Logistic Regression
Model assessment & Prediction

March 9, 2021

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 restricted cubic splines to check linearity of age in an adjusted model from the WCGS:

```
. mkspline agesp=age, cubic
. 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)

- 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`)
  - Use the linear regression approach (lecture 6) to provide approximate variance inflation measures for binary outcomes

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 (one measure 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:

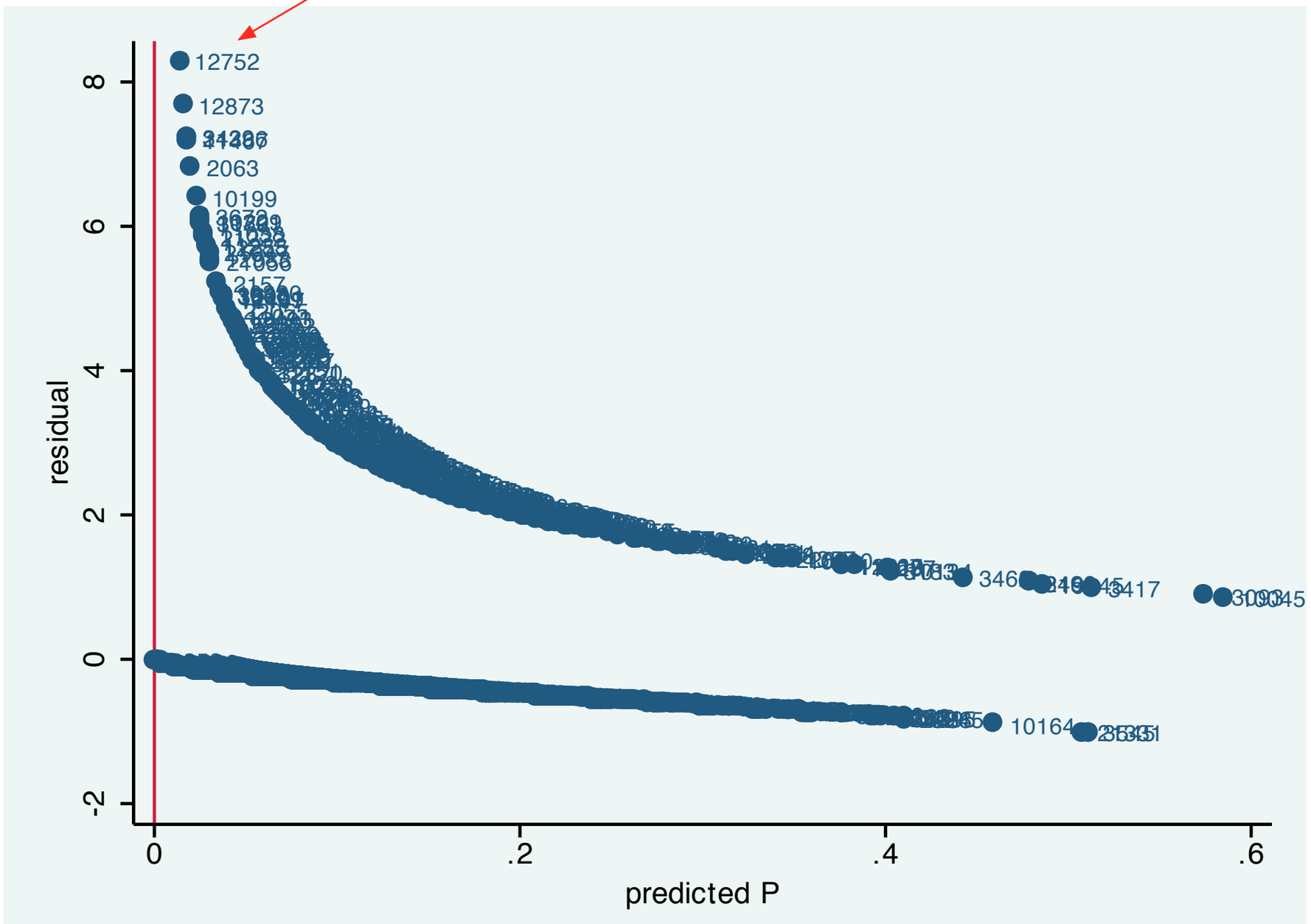
```
. list id chd69 age arcus dibpat chol r if _n<=10

     +-----------------------------------------------------------+
     |   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):

```
twoway (scatter r p, xline(0) mlab(id)) , plotregion(style(none)) xtitle("predicted P") ytitle("residual")
```

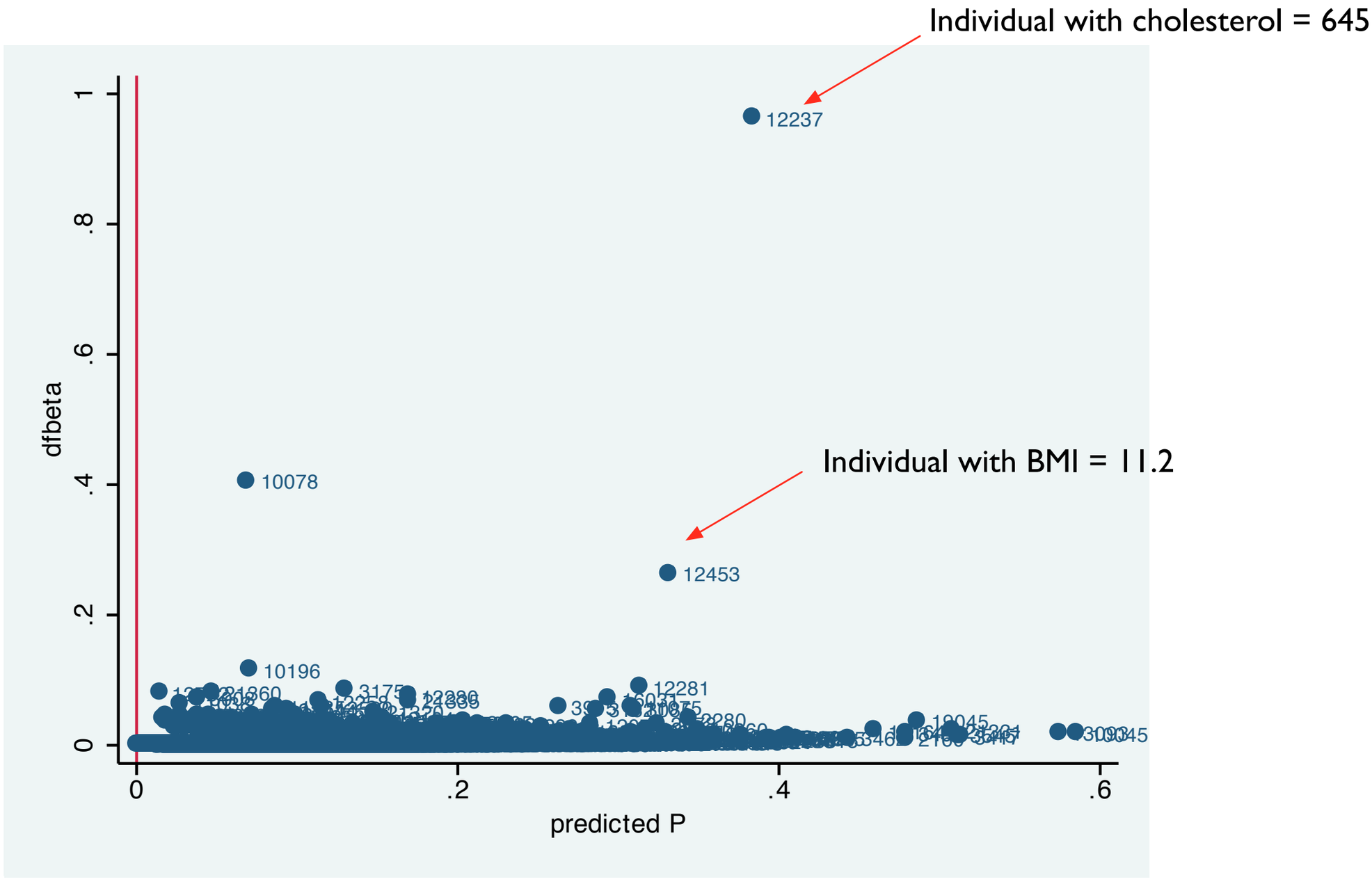
Points labeled by participant ID #



Residuals are harder to interpret than with linear models

Regression diagnostics for logistic models

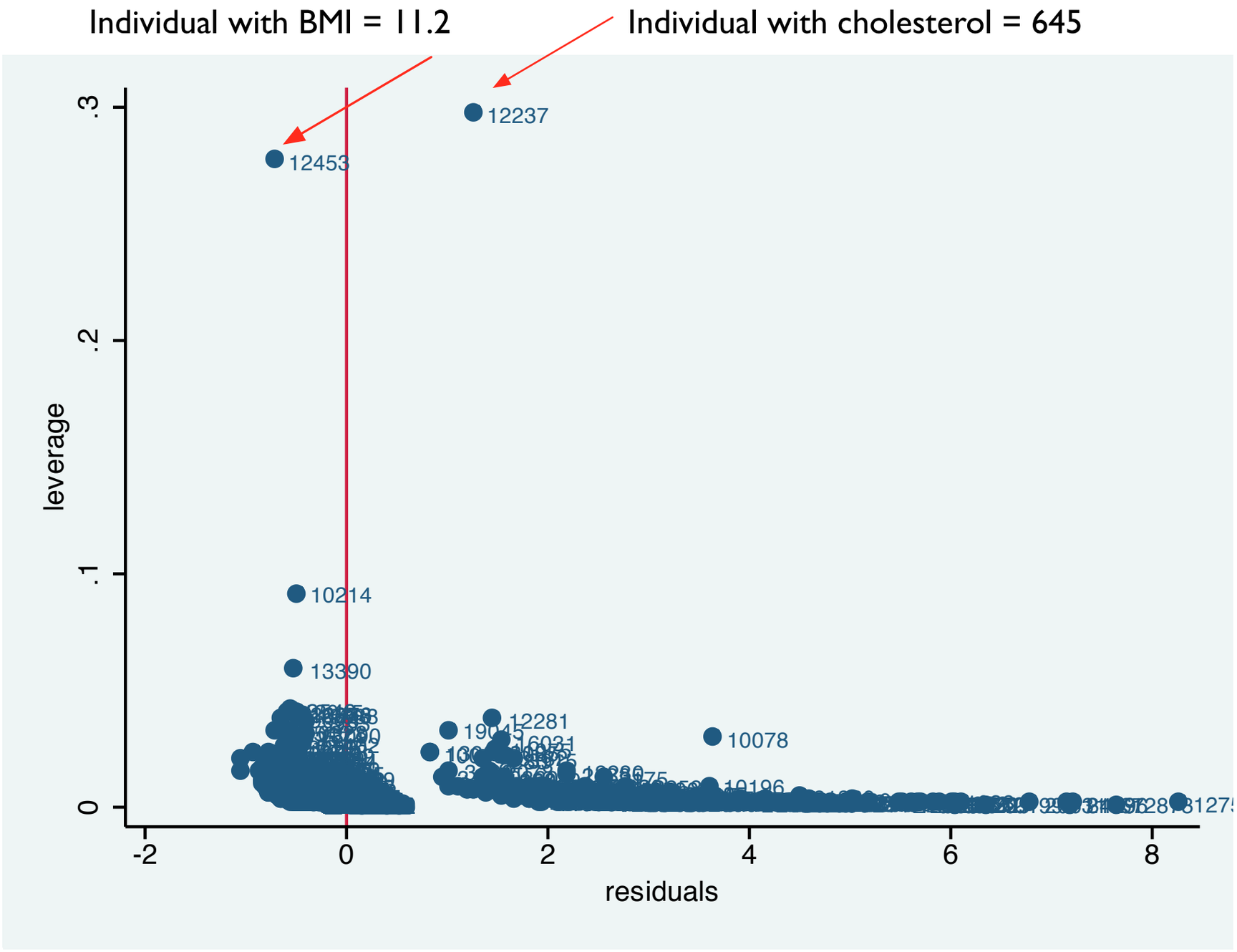
dfbeta's versus fitted values (probability scale):



Note: Default dfbeta values for logistic models are computed by observation (not for each predictor as in linear models). A downloadable Stata module `ldfbeta` provides approximate dfbeta values comparable to those from linear regression.

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:

predictor	Coef. all obs	Coef. 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 moderate 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)

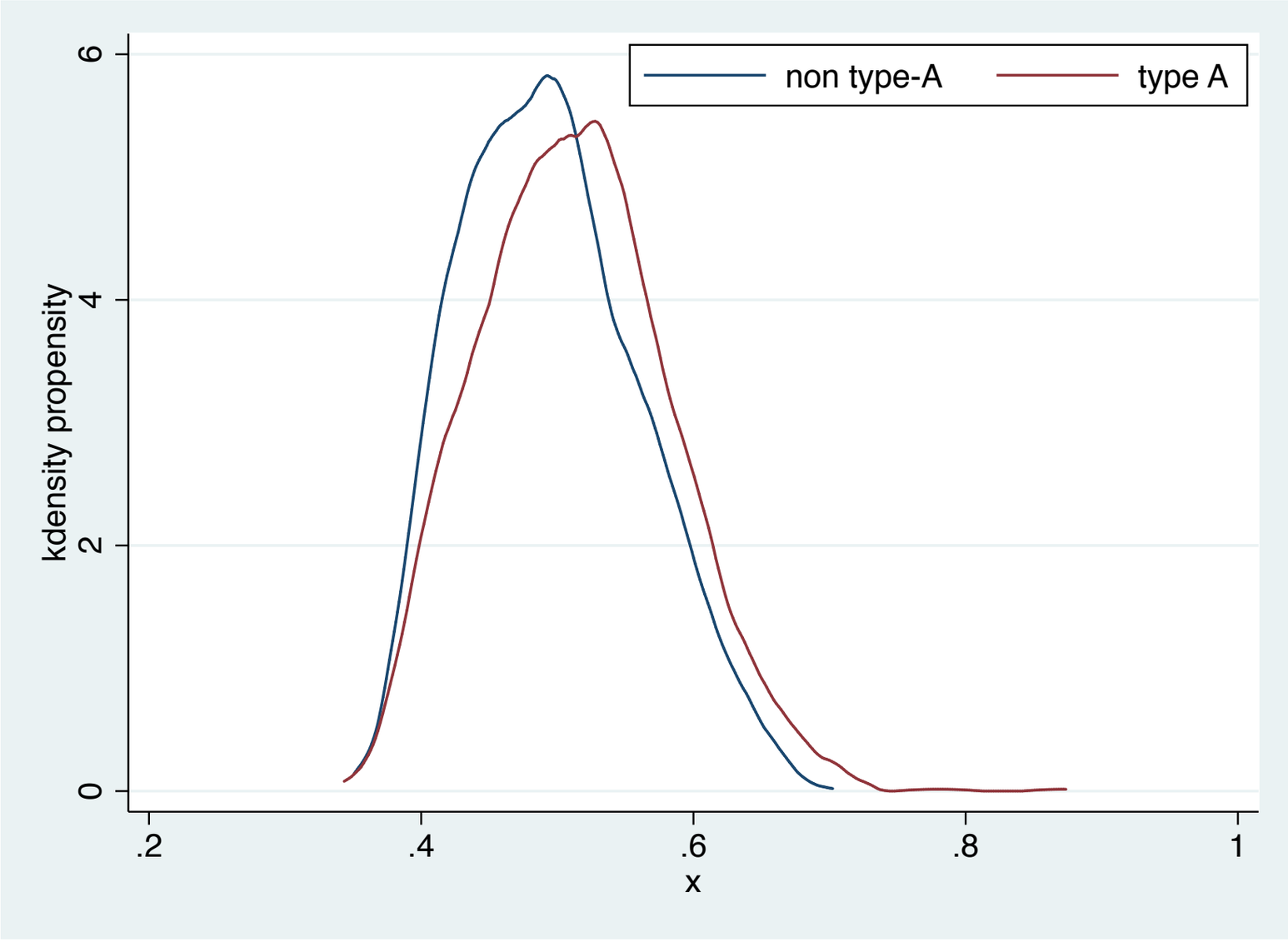
- Fit a model for the relationship between primary (binary) predictor and potential confounders, and use predicted probability to assess overlap:

```
. 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 logistic model and/or omitted/mis-modeled predictors may be issues to consider)

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
- *Note*: The variable `_hat` is just the estimated right-hand side of the fitted model (i.e. obtained using `predict` with the `xb` option (and using the same observations used to fit the original model); `_hatsq` is the square of `_hat`

Regression diagnostics for logistic models

Model misspecification:

- Recall that for linear models, *robust standard errors* were useful if constant variance assump. was violated

- Use of robust SEs for misspecified logistic regression models:

  - may better estimate variability, but *regression coefficients will be biased due to the misspecified model*

- Robust SEs remain useful for well-specified binary regression models in some settings:
  - dependent outcomes (e.g. clustering, repeated measures (Biostat 209)
  - marginal causal effect estimates involving inverse probability weights (Biostat 215)
  - log relative risk models based on Poisson distribution (lecture 11)

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 outcomes in group (1)

expected # outcomes in group (1) - based on *mean* predicted prob. of 0.0103 (next slide)

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 be effective in discriminating between individual 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)

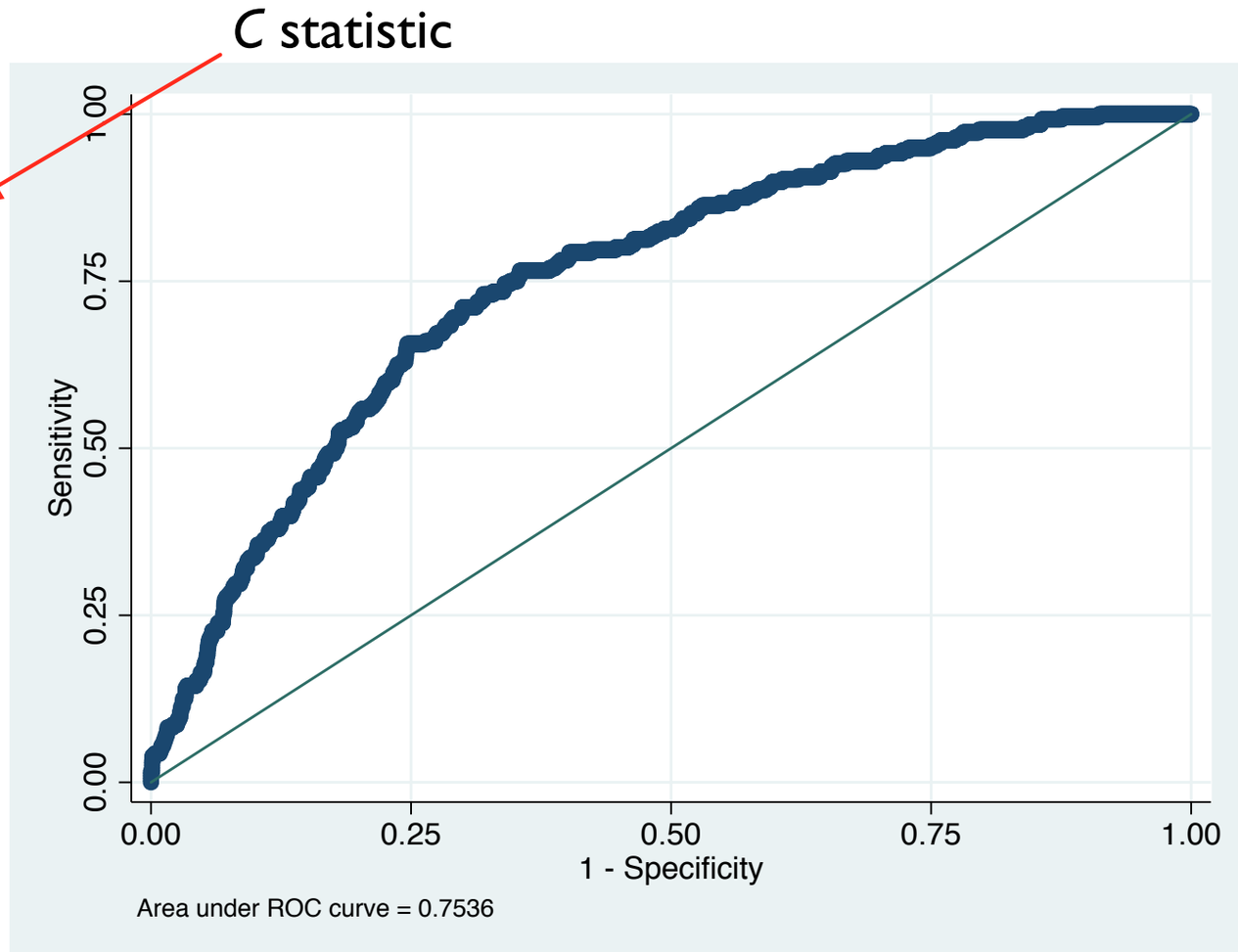
- Discrimination performance for observations used to fit the model provide an *optimistic* assessment relative to assessments based on external data

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
* (results for model fitted above)
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

Some issues not addressable using standard model assessment techniques:

- Measurement error in predictors: typically 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

Additional notes on predictor selection for logistic regression

- Approaches & goals mirror those used for linear models
- Goal 1: Outcome prediction (discussed further below)
  - 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 causal models & DAGs)
  - change in coefficient criteria inappropriate for logistic models (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
  - Stepwise methods have limitations
  - Recently developed shrinkage techniques (e.g. LASSO regression) provide possible alternatives

Additional notes on predictor selection for logistic regression

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

  - rough rule of thumb: ~ 1 predictor for every 10 events in your sample
    Example: WCGS has ~250 events implying ≤ 25 predictors can be supported
  - propensity score methods can be useful in rare outcome settings (biostat 215)

- Omitting predictors that are strongly predictive of the outcome but are unassociated with the primary predictor can attenuate the estimated coefficient for the primary predictor in logistic regression models (related to non-collapsibility). Including such predictors can result in decreased precision for effect estimates.

  - can be an important consideration in randomized trials
  - as a result, methods based on estimating average causal effects of treatment (e.g. using inverse weighting via prop. scores or `margins`) may be preferred to reporting conditional OR as measure of effect

References:

1. Rosenblum M, van der Laan MJ. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. Int J Biostat. 2010;6(1):Article 13.
2. http://thestatsgeek.com/2014/02/01/adjusting-for-baseline-covariates-in-randomized-controlled-trials/

Prediction for binary outcomes

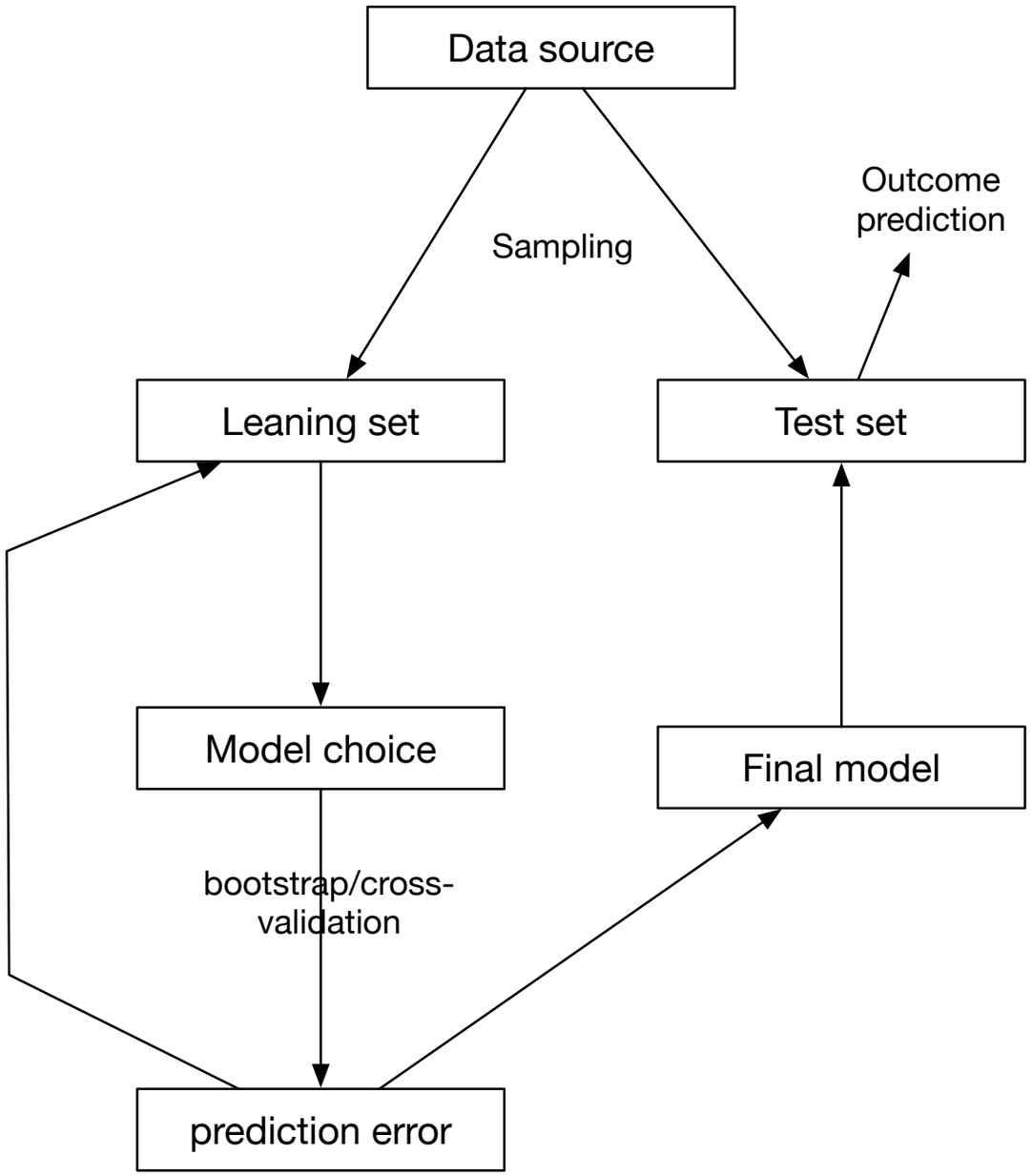
Logistic regression prediction

- 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 model development can be summarized as follows:

1. Identify candidate predictors, assess missing values, identify meaningful interactions and target possible nonlinearities
2. Develop model(s), using appropriate variable selection strategies (see Lect. 7 & Chap. 10)
3. For each candidate model, 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
4. Choose best model (min. pred. error); assess performance in validation data

- *Note*: Alternative tools (e.g. classification trees, random forests, support vector machines) often yield better classification performance than logistic models

Example prediction tool development process (from lecture 7)



Prediction for binary outcomes

Example: Predicting CHD based on WCGS

- Randomly split the dataset into learning and test/validation samples
- Use the learning sample to develop a logistic regression prediction tool
- Validate the prediction tool by predicting the (known) binary outcomes in the test sample

UCSF courses covering more advanced topics in prediction:

- Biostatistics 210
  http://ticr.ucsf.edu/courses/schedule/biostativ.html
- Biostatistics 216
  http://ticr.ucsf.edu/courses/schedule/machine_learning.html

Prediction for binary outcomes

WCGS example:

- The variable `group` has two levels, each designates a randomly selected sub-sample of individuals from the WCGS dataset - group 1(2) is learning(validation) sample

- Steps 1 & 2: The prediction model is obtained by fitting the candidate model using the learning sample (75% of the original sample)

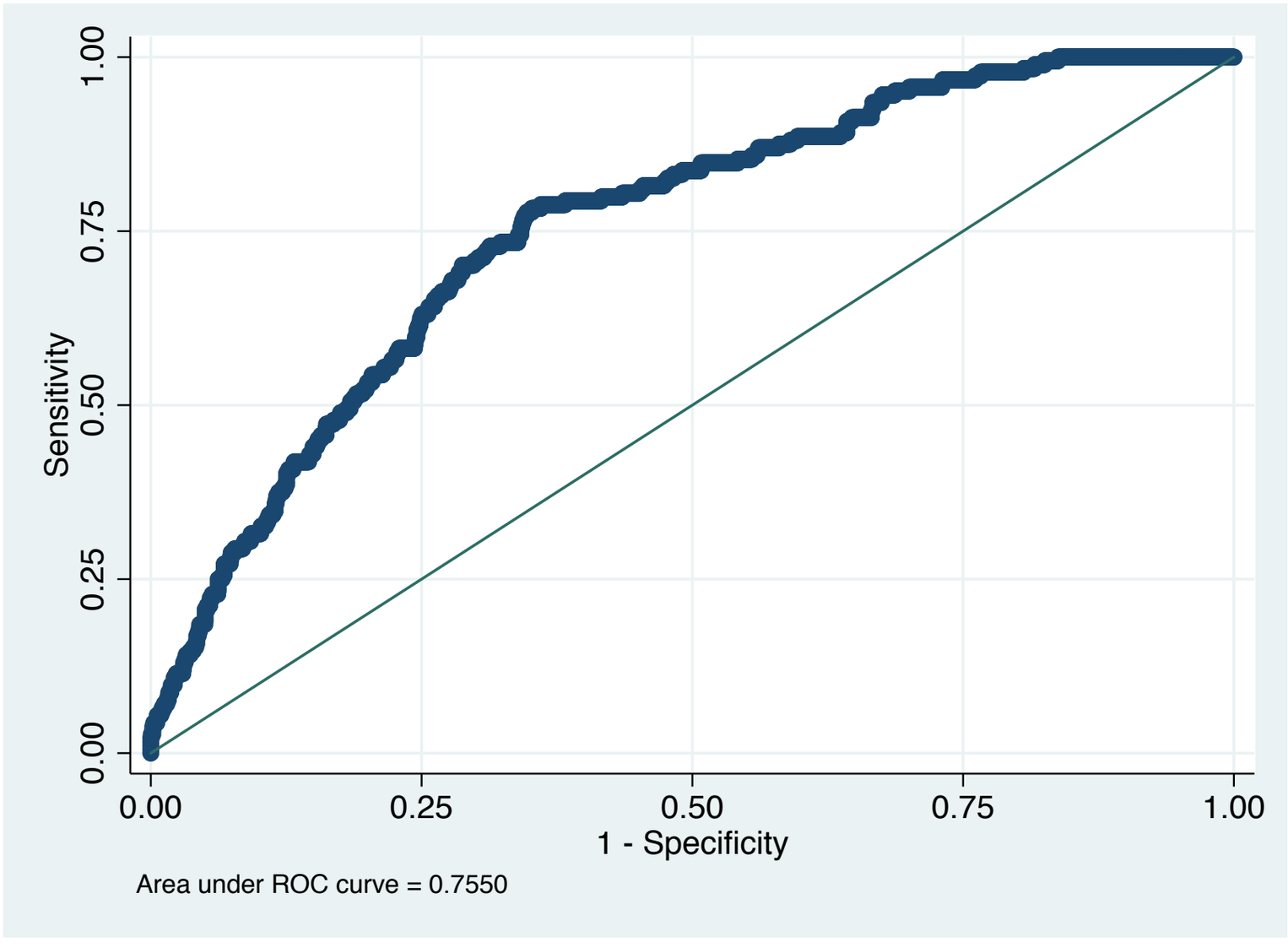
```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp if group==1, coef
Logistic regression                               Number of obs     =      2,353
                                                  LR chi2(8)        =     148.24
                                                  Prob > chi2       =     0.0000
Log likelihood = -571.41457                       Pseudo R2         =     0.1148

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   .6857771   .1408158     4.87   0.000     .4097833     .961771
     chol_50 |   .5416281   .0926738     5.84   0.000     .3599909    .7232653
      sbp_50 |    1.24678   .2562094     4.87   0.000     .7446189    1.748941
      bmi_10 |   1.048262   .3584687     2.92   0.003     .3456762    1.750848
       smoke |   .5405078   .1656906     3.26   0.001     .2157603    .8652553
      dibpat |   .6606274   .1683372     3.92   0.000     .3306925    .9905624
     bmichol |   -1.04987   .3178348    -3.30   0.001    -1.672814   -.4269249
      bmisbp |  -1.653018   .8538402    -1.94   0.053    -3.326514    .0204776
       _cons |  -3.403915    .173878   -19.58   0.000    -3.744709    -3.06312
------------------------------------------------------------------------------
```

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

- Evaluate overall prediction performance in learning sample
- **`.lroc if group==1`**

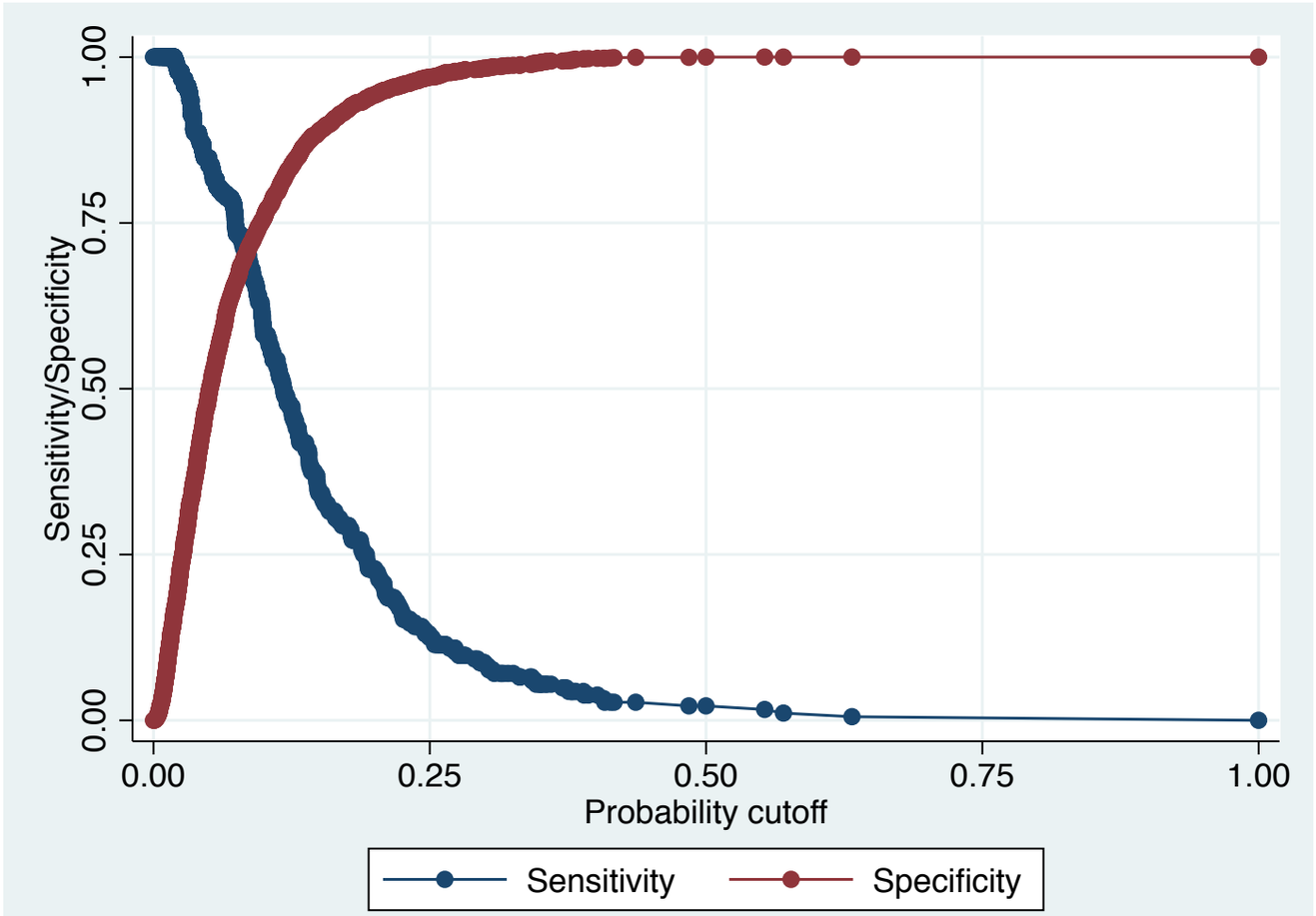


- Interpretation: *C* statistic = 0.7550 - represents (*over-optimistic)* classification performance in the data used to estimate the model coefficients

Prediction for binary outcomes

- If model is to be used for classification purposes, a classification cut-off for predicted probabilities P needs to be developed

```
. predict p
. lsens if group==1
```



Cut-off of ~ 0.085 leads to approximately equal levels of sens. & spec.

Note: procedures for assigning cut-offs based on relative importance weighting of sens. & spec. are often useful (e.g. the *Youden index*)

R*eference*: Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology. 2005;16:73-81.

Prediction for binary outcomes

Split sample prediction example:

- Step 3: Assess out-of-sample prediction performance using cross-validation/ bootstrapping
- *Prediction error* is assessed by the rates of misclassification of outcomes, summarized via *C* statistic
- Misclassification for a particular classification rule 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*: Classify indiv. with $P > 0.085$ as positive outcomes based on the development sample analysis.

Prediction for binary outcomes

Optimism-corrected measures of model performance (based on learning sample):

- Cross-validation:
  - divide learning sample into 5-10 random subsets
  - leave out each subset, fit model on remainder
  - evaluate discrimination performance on left-out data
  - estimate optimism-corrected performance using cross-validated measure
- Bootstrapping:
  - in each of many (e.g. 500) bootstrap samples (i.e. drawn with replacement):
  - fit model to bootstrap sample, obtain outcome predictions
  - evaluate discrimination performance of fitted model in bootstrap & original samples
  - averaged sample-specific optimism over bootstrap replications used to provide overall optimism-corrected performance measure

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 (over-optimistic) 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: restrict to learning sample, divide into 10 mutually exclusive subsets
. keep = group==1
. xtile cvgroup = uniform(), nq(10)
. gen cv_fitted = .
(2353 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 (i.e. optimism-corrected) area under ROC curve
. roctab chd69 cv_fitted
                    ROC                    -Asymptotic Normal--
        Obs       Area     Std. Err.      [95% Conf. Interval]
       -------------------------------------------------------
       2,353     0.7380       0.0177        0.70338     0.77258
```

Prediction for binary outcomes

Example: bootstrap optimism for WCGS model based on learning sample

```
* create a sequential ID variable indexing original observations
. gen sid=_n
* set program variables (note that these need to be created first and set to missing values)
. capture drop id bwt Cb Co optim cons
* dataset includes 2353 observations with non missing values on outcome & predictors
. forvalues j=1(1)500{
  2. * use sample to get sample weights for a single sample of 2353 with-replacement
.     bsample 2353 ,weight(bwt), in 1/2353
  3.     sort sid
  4. * fit & evaluate model on current bootstrap sample
. quietly logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp
[fweight=bwt]
  5.     quietly lroc, nograph
  6.     sort sid
  7.     replace Cb = r(area) in `j'/`j'
  8. * evaluate model on original data
.     quietly lroc [fweight=cons], nograph all
  9.     sort sid
 10.     replace Co = r(area) in `j'/`j'
 11.     replace optim = Cb-Co in `j'/`j'
 12.   }
. summarize optim
. scalar O=r(mean)
* display average optimism & optimism-corrected C-statistic
. dis O
.01050985
. dis 0.7536-O
.74309015
```

Prediction for binary outcomes

- Classification performance in learning sample for a particular cut-off for P:

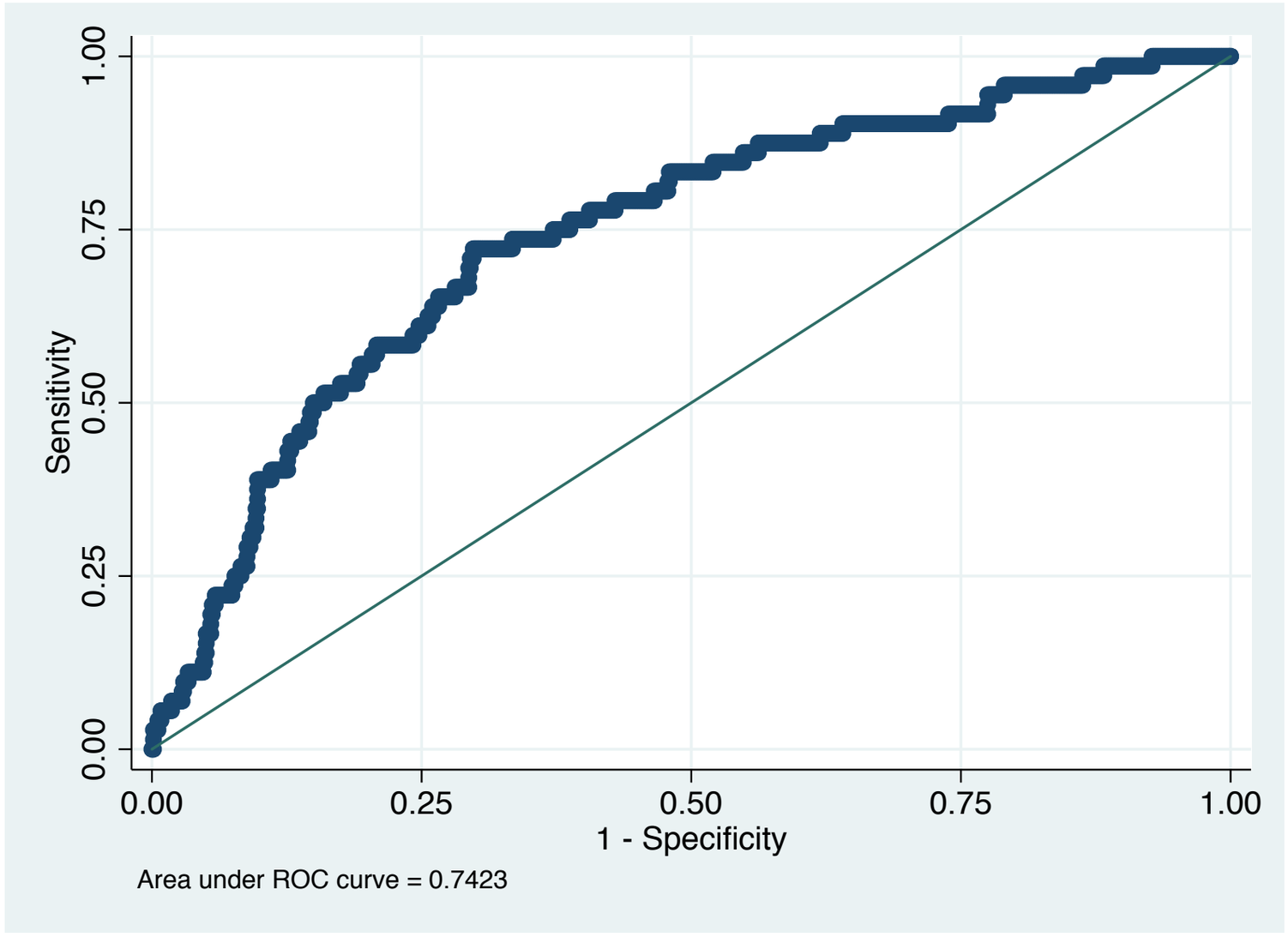
```
. estat classification if group==1, cutoff(0.085)

Logistic model for chd69

              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |       129            635  |        764
     -     |        55           1534  |       1589
-----------+--------------------------+-----------
   Total   |       184           2169  |       2353

Classified + if predicted Pr(D) >= .085
True D defined as chd69 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   70.11%
Specificity                     Pr( -|~D)   70.72%
Positive predictive value       Pr( D| +)   16.88%
Negative predictive value       Pr(~D| -)   96.54%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   29.28%
False - rate for true D         Pr( -| D)   29.89%
False + rate for classified +   Pr(~D| +)   83.12%
False - rate for classified -   Pr( D| -)    3.46%
--------------------------------------------------
```

Prediction for binary outcomes

- Step 4: Assess prediction performance in validation sample

  - ROC curve for the validation sample considering all cut-offs for P:

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



C statistic (0.742) is slightly smaller than the value obtained using the development sample (0.755), reflecting correction for optimism

Prediction for binary outcomes

- Classification performance in validation sample for classification rule based on cut-off developed above:

```
. estat classification if group==2, cutoff(0.085)

Logistic model for chd69

              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |        52            214  |        266
     -     |        20            501  |        521
-----------+--------------------------+-----------
   Total   |        72            715  |        787

Classified + if predicted Pr(D) >= .085
True D defined as chd69 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   72.22%
Specificity                     Pr( -|~D)   70.07%
Positive predictive value       Pr( D| +)   19.55%
Negative predictive value       Pr(~D| -)   96.16%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   29.93%
False - rate for true D         Pr( -| D)   27.78%
False + rate for classified +   Pr(~D| +)   80.45%
False - rate for classified -   Pr( D| -)    3.84%
--------------------------------------------------
Correctly classified                        70.27%
--------------------------------------------------
```

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"

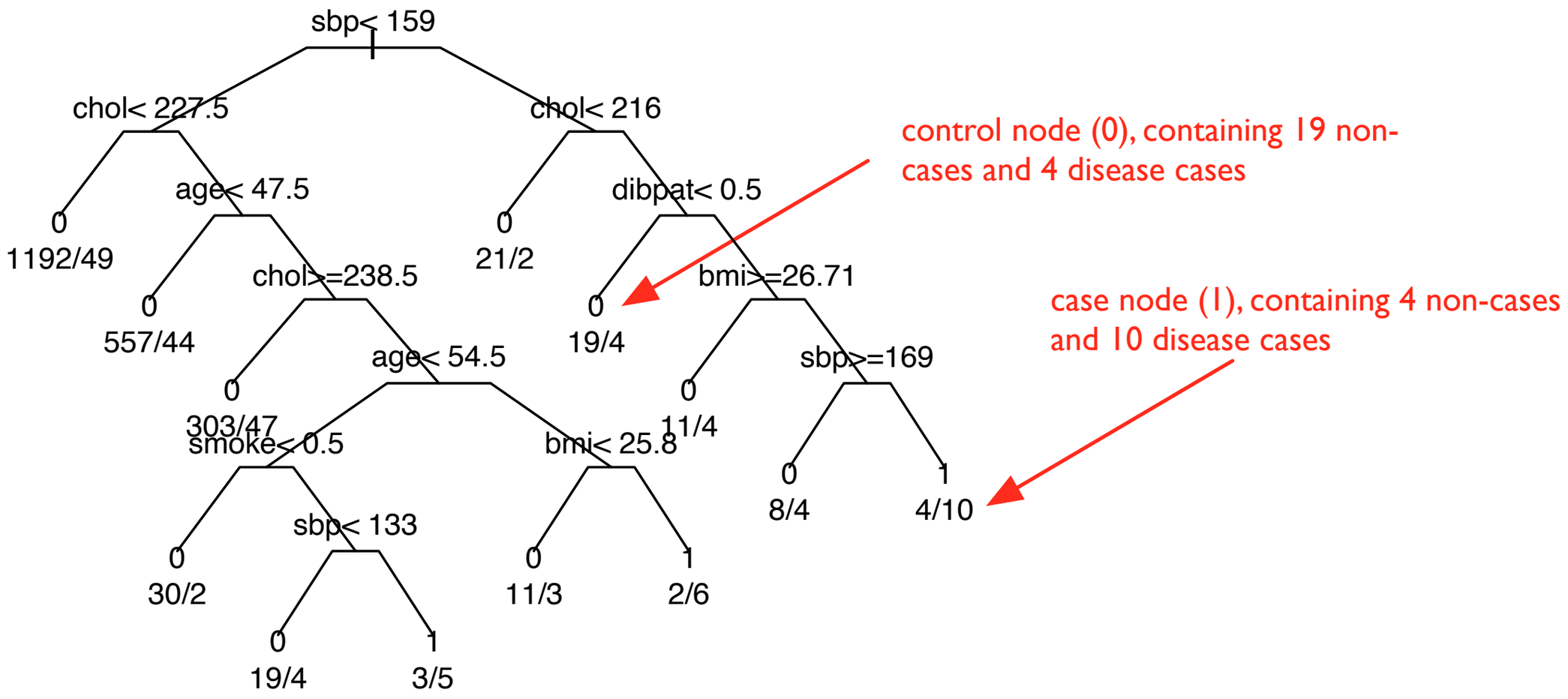
Advantages:

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

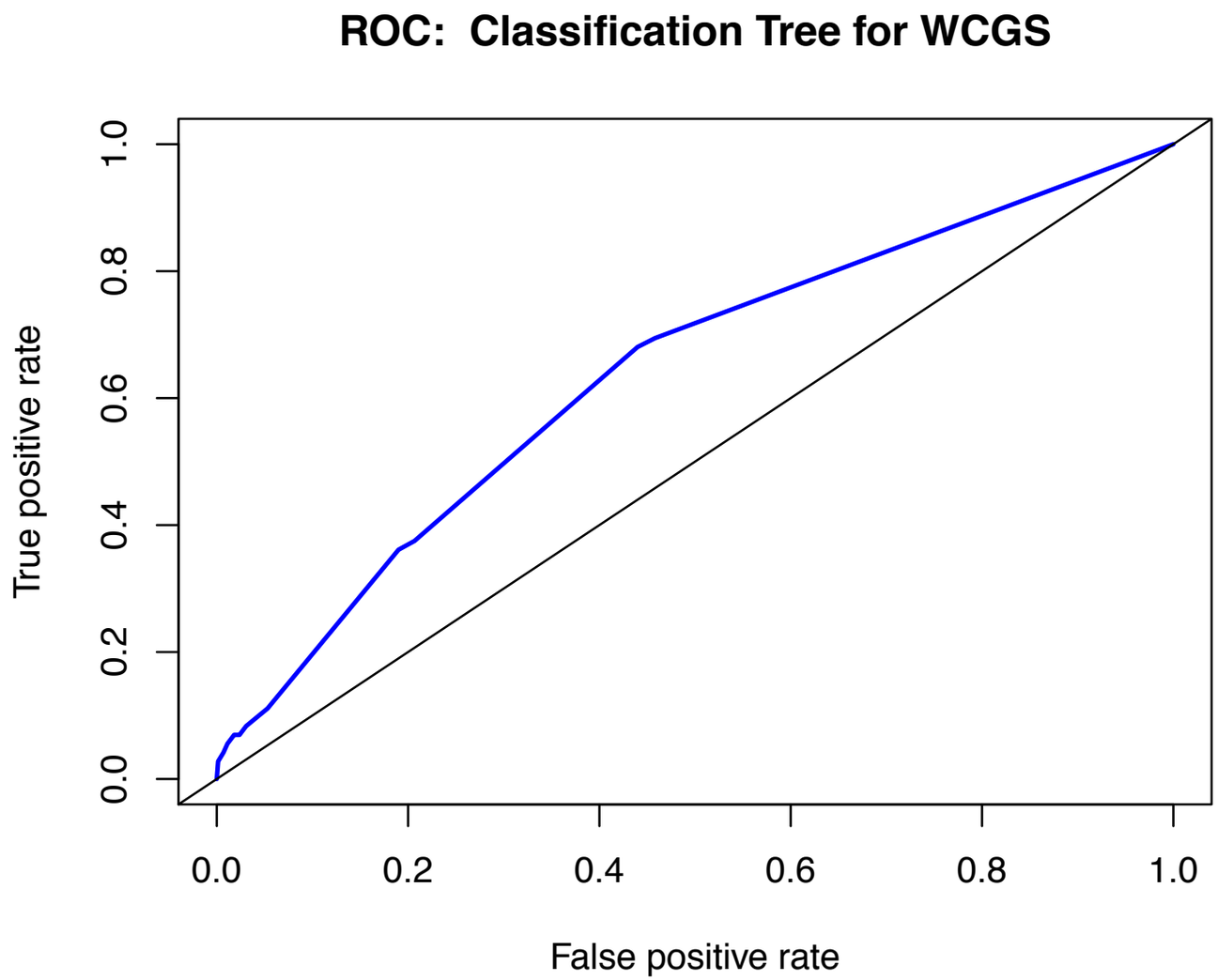
Prediction for binary outcomes

Classification tree example: based on WCGS learning sample

- Tree constructed using the `rpart` function in the statistical package R
- Terminal nodes labeled by outcome class (CHD; 1=positive, 0=negative)
- Default is to weight sensitivity and specificity as equally important



WCGS classification tree applied to validation sample: AUC = 0.65



Alternative regression models for binary outcomes

As discussed previously, there are situations where an alternative model for a binary outcome may be appropriate:

- When risk is thought to be related to predictors in a fashion other than assumed in the logistic model (i.e. linear and additive in the log odds)
  - For example, relative risk and risk difference-based models
- When estimates of predictor effects other than odds ratios are required (e.g. *RR*)
- In cases where external information suggests a specific formulation
  - HIV transmission example discussed in section 5.5.1 of text

Example: relationship between probability of coronary heart disease age and BMI in the WCGS study (excluding the extremely low value of BMI noted above)

Logistic regression via `binreg`:

```
. binreg chd69 age bmi if bmi>12 & bmi!=., or

Generalized linear models                         No. of obs      =      3153
Optimization     : MQL Fisher scoring             Residual df     =      3150
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  1727.411122                   (1/df) Deviance =  .5483845
Pearson          =   3155.69054                   (1/df) Pearson  =  1.001807

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  BIC             = -23649.33
------------------------------------------------------------------------------
             |                 EIM
      chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.076649   .0122104     6.51   0.000     1.052981    1.100849
         bmi |    1.08468   .0264501     3.33   0.001     1.034058     1.13778
       _cons |   .0003617   .0002961    -9.68   0.000     .0000727    .0017999
------------------------------------------------------------------------------
```

Note: The `binreg` comand in Stata allows fitting a number of alternative *generalized linear regression* models for binary outcomes, including *logit* (i.e. log odds), *log* (log probability), and *linear* (i.e. linear in probability).

Generalized linear models (fitted in Stata with `glm`) include binary and linear regression models studied in this course, as well as poisson regression and a number of other examples. These models are specified by the distributional `family` of the outcome, and the `link` between the mean and predictors For the previous example:

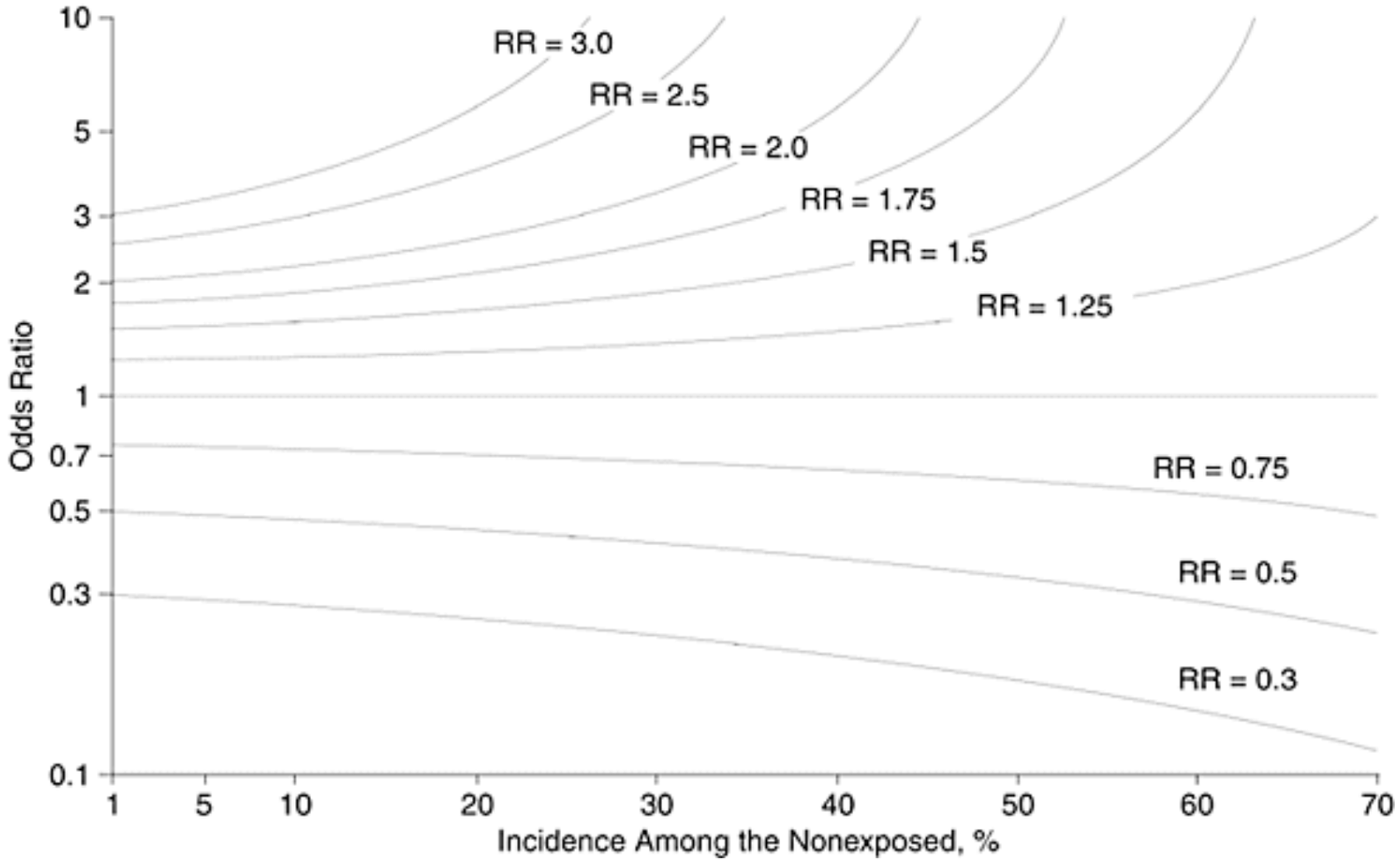
```
. glm chd69 age bmi if bmi>12 & bmi!=., family(binomial) link(logit) eform

Generalized linear models                         No. of obs      =      3,153
Optimization     : ML                             Residual df     =      3,150
                                                  Scale parameter =          1
Deviance         =  1727.411122                   (1/df) Deviance =   .5483845
Pearson          =  3155.775804                   (1/df) Pearson  =   1.001834

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =   .5497657
Log likelihood   =  -863.705561                   BIC             =  -23649.33
------------------------------------------------------------------------------
             |                 OIM
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.076649   .0122105     6.51   0.000      1.05298    1.100849
         bmi |    1.08468   .0264503     3.33   0.001     1.034057     1.13778
       _cons |   .0003617   .0002961    -9.68   0.000     .0000727    .0017999
------------------------------------------------------------------------------
```

- `logistic`, `binreg` & `glm` can all be used to fit the logistic model
- `binreg` also allows relative and risk diff. models to be fitted
- `glm` applies to the entire family of generalized linear models

Problem: Want to summarize age adjusted effect of BMI on CHD using a *relative risk* rather than an odds ratio

- *Why not use the logistic model and either assume that estimated ORs approximate RRs, or use the probability form of the model to estimate RRs?*

  - Odds ratios only approximate relatives risks in the rare outcome setting (see *figure*)
  - Relative risks estimated using the probability form of the logistic model depend on other predictors in the model



Problem: Want to summarize age adjusted effect of BMI using a *relative risk* rather than an odds ratio

- *Solution*: Use a *log relative risk regression* model based on the log of the outcome probability:

$$\log P(\texttt{age},\texttt{bmi}) = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{bmi}$$

- The coefficient β_2 is interpreted as the log relative risk associated with a unit change in `bmi`, holding `age` fixed.

- The coefficient β_1 is interpreted as the log relative risk associated with a unit change in `age`, holding `bmi` fixed.

- *Note*: recall that the coefficients for this model have to be estimated so that the left-hand side specifies a valid outcome probability (i.e. $0 \leq P \leq 1$, or equivalently $-\infty < \log P \leq 0$) for all values of `bmi` and `age`. This *constraint sometimes causes fitting problems* not encountered with the logistic model.

- Similar constraints and numerical difficulties arise in fitting additive risk difference models

Relative risk regression (*continued*)

- log relative risk models can be fitted in Stata using the `binreg` command:

```
. binreg chd69 age bmi if bmi>12 & bmi!=., rr iterate(100)
Generalized linear models                         No. of obs      =      3153
Optimization     : MQL Fisher scoring             Residual df     =      3150
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  1727.536638                   (1/df) Deviance =  .5484243
Pearson          =  3149.490834                   (1/df) Pearson  =  .9998384
Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u)                   [Log]

                                                  BIC             = -23649.21
------------------------------------------------------------------------------
             |                 EIM
      chd69 | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.068756   .0107595     6.61   0.000     1.047874    1.090053
         bmi |   1.073924   .0228021     3.36   0.001      1.03015    1.119558
       _cons |   .0005968   .0004258   -10.41   0.000     .0001474    .0024165
------------------------------------------------------------------------------
```

- Results quite similar to logistic model in this case

- Because of the constraint mentioned above, sometimes Stata will fail to estimate the parameters from a log rel. risk model. As shown in Lab 10, using the `iterate(100)` option stops estimation after 100 steps if convergence problems are encountered

Relative risk regression (*continued*)

- Approximate log relative risk regression for binary outcomes using the `poisson` regression command:

```
. poisson chd69 age bmi if bmi>12 & bmi!=., irr robust

Poisson regression                                Number of obs   =       3153
                                                  Wald chi2(2)    =      57.00
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -876.93885                 Pseudo R2       =     0.0270
------------------------------------------------------------------------------
             |               Robust
       chd69 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.068913   .0106896     6.66   0.000     1.048165     1.09007
         bmi |   1.075168   .0218344     3.57   0.000     1.033214    1.118825
       _cons |   .0005759   .0004083   -10.52   0.000     .0001435    .0023113
------------------------------------------------------------------------------
```

- This approach is generally recommended for models involving multiple predictors (especially those including continuous variables)

- Results very close to the estimates on the previous slide
- `robust` variance option *required* to obtain appropriate standard error estimates

Reference: Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. Amer J Epidemiol 2004; 159:702-6.

Example: coronary heart disease, age and BMI in the WCGS study (*continued*)

Problem: Want to summarize age adjusted effect of BMI using a *risk difference* (or excess risk) rather than an odds ratio

- *Solution*: *risk difference regression* model based directly on the outcome probability P:

$$P(\texttt{age},\texttt{bmi}) = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{bmi}$$

- The coefficient β_2 is interpreted as the risk difference associated with a unit change in `bmi`, holding `age` fixed.

- The coefficient β_1 is interpreted as the risk difference associated with a unit change in `age`, holding `bmi` fixed.

- *Note*: model coefficients have to be estimated so that the left-hand side still defines a valid outcome probability (i.e. $0 \leq P \leq 1$) for all values of `bmi` and `age`. Fitting problems are common, especially for continuous predictors.

Risk difference regression (*continued*)

- Here is the `binreg` risk difference model fit of the last example:

```
. binreg chd69 age bmi if bmi>12 & bmi!=., rd iterate(100)
Iteration 1:   deviance =  1731.716
Iteration 2:   deviance =  1731.192
Iteration 3:   deviance =  1731.189
Iteration 4:   deviance =  1731.189
Iteration 5:   deviance =  1731.189
.
.
.
Iteration 95:  deviance =  1731.182
Iteration 96:  deviance =  1731.182
Iteration 97:  deviance =  1731.182
Iteration 98:  deviance =  1731.181
Iteration 99:  deviance =  1731.181
Iteration 100: deviance =  1731.181
convergence not achieved
r(430);
```

- No convergence of the estimation process in 100 iterations (convergence is never achieved in this example, even if the iteration limit is increased to >1000).

- Model excluding BMI or using a categorical version converges (see next page):

Risk difference regression (*continued*)

- Using bmi recoded into 4 groups based on quartiles:

```
. xtile bmicat=bmi if bmi>12,nq(4)
. binreg chd69 age i.bmicat, rd

Generalized linear models                         No. of obs      =      3153
Optimization     : MQL Fisher scoring             Residual df     =      3148
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  1730.073128                   (1/df) Deviance =  .5495785
Pearson          =  3153.048459                   (1/df) Pearson  =  1.001604

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = u                       [Identity]
                                                  BIC             = -23630.56
------------------------------------------------------------------------------
             |                 EIM
      chd69 | Risk Diff.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0051902    .000893     5.81   0.000        .00344    .0069405
             |
      bmicat |
           2 |   .0170412   .0118722     1.44   0.151    -.0062278    .0403102
           3 |   .0255155   .0121295     2.10   0.035      .001742     .049289
           4 |   .0380856   .0129088     2.95   0.003     .0127848    .0633865
             |
       _cons |  -.1784397   .0398784    -4.47   0.000    -.2565999   -.1002795
------------------------------------------------------------------------------
```

- Categorization leads to reasonable estimates, but a somewhat different interpretation

Example: Binary regression models for infectious disease transmission

Problem: Want to link a binary infection outcome to exposure to infection (and, possibly additional predictors characterizing susceptibility and/or infectiousness).

- Assume exposure consists of potentially infectious contacts (expressed as counts); outcome is acquisition of infection following exposure

- Assume that the risk of infection carried by a single infectious contact is λ

- The probability of escaping infection following k independent contacts is $(1 - \lambda)^k$

- The corresponding probability of acquiring infection is then $P = 1 - (1 - \lambda)^k$

- Via the *complementary log-log transformation* this has the linear form:

$$\log\left[-\log(1 - P)\right] = \log\left[-\log(1 - \lambda)\right] + \log(k)$$

- Suggests a linear model with intercept $\log$[-log$(1 - \lambda)$], and predictor $\log(k)$ with coefficient *fixed* at 1.

- A predictor with a fixed coefficient (i.e. assumed known) is called an *offset* in regression models.

Example: Estimate the per-contact risk of HIV from the CDC transmission study

```
. glm hivp, family(binomial) link(cloglog) offset(logcontacts)
Generalized linear models                         No. of obs      =         31
Optimization     : ML                             Residual df     =         30
                                                  Scale parameter =          1
Deviance         =   40.8340195                   (1/df) Deviance =   1.361134
Pearson          =  84.90572493                   (1/df) Pearson  =   2.830191

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(-ln(1-u))            [Complementary log-log]

                                                  AIC             =   1.381743
Log likelihood   = -20.41700975                   BIC             =   -62.1856
------------------------------------------------------------------------------
             |                 OIM
        hivp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -7.033126   .3803284   -18.49   0.000    -7.778556   -6.287696
 logcontacts |          1  (offset)
------------------------------------------------------------------------------

. * estimated per-contact risk (λ):
. dis 1-exp(-exp(_b[_cons]))
.00088178
```

- Small sample size suggests that a bootstrap confid. interval may be appropriate here (section 5.5.1 of textbook)

Conclusions about alternate binary regression approaches:

- Alternative models (e.g. relative risk / risk difference) may be preferable in some instances
  - requested by reviewers
  - scientific grounds
  - marked improvement in fit over the logistic

- Alternate models models may raise numerical estimation problems that typically don't affect logistic regression
  - The Poisson regression approach illustrated on slide 53 is the recommended way to obtain estimated relative risks for binary outcomes and multiple predictors (*not valid for case-control studies*)

- The logistic model should generally be the default choice unless these other factors are compelling
  - recall that the logistic model can also be used as a basis for causal inference in many situations (e.g. for estimating average causal effects, propensity scores, marginal structural models, etc.)

Alternatives to logistic regression for small samples & rare outcomes

As sample size shrinks:

- Reliability of inferences from standard methods (Wald (Z) test, 95% confidence intervals and even likelihood ratios diminishes
- difficulties in estimating coefficients become more common
- *Exact logistic regression* and *bootstrap methods* can both help
- What is a "small" sample?
- A related issue is *rare outcomes*, which can be an issue even in large samples
  - can cause estimation issues (hence the 10 outcomes per predictor rule of thumb)
  - frequently results in infinite OR estimates

Example: relationship between transmission of HIV to female partners of HIV infected men in CDC transmission study

Problem: Sample of $n = 31$ - calls validity of approx. inferences from `logistic` into question

```
. logistic hivp sexd60 aids
Logistic regression                               Number of obs   =         31
                                                  LR chi2(2)      =       6.81
                                                  Prob > chi2     =     0.0331
Log likelihood = -13.151565                       Pseudo R2       =     0.2058
------------------------------------------------------------------------------
        hivp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sexd60 |   7.081807     9.7318     1.42   0.154      .479099    104.6798
        aids |   12.80809    16.1979     2.02   0.044     1.073995    152.7449
------------------------------------------------------------------------------
```

- Wide confidence intervals demonstrate lack of precision of estimates
- Approximate p-values may be inaccurate for such a small sample with few outcomes

Example: relationship between transmission of HIV to female partners of HIV infected men in CDC transmission study (*continued*)

Exact logistic regression

- Best suited for small samples (<200) and few (predominantly categorical) predictors
- Estimation and inference are permutation based, and don't rely on distributional assumptions

```
. exlogistic hivp sexd60 aids
```

(*Note*: `i.` notation doesn't work)

```
Exact logistic regression                       Number of obs =         31
                                                Model score   =  6.582907
                                                Pr >= score   =     0.0258
------------------------------------------------------------------------------
        hivp | Odds Ratio       Suff.  2*Pr(Suff.)     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sexd60 |   5.403325           6      0.3047      .4320873    336.6659
        aids |    10.2224           3      0.1021      .7104864    589.9442
------------------------------------------------------------------------------
```

- Output provides estimated ORs, *p*-values for tests of null-hypoth. that true coeff. are zero (labeled "`2*Pr(Suff.)`"), and exact 95% CIs.
- In this example, results indicate non-significant results at the 5% level

Modifications of logistic regression for rare outcomes

Even in large samples, outcomes may be rare

- can cause estimation issues for the logistic model (hence the 10 outcomes per predictor rule of thumb),
- can lead to biased/infinite OR estimates (bootstrap not very helpful in these cases)

- When the exact logistic approach becomes computationally infeasible, the *penalized likelihood* version of the logistic model can be used to obtain more reliable inferences

- In Stata, the downloadable `firthlogit` package can be used to fit penalized logistic models
  - *Note*: does not accept the `i.` syntax for categorical predictors and will not work with some post-estimation commands like `margins`

Reference: Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med. 2002;21:2409-19.

Example: ICU data from homework #4

Problem: Sample of $n = 200$ with 40 outcomes. Model in question #3 involved 11 predictors:

```
. logistic sta i.crn i.typ i.can i.loc i.agecat i.syscat
note: 1.loc != 0 predicts success perfectly; 1.loc dropped and 5 obs not used
Logistic regression                               Number of obs   =        195
                                                  LR chi2(10)     =      62.06
                                                  Prob > chi2     =     0.0000
Log likelihood = -60.742305                       Pseudo R2       =     0.3381
------------------------------------------------------------------------------
         sta | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.crn |   2.840999   2.165964     1.37   0.171     .6375571    12.65969
       1.typ |   52.63315   69.20641     3.01   0.003     3.999717     692.611
       1.can |   13.62877   13.28441     2.68   0.007     2.017263     92.0769
             |
       1.loc |
         1   |          1  (empty)
         2   |   16.38254   15.17629     3.02   0.003     2.665963     100.672
             |
     agecat |
         2   |   4.240071   3.244213     1.89   0.059     .9464375    18.99566
         3   |   5.588847   4.478942     2.15   0.032     1.161885    26.88322
         4   |   13.20398   9.938337     3.43   0.001     3.020121    57.72781
             |
     syscat |
         2   |   .7691673   .4733606    -0.43   0.670     .2302368    2.569608
         3   |   1.059359   .6293668     0.10   0.923     .3306305    3.394246
         4   |   .0768537   .0754065    -2.62   0.009     .0112329     .525821
             |
       _cons |   .0011421   .0018182    -4.26   0.000     .0000504    .0258683
------------------------------------------------------------------------------
```

No inference about level 1 of loc

Example: ICU data from homework #4

Solution: Use `firthlogit`, after generating binary predictors for levels of `loc`, `agecat` & `syscat`:

```
. firthlogit sta crn typ can locc2 locc3 agec2 agec3 agec4 sysc2 sysc3 sysc4, or

initial:       penalized log likelihood = -91.758995
```

… (additional iteration output omitted)

```
Iteration 5:   penalized log likelihood = -56.752069

                                                Number of obs    =        200
                                                Wald chi2(11)    =      30.23
Penalized log likelihood = -56.752069           Prob > chi2      =     0.0015
------------------------------------------------------------------------------
         sta | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         crn |   2.585408   1.824617     1.35   0.178      .6483431    10.30987
         typ |   23.81719   26.04578     2.90   0.004      2.792809    203.1139
         can |   9.392172   8.038851     2.62   0.009      1.754747    50.27102
       locc2 |   1622.537   3584.976     3.35   0.001      21.35382      123286
       locc3 |   11.43597   9.426443     2.96   0.003      2.273238    57.53094
       agec2 |   3.620715   2.560872     1.82   0.069      .9052191    14.48222
       agec3 |   4.738263   3.503944     2.10   0.035      1.112131     20.1875
       agec4 |   10.26244     7.1259     3.35   0.001      2.631508    40.02182
       sysc2 |   .7950167   .4641699    -0.39   0.694      .2531645    2.496604
       sysc3 |   1.059447   .5989035     0.10   0.919      .3498615    3.208204
       sysc4 |   .1157282   .1005024    -2.48   0.013      .0210972     .634823
       _cons |   .0032221   .0043901    -4.21   0.000       .000223    .0465493
------------------------------------------------------------------------------
```

Estimated OR for level 1 of `loc`

Logistic regression for small samples/rare outcomes

Conclusion:

- exact logistic results preferred for small samples ($n < 200$), especially when standard approach fails to yield estimates
- penalized likelihood approach useful for problems with rare outcomes, and in cases where the exact approach is computationally infeasible
- bootstrap inference useful for small-moderate samples with adequate number of outcomes, and where exact methods are computationally infeasible
- conventional logistic regression (or the Poisson-based log relative risk model) preferred for moderate-large samples with adequate numbers of outcomes