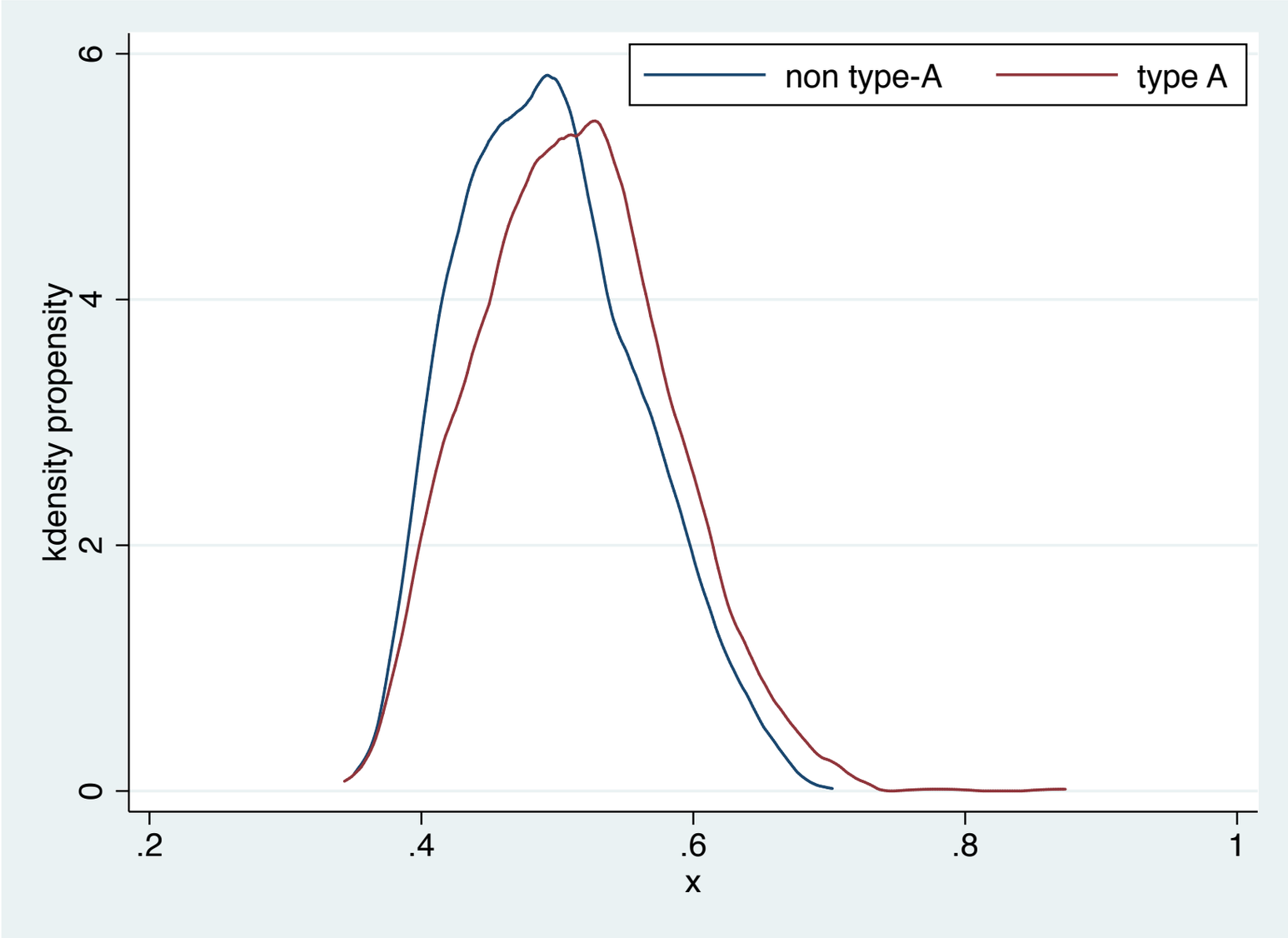
- Fit a model for the relationship between primary (binary) predictor and potential confounders, and use predicted probability to assess overlap:

```
. quietly logistic dibpat age_10 chol_50 sbp_50 bmi_10 i.smoke bmichol bmisbp if (id!
=12237 & id!=12453)
. predict propensity
(option pr assumed; Pr(dibpat))
(12 missing values generated)
. graph twoway kdensity propensity if dibpat == 0 || kdensity propensity if dibpat ==
1, legend(order(1 "non type-A" 2 "type A" ) pos(2) ring(0))
```

Regression diagnostics for logistic models

Example: behavior pattern in WCGS

Overlap plot of propensity scores:



- No evidence for serious lack of overlap

Regression diagnostics for logistic models

Model misspecification:

- The logistic model may not be the right model for various reasons:
  - fundamental form of the model is wrong (e.g. a risk difference or relative risk model is more appropriate)
  - key features or variables are omitted
  - included predictors are incorrectly specified (e.g. omitted interactions or nonlinear terms)
- The effects of misspecification vary depending on the cause and severity of the problem and include biased or inappropriate parameter estimates, and incorrect measures of variability

Regression diagnostics for logistic models

Model misspecification:

- The `linktest` command provides a preliminary means of assessing adequacy of the logistic model. This command addresses the previously fitted model:

```
. quietly logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp if (id!
=12237 & id!=12453)
. linktest

Logistic regression                                     Number of obs   =       3140
                                                        LR chi2(2)      =     201.07
                                                        Prob > chi2     =     0.0000
Log likelihood = -786.47515                             Pseudo R2       =     0.1133

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   .5834516   .3023552     1.93   0.054    -.0091536    1.176057
      _hatsq |  -.0960448   .0680928    -1.41   0.158    -.2295042    .0374146
       _cons |  -.3801513   .3193015    -1.19   0.234    -1.005971    .2456682
------------------------------------------------------------------------------
```

- `linktest` focuses on the way the outcome is modeled as a function of the predictors
- a significant result of the Wald test for the predictor `_hatsq` indicates that the model may not be a satisfactory representation of the relationship between outcome and predictors
- a significant result does not reveal the nature of the misspecification (i.e. the choice of the

Regression diagnostics for logistic models

Model misspecification:

- `linktest` repeated for previous model excluding interactions:

```
. quietly logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat if (id!=12237 & id!
=12453)
. linktest

Logistic regression                                     Number of obs   =       3140
                                                        LR chi2(2)      =     189.99
                                                        Prob > chi2     =     0.0000
Log likelihood = -792.01231                             Pseudo R2       =     0.1071
------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   .2879463   .3207197     0.90   0.369    -.3406528    .9165455
      _hatsq |  -.1681891   .0742107    -2.27   0.023    -.3136395   -.0227388
       _cons |  -.6353921    .332168    -1.91   0.056    -1.286429    .0156453
------------------------------------------------------------------------------
```

- excluding interaction terms from previous model gives significant evidence of misspecification

Evaluation of model performance

- Even when the logistic model provides a reasonable representation for the relationship between outcome and predictors, the model may fail to provide reasonable estimates of average outcomes and predictions of individual outcomes

- calibration & discrimination statistics are used to assess performance

  - discrimination: how well does model distinguish outcomes between individuals (e.g. cases from controls, high-risk patients from low, early from late events)?

  - calibration: how accurately does model estimate outcome probabilities?

Evaluation of model performance

Overall measures of performance for logistic models

- Default likelihood ratio test:
  - measures difference in log likelihood for fitted model compared to model excluding all predictors.
- Pseudo R^2:
  - Measures proportionate change in log likelihood for fitted model compared to model excluding all predictors.
  - Not really equivalent to the R^2 measure from linear models.

$$Pseudo\ R^2 = 1 - \left(\frac{LL_1}{LL_0}\right)$$

LL_1 : fitted model; LL_0 : constant-only model

- *AIC & BIC*: Obtained after a model fit using the `estat ic` command

These measures more useful in relative comparison between models rather than as direct measures of calibration / discrimination

Evaluation of model performance

Goodness of fit tests

A common way to assess calibration performance of a logistic regression model is via a test of "goodness of fit".

- The null hypothesis of these tests is that the outcome probabilities estimated by the model agree with the empirical outcome probabilities.

- The test statistic is constructed by grouping the data and comparing observed (empirical) and expected (based on the estimated model) outcome probabilities for each group using a chi-squared statistic similar to the one used for two-way frequency tables. (i.e. this approach measures model calibration.)

- Rejection of the null hypothesis indicates appreciable discrepancy between estimated and empirical outcome probabilities. (i.e. the model doesn't fit well.)

- Failure to reject the null hypothesis doesn't necessarily indicate that the model fits the data well. (Only that it doesn't fit badly.)

- The Hosmer-Lemeshow goodness of fit test is perhaps the most commonly used test of this type in the medical and epidemiological literature.

Evaluation of model performance

Goodness of fit tests

Example: WCGS model (fitted above)

```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp if (id!
=12237 & id!=12453)
Logistic regression                               Number of obs   =       3140
                                                  LR chi2(8)      =     198.96
                                                  Prob > chi2     =     0.0000
Log likelihood = -787.53196                       Pseudo R2       =     0.1121
------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   1.807172   .2171423     4.92   0.000     1.427981    2.287053
     chol_50 |   1.771936    .138068     7.34   0.000     1.520978    2.064301
      sbp_50 |   2.883486   .6100443     5.01   0.000      1.90473     4.36518
      bmi_10 |   2.864043   .8610826     3.50   0.000     1.588779    5.162924
       smoke |   1.835662   .2588757     4.31   0.000      1.39236    2.420105
      dibpat |   2.064876   .2991801     5.00   0.000     1.554401    2.742992
     bmichol |    .418712    .116608    -3.13   0.002     .2425842    .7227172
      bmisbp |   .1700948   .1225737    -2.46   0.014     .0414284    .6983674
       _cons |    .032955   .0049595   -22.68   0.000     .0245369    .0442611
------------------------------------------------------------------------------
```

Evaluation of model performance

The Hosmer-Lemeshow goodness of fit test is performed by the `estat` command in Stata after a regression model has been fitted

```
. estat gof, table group(10)

Logistic model for chd69, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |    Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+---------+-------+-------+-------+-------+-------|
  |     1 |  0.0156 |     1 |   3.2 |   313 | 310.8 |   314 |
  |     2 |  0.0252 |     6 |   6.4 |   308 | 307.6 |   314 |
  |     3 |  0.0343 |    11 |   9.3 |   303 | 304.7 |   314 |
  |     4 |  0.0451 |    13 |  12.4 |   301 | 301.6 |   314 |
  |     5 |  0.0573 |    17 |  16.0 |   297 | 298.0 |   314 |
  |-------+---------+-------+-------+-------+-------+-------|
  |     6 |  0.0728 |    12 |  20.4 |   302 | 293.6 |   314 |
  |     7 |  0.0962 |    27 |  26.5 |   287 | 287.5 |   314 |
  |     8 |  0.1263 |    43 |  34.7 |   271 | 279.3 |   314 |
  |     9 |  0.1782 |    50 |  46.7 |   264 | 267.3 |   314 |
  |    10 |  0.6178 |    76 |  80.3 |   238 | 233.7 |   314 |
  +--------------------------------------------------------+
      number of observations =      3140
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =         8.50
                Prob > chi2 =         0.3866
```

observed non-outcomes in group (1), based on *mean* predicted prob. (0.0156 is the group maximum)

expected # non-outcomes in group (1)

Conclusion: The Hosmer-Lemeshow test fails to reject the null hypothesis that estimated and observed probabilities agree (P=0.39) .

Evaluation of model performance

Verification of table returned from H-L test (compare to previous slide)

```
. predict pr if (id!=12237 & id!=12453)
. sort pr
. xtile prg=pr, nq(10)
. tabstat pr, by(prg) stat(n mean min max)
Summary for variables: pr
     by categories of: prg (10 quantiles of pr)

     prg |         N      mean       min       max
---------+--------------------------------------------
       1 |       314   .010283  .0007539  .0156125
       2 |       314  .0204795  .0156859  .0250772
       3 |       314  .0295139  .0252241   .034299
       4 |       314   .039636  .0343086  .0450958
       5 |       314  .0510451  .0450968  .0572178
       6 |       314  .0648687  .0572933  .0727661
       7 |       314  .0842371  .0727699  .0960853
       8 |       314  .1106064   .096269  .1262841
       9 |       314  .1487612  .1263497  .1780666
      10 |       314  .2558558  .1782435  .6178426
---------+--------------------------------------------
   Total |      3140  .0815287  .0007539  .6178426
------------------------------------------------------
. dis 314*.010283
3.228862
. dis 314*(1-.010283)
310.77114
```

mean predicted prob. for group (1)

max predicted prob. for group (1)

expected # outcomes in group (1)

expected # non-outcomes in group (1)

Evaluation of model performance

Goodness of fit tests (continued)

- The H-L test has relatively low power. Further, the choice of number of groups is subjective. Too many is bad, because the data in some groups become sparse and instability results (similar to the conventional Chi-squared test). In general, ten groups are recommended.

- In the absence of other model diagnostics, goodness of fit tests are not sufficient to judge how well a particular model fits the data. However, when used along with other diagnostics (e.g. residual and influence analysis) they provide a useful (albeit crude) tool in summarizing the agreement of a model with the data (i.e. calibration performance).

Evaluation of model performance

Discrimination:

- A model that appears to fit well using the H-L test may not provide reasonable predictions of individual outcomes, and to discriminate between outcomes and non-outcomes on the basis of predictors. Measures of discrimination can be used to assess predictive performance of a model in this sense.

- Tools for evaluating for discrimination focus on receiver operating characteristic (ROC) curves

- ROC curves are based on estimated sensitivity and specificity of model-based predictions of individual outcomes, obtained by varying the classification cut-off for defining a "case" between 0 and 1

- The *C* statistic measures the area under the ROC curve for the model (A *C* statistic of 0.5 reflects no discrimination, while 1 reflects perfect discrimination)

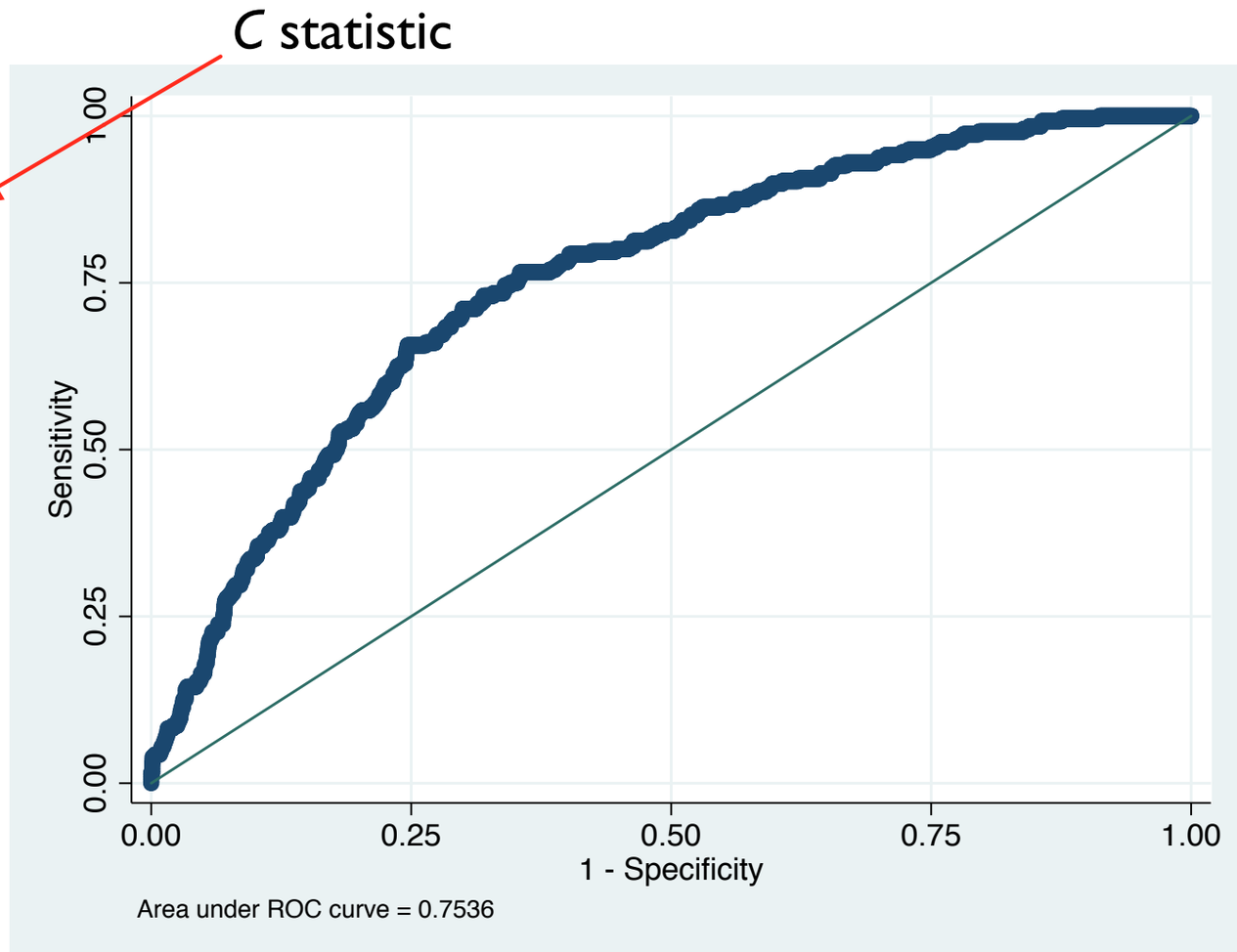
- Discrimination performance for observations used to fit the model provide an *optimistic* assessment relative to assessments based on external data

Evaluation of model performance

Discrimination:

- ROC curves and *C* statistics are obtained using the `lroc` or `roctab` commands (after a regression model has been fit):

```
. lroc

Logistic model for chd69

number of observations =      3140
area under ROC curve   =    0.7536
```



Conclusion: Discrimination is better than chance alone, but misclassification still high

Evaluation of model performance

Discrimination (continued):

Evaluate incremental improvement in discrimination from additional predictor(s)

```
. quietly logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat if (id!=12237 &
id!=12453)
. predict p0
. quietly logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp if
(id!=12237 & id!=12453)
. predict p1
. roccomp chd69 p0 p1
```

	Obs	ROC Area	Std. Err.	-Asymptotic Normal-- [95% Conf.	Interval]
p0	3142	0.7490	0.0151	0.71949	0.77858
p1	3142	0.7540	0.0149	0.72485	0.78314

```
Ho: area(p0) = area(p1)
    chi2(1) =     2.29       Prob>chi2 =   0.1299
```

- Null hypothesis: area under ROC curves for two models equal
- Conclusion: Discrimination not significantly improved by inclusion of additional predictors for interactions between BMI, cholesterol & SBP .

Evaluation of model performance

Some issues not addressable using standard model assessment techniques:

- Measurement error in predictors: leads to attenuated estimates of regression coefficients. Differential measurement errors can lead to biases in either direction.
  - example: systematic or random recording errors in SBP measurements in WCGS
  - example: recall errors in reported number of sexual contacts in CDC heterosexual HIV transmission study

- Misclassified binary outcomes: may bias coefficient estimates
  - non-differential misclassification leads to attenuated estimates
  - differential misclassification can lead to bias in either direction

Predictor selection for logistic regression

- Recall three inferential goals:

  1. predict future outcomes
  2. estimate causal effect of a primary predictor
  3. identify important risk factors for an outcome

- Issues & approaches to all three goals for logistic models mirror those described for standard linear models

Prediction for binary outcomes

Logistic regression prediction

- By specifying valid values for the predictors in any logistic model, outcome risk (P) can be estimated for arbitrary individuals

  - Model can also be used for outcome classification (e.g. for diagnosis)

- Steps in developing prediction models can be summarized as follows:

  1. Identify outcome, candidate predictors, assess missing values, identify meaningful interactions and target possible nonlinearities

  2. Develop model, including variable selection strategies

  3. Check model assumptions (model assessment)

  4. Given "final" model, perform validation to assess calibration & discrimination performance (*Ref*: Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.) ds

  5. Perform external validation

- *Note*: Alternative tools (e.g. classification trees, random forests, support vector machines) often used in place of logistic models

Prediction for binary outcomes

Example: Split sample prediction for WCGS

- Randomly split the dataset into two groups
- Use the first group as a test sample to develop a logistic regression prediction tool
- Validate the prediction tool by predicting the (known) binary outcomes in the second validation sample group
- *Note*: split sample approach is a good illustration of the overall process, but alternatives to validation assessment using bootstrapping / cross validation generally are more efficient in practice

Prediction for binary outcomes

Split sample prediction example:

- The variable `group` has two levels, each designates a randomly selected group of individuals from the WCGS dataset. (Method for obtaining groups will be covered in lab.)

- Step 1: The prediction model is obtained by fitting the candidate model using one group containing 75% of the original sample

```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp if group==1, coef
Logistic regression                                     Number of obs   =       2353
                                                        LR chi2(8)      =     148.24
                                                        Prob > chi2     =     0.0000
Log likelihood = -571.41457                             Pseudo R2       =     0.1148

------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   1.985314   .2795635     4.87   0.000     1.506491    2.616326
     chol_50 |   1.718803   .1592879     5.84   0.000     1.433316    2.061153
      sbp_50 |   3.479123    .891384     4.87   0.000     2.105639    5.748514
      bmi_10 |   2.852689     1.0226     2.92   0.003     1.412945    5.759484
       smoke |   1.716878   .2844706     3.26   0.001     1.240805    2.375613
      dibpat |   1.936007    .325902     3.92   0.000     1.391932    2.692748
     bmichol |   .3499834   .1112369    -3.30   0.001      .187718    .6525126
      bmisbp |   .1914711   .1634857    -1.94   0.053     .0359181    1.020689
```

Formula for P:

$$P = \frac{\exp(\beta_0 + \beta_1 \texttt{age\_10} + \beta_2 \texttt{chol\_50} + \beta_3 \texttt{sbp\_50} + \beta_4 \texttt{bmi\_10} + \beta_5 \texttt{dibpat} + \beta_6 \texttt{smokeb} + \beta_7 \texttt{bmichol} + \beta_8 \texttt{bmisbp})}{1 + \exp(\beta_0 + \beta_1 \texttt{age\_10} + \beta_2 \texttt{chol\_50} + \beta_3 \texttt{sbp\_50} + \beta_4 \texttt{bmi\_10} + \beta_5 \texttt{dibpat} + \beta_6 \texttt{smokeb} + \beta_7 \texttt{bmichol} + \beta_8 \texttt{bmisbp})}$$

Prediction for binary outcomes

Split sample prediction example:

- Step 2: Use `predict` to estimate predicted response probabilities (P) for the second (validation) group (25% of original sample)

```
. predict phat if group==2
. list phat age sbp chol dibpat smoke bmi if group==2
```

```
     +-----------------------------------------------------------+
     |     phat   age   sbp   chol   dibpat        smoke      bmi |
     |-----------------------------------------------------------|
2365. | .0308222    39   130    266    B3,B4   nonsmoker   24.02119 |
2366. | .0090411    41   126    220    B3,B4   nonsmoker   18.24394 |
2367. | .0694368    50   124    234    A1,A2   nonsmoker   23.80396 |
2368. | .0663964    44   116    206    A1,A2      smoker   25.84559 |
2369. | .0526029    40   126    269    B3,B4      smoker   24.80655 |
     |-----------------------------------------------------------|
(first 5 observations shown)
```

- Interpretation: Predicted risk (P) of developing CHD in a ten year interval for an individual with the specified values of the predictors
- Model can be used to develop *risk scores* (Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Stat Med. 2004;23:1631-60.)
- For classification, need cut-off based on P to assign individual outcomes (i.e. 1 or 0)

Prediction for binary outcomes

Split sample prediction example:

- Step 3: Assess prediction performance using validation sample of individuals not used in model development

- *Prediction error* is assessed by the rates of misclassification of outcomes in validation sample, summarized via *C* statistic

- Misclassification is summarized using quantities like:

  - Sensitivity: The proportion of individuals with the outcome that are correctly classified.

  - Specificity: The proportion of individuals without the outcome that are correctly classified.

  - *Example*: Designate indiv. with $P > 0.2$ would as positive outcomes. The `estat` command can be used to give estimates of misclassification error for the second group of individuals:

Prediction for binary outcomes

Split sample prediction example:

- Classification performance for a particular cut-off for P:

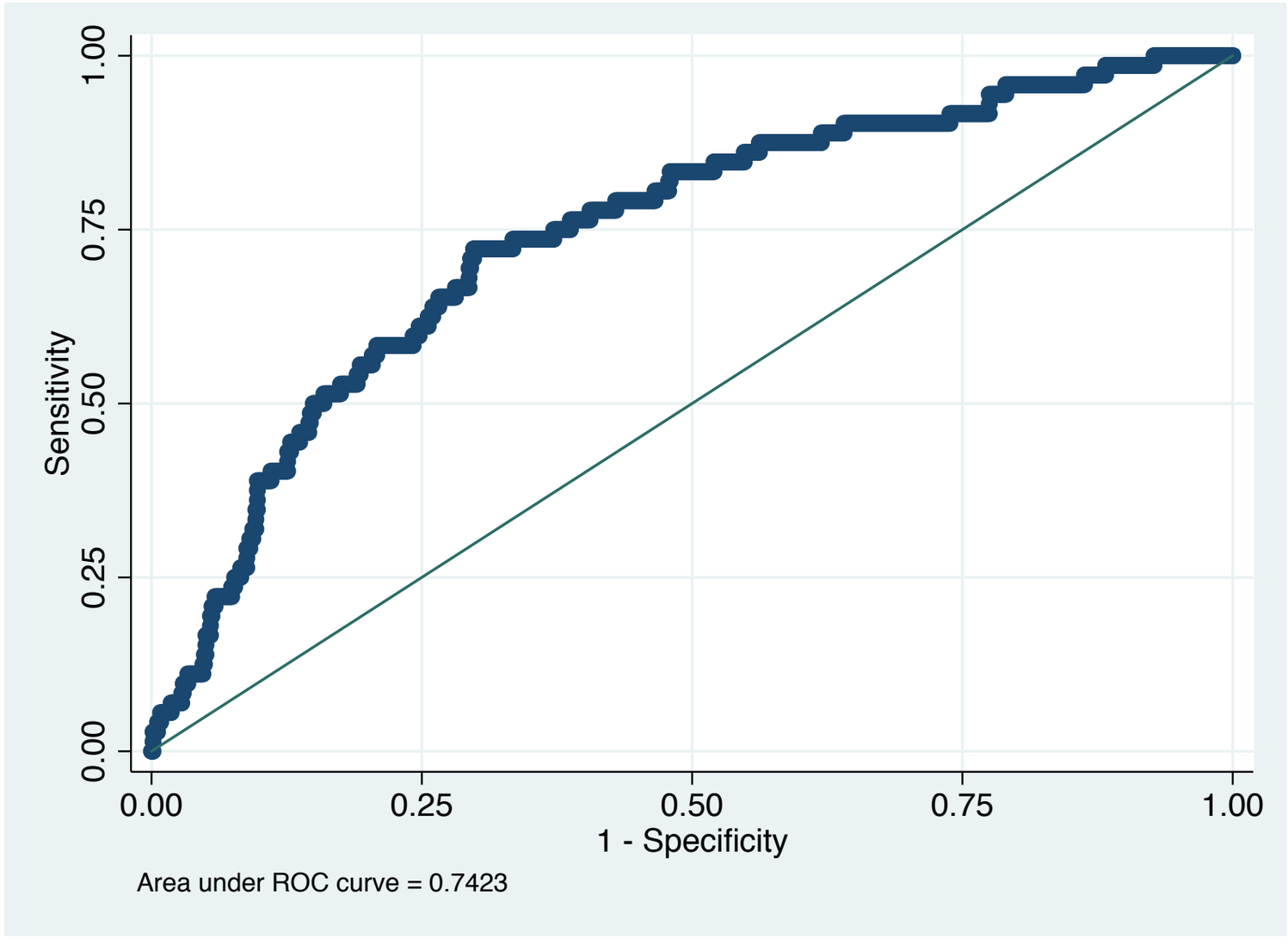
```
. estat classification if group==2, cutoff(0.2)
Logistic model for chd69
              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |        16             42  |         58
     -     |        56            673  |        729
-----------+--------------------------+-----------
   Total   |        72            715  |        787

Classified + if predicted Pr(D) >= .2
True D defined as chd69 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   22.22%
Specificity                     Pr( -|~D)   94.13%
Positive predictive value       Pr( D| +)   27.59%
Negative predictive value       Pr(~D| -)   92.32%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    5.87%
False - rate for true D         Pr( -| D)   77.78%
False + rate for classified +   Pr(~D| +)   72.41%
False - rate for classified -   Pr( D| -)    7.68%
```

Prediction for binary outcomes

Split sample prediction example:

- ROC curve considering all cut-offs for P:

```
. lroc if group==2
```



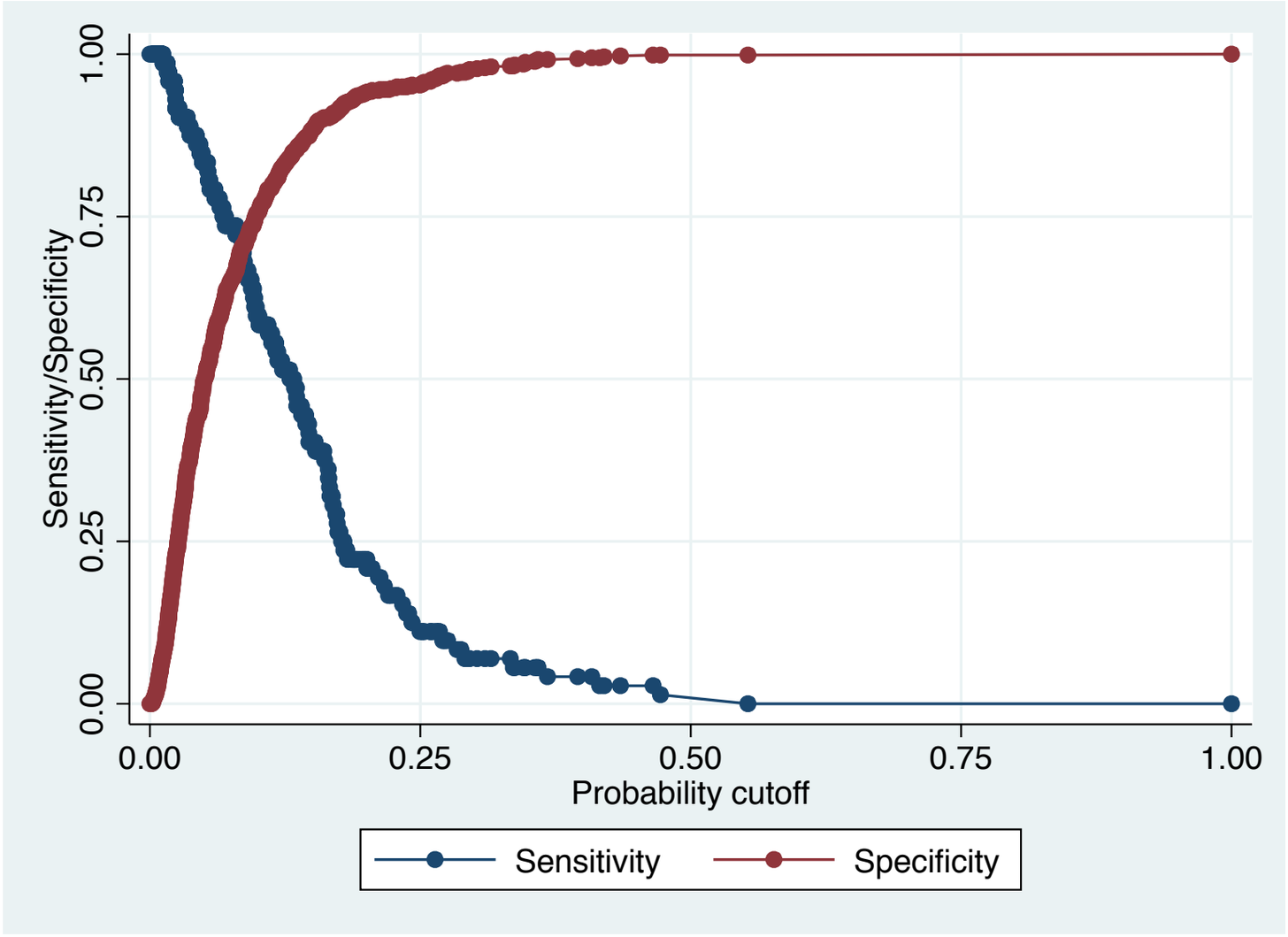
C statistic (0.74) is slightly smaller than the value obtained using the full sample (0.75), reflecting correction for optimism

Prediction for binary outcomes

Split sample prediction example:

- `lsens` for inspecting sensitivity & specificity for all P cut-offs:

```
. lsens if group==2
```



Cut-off of ~ 0.2 leads to approximately equal levels of sensit. & specif.

Prediction for binary outcomes

Preferred alternatives to split sample approach:

- Cross-validation:
  - divide data into 5-10 random subsets
  - leave out each subset, develop model on remainder
  - evaluate discrimination on left-out data
- Bootstrapping:
  - in each of many (e.g. 200) bootstrap samples:
  - fit model to bootstrap sample, obtain predictions
  - evaluate prediction error (PE) in bootstrap, original samples
  - average difference in paired PEs estimates optimism

Prediction for binary outcomes

Example: ten-fold cross-validation for WCGS model

```
. quietly logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp
. predict fitted, pr
(12 missing values generated)
. * Naive estimate of area under the ROC curve
. roctab chd69 fitted

                    ROC                    -Asymptotic Normal--
        Obs       Area     Std. Err.      [95% Conf. Interval]
       -------------------------------------------------------
       3140     0.7536       0.0149        0.72440     0.78287
. * Step 1: divide data into 10 mutually exclusive subsets
. xtile cvgroup = uniform(), nq(10)
. gen cv_fitted = .
(3152 missing values generated)
. forvalues i = 1/10 {
  2. * Step 2: estimate model omitting each subset
. quietly logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp if
cvgroup~=`i'
  3. quietly predict cv_fittedi, pr
  4. * Step 3: save cross-validated statistic for each omitted subset
. quietly replace cv_fitted = cv_fittedi if cvgroup==`i'
  5. quietly drop cv_fittedi
  6. }
.
. * Step 4: calculate cross-validated area under ROC curve
. roctab chd69 cv_fitted

                    ROC                    -Asymptotic Normal--
        Obs       Area     Std. Err.      [95% Conf. Interval]
       -------------------------------------------------------
       3140     0.7426       0.0151        0.71296     0.77226
```

Prediction for binary outcomes

Alternatives to logistic regression: classification trees (recursive partitioning):

1. For each predictor in succession, split the set of outcomes at each possible value of the predictor into two groups and select the value which best classifies the outcomes.

2. Choose the best "split" from among all candidate predictors, and use it to separate the individuals into two groups (nodes)

3. Continue to apply (1) and (2) to all of the nodes produced in preceding splits, "growing" a tree which places each individual in a terminal node

4. Stop growing when the terminal nodes reach a specified level of "purity"

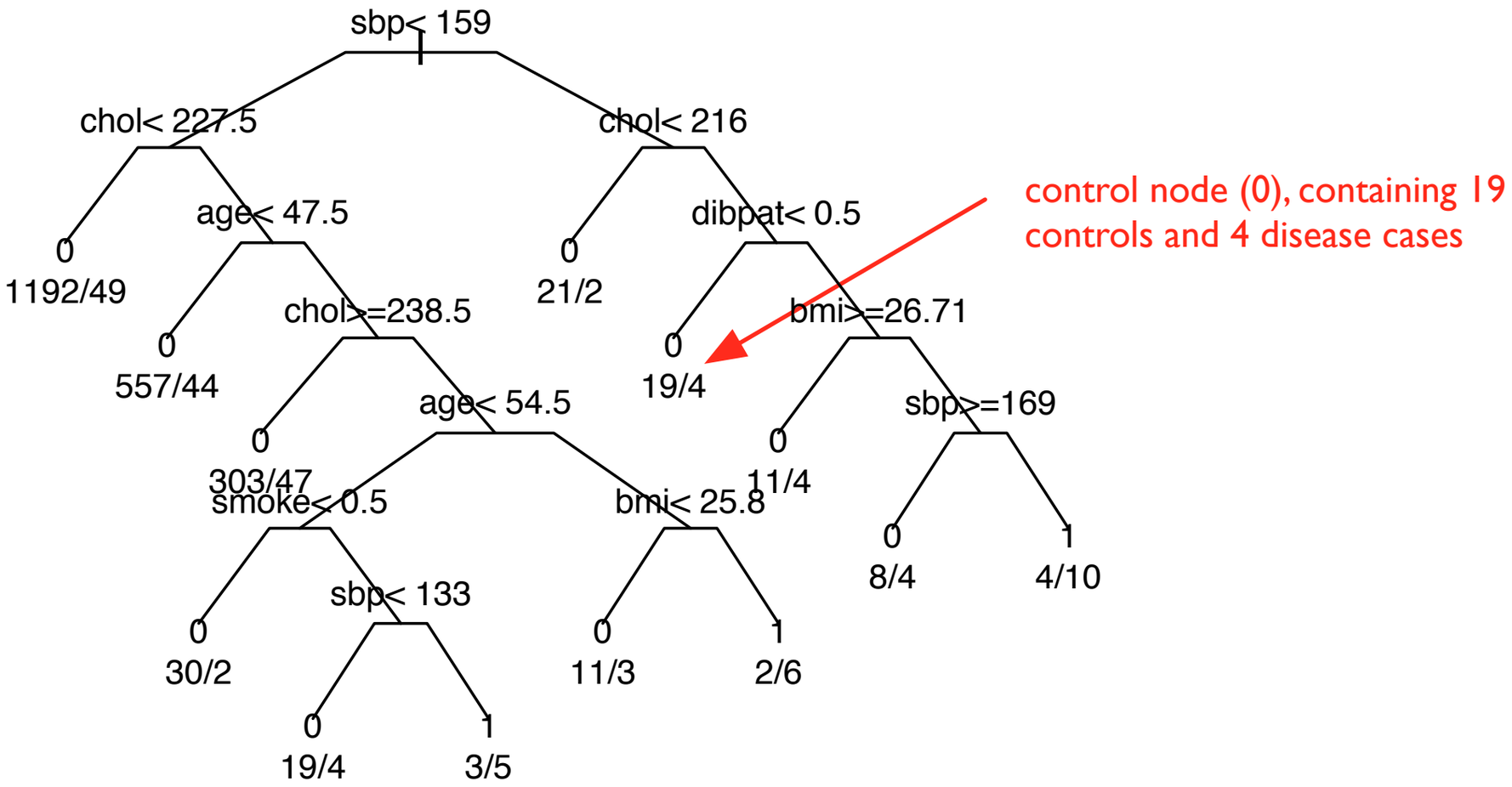
Advantages:

- Nonparametric
- Built in cross-validation assessment of prediction error
- Resistant to outliers
- Flexible in accommodating missing predictor values

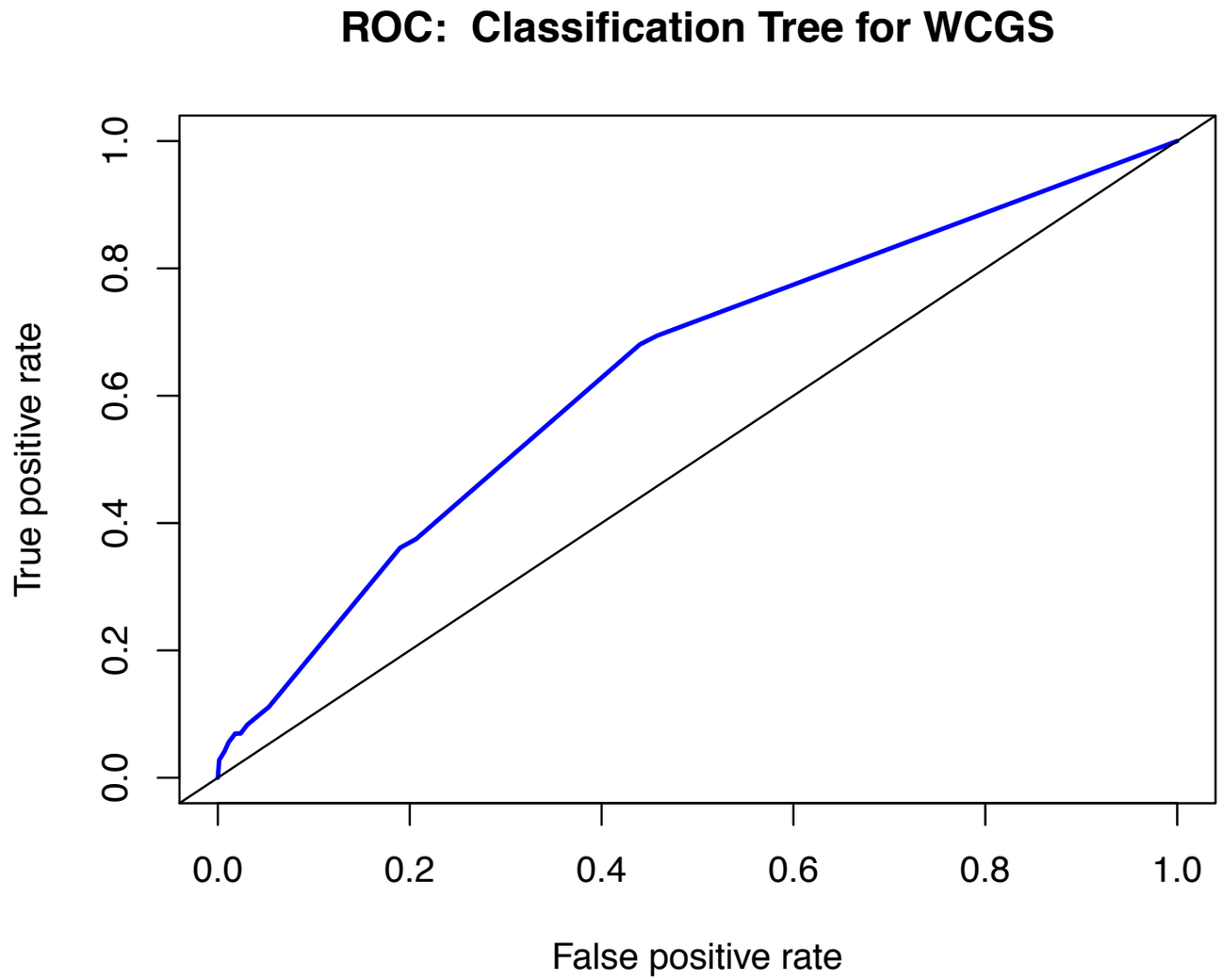
Prediction for binary outcomes

Classification tree example: based on WCGS participants from previous examples

- Tree constructed using the `rpart` function in the statistical package R
- Terminal nodes labeled by outcome class (CHD; 1=positive, 0=negative)



WCGS classification tree example: AUC = 0.65



Additional issues in logistic model construction & evaluation:

- Methods for *predictor selection* in logistic models mirror those covered for linear models with normally distributed outcomes

- In samples with small numbers of events, models shouldn't contain too many predictors

  - rough rule of thumb: ~ 1 predictor for every 10 events
  - propensity score methods can be useful in rare outcome settings

- omitting predictors that are strongly predictive of the outcome but are unassociated with the primary predictor can attenuate the estimated effect of the primary predictor in logistic models - including such predictors results in decreased precision for effect estimates

  - can be important in randomized trials
  - methods based on estimating average causal effects of treatment (e.g. using prop. scores) may be preferred

- logistic models for dependent outcomes will be covered in Biostat 209

Additional notes on predictor selection for logistic regression

- Approaches & goals mirror those used for linear models
- Goal 1: Outcome prediction
  - select to avoid overfitting & minimize optimism-corrected prediction error
- Goal 2: Causal effect of a predictor of primary interest
  - focus on confounding control (ideally motivated by DAGs)
  - change in coefficient criteria affected by non-collapsibility of *OR*s (Greenland S. Invited Commentary: Variable Selection versus Shrinkage in the Control of Multiple Confounders. Am. J. Epidemiol. 2008; 167:523-529.)
- Goal 3: Selecting multiple important predictors
  - Backwards stepwise, Shrinkage (LASSO)