

Logistic Regression

Model assessment & Prediction

March 14, 2017

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
Pseudo R2	=	0.1103

Log likelihood = -791.44399

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

- Models with product terms often introduce collinearity
 - 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

Pseudo R2 = 0.0301

Log likelihood = -858.93362

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 main arcus effect, and now the model coefficients are more interpretable.
- max. *VIF* reduced to 1.6 after centering
- **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:

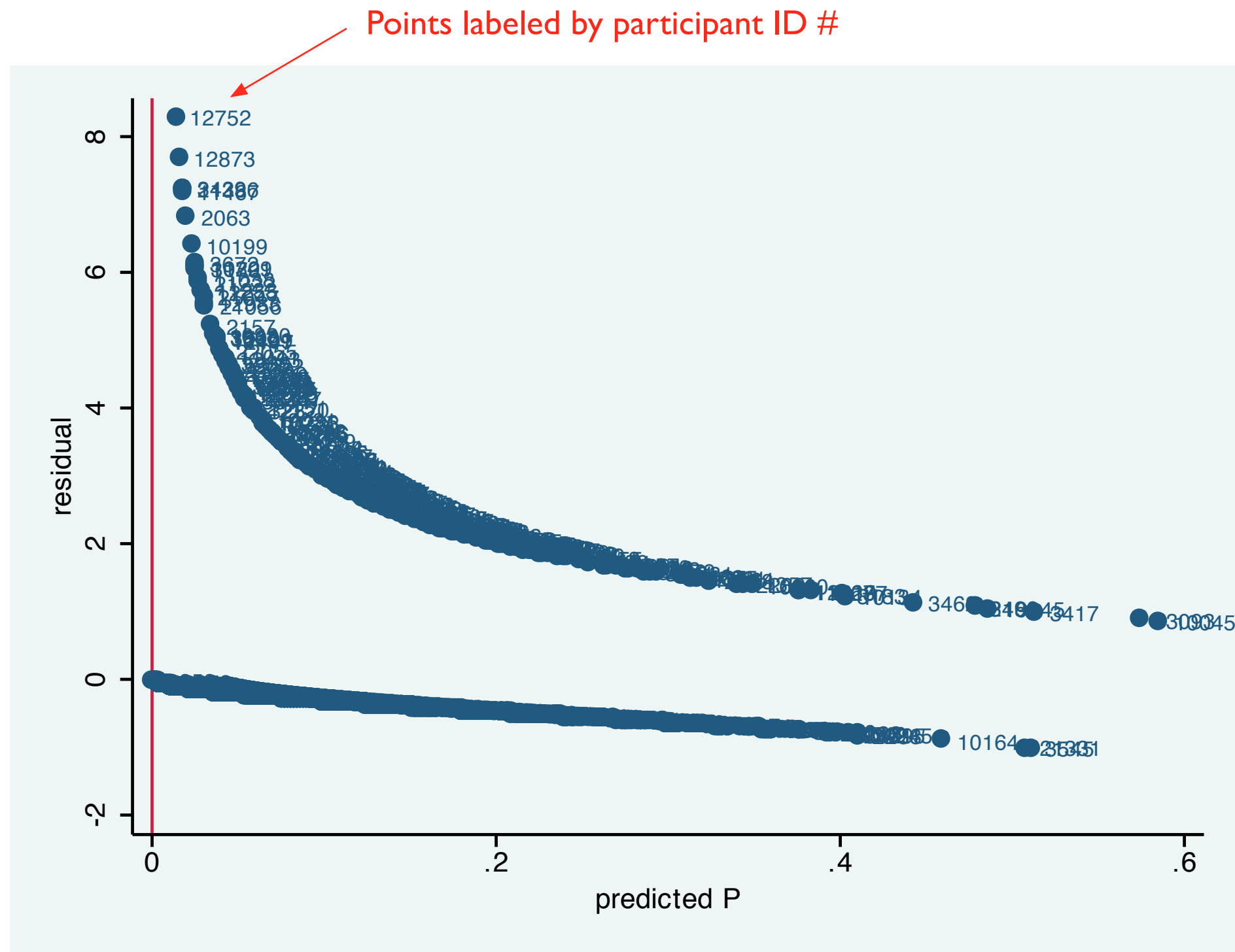
```
. list id chd69 age arcus dibpat chol r if _n<=10
```

	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

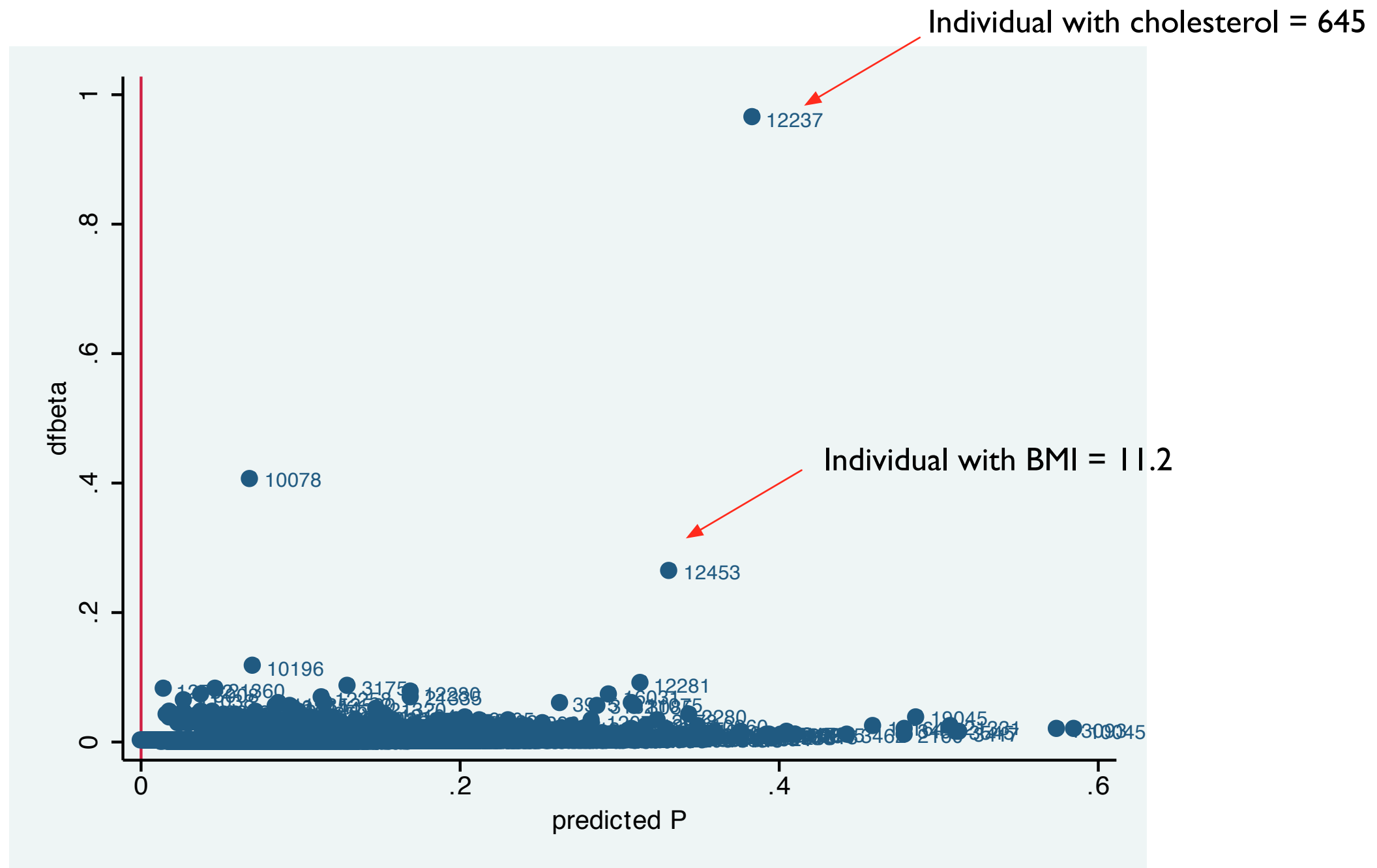
Residuals versus fitted values (probability scale):

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



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.

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 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

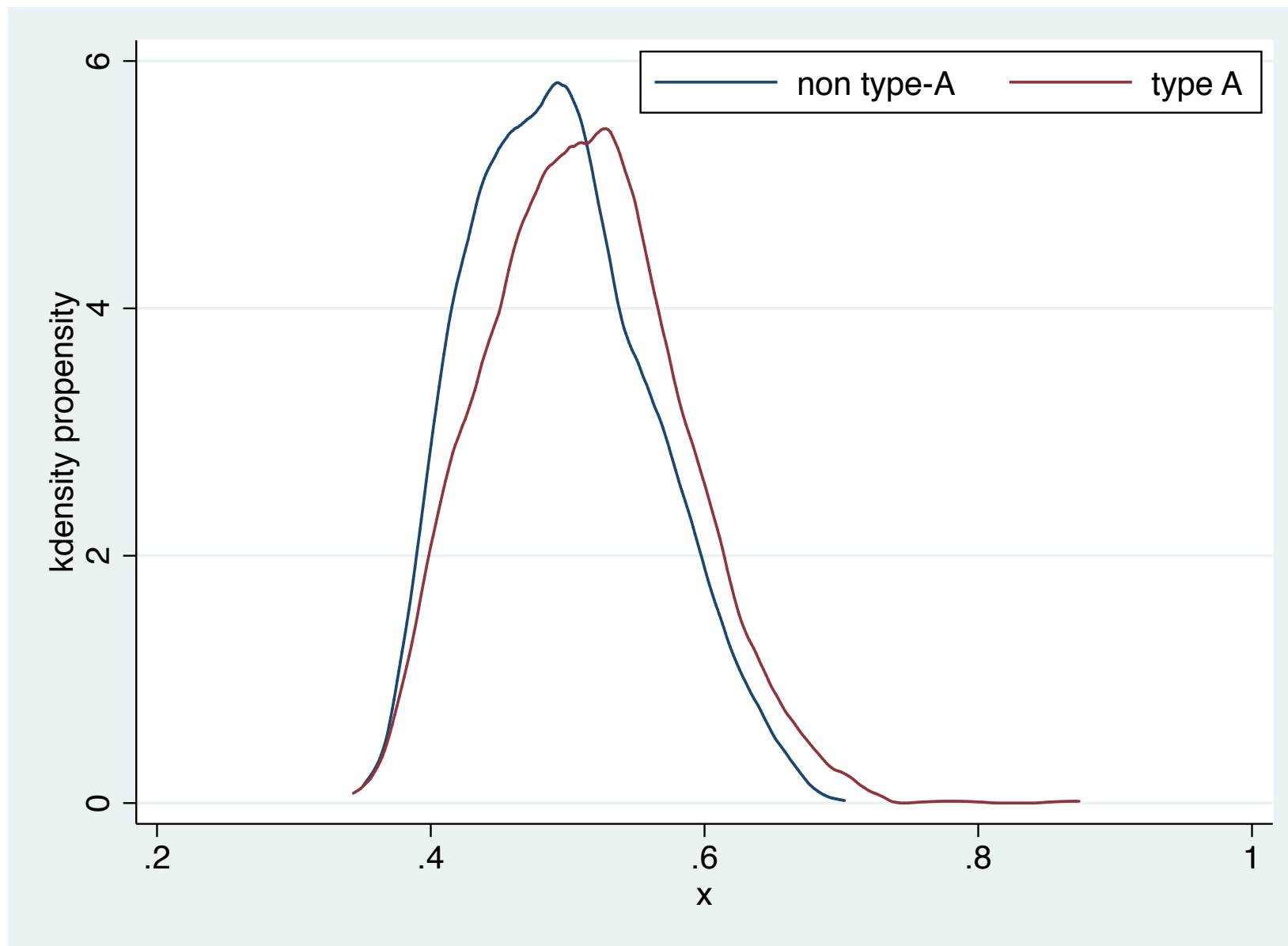
- Use of propensity scores (see lecture 6 & Sect. 9.4) from model for the log odds of reporting type A behavior (using the predictors in the model fitted in the last slide)
- Fit a model for the relationship between primary (binary) predictor and potential confounders, and use predicted probability to 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 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
```

-----+-----		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
chd69							
-----+-----							
_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`

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.
$$Pseudo R^2 = 1 - \left(\frac{LL_1}{LL_0} \right) \quad LL_1 : \text{fitted model}; LL_0 : \text{constant-only model}$$
- **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

Pseudo R2 = 0.1121

Log likelihood = -787.53196

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

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

expected # non-outcomes
in group (1)

(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

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

mean predicted prob. for group (1)

max predicted prob. for group (1)

```
. dis 314*.010283
```

3.228862 expected # outcomes in group (1)

```
. dis 314*(1-.010283)
```

310.77114 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

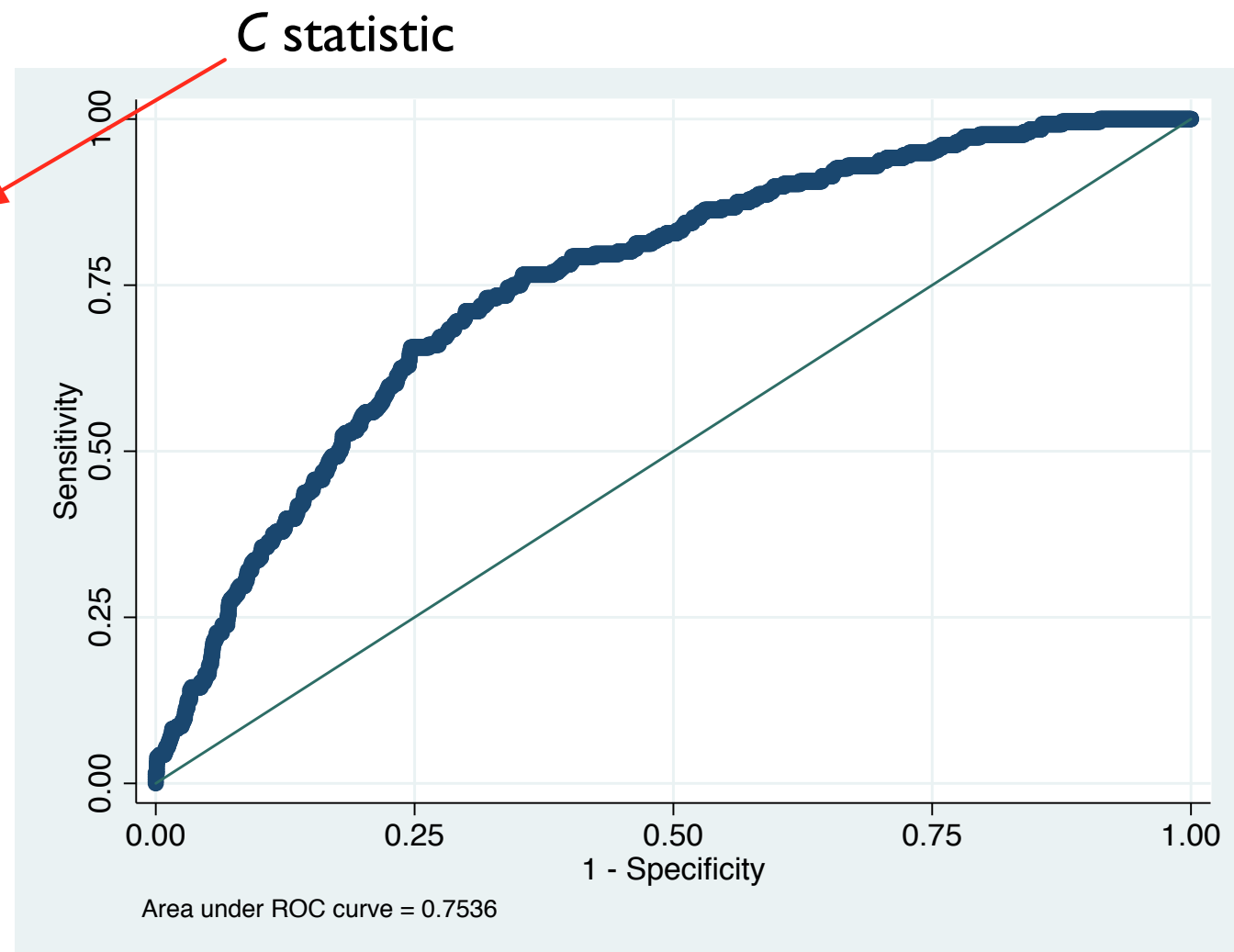
Discrimination:

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

```
. lroc
```

```
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:** 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

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 (see lecture 7 & Chapter 10)
 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 yield better classification performance than 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 effective in practice (take biostat 210 for more details)

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 (randomly selected) 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      =      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

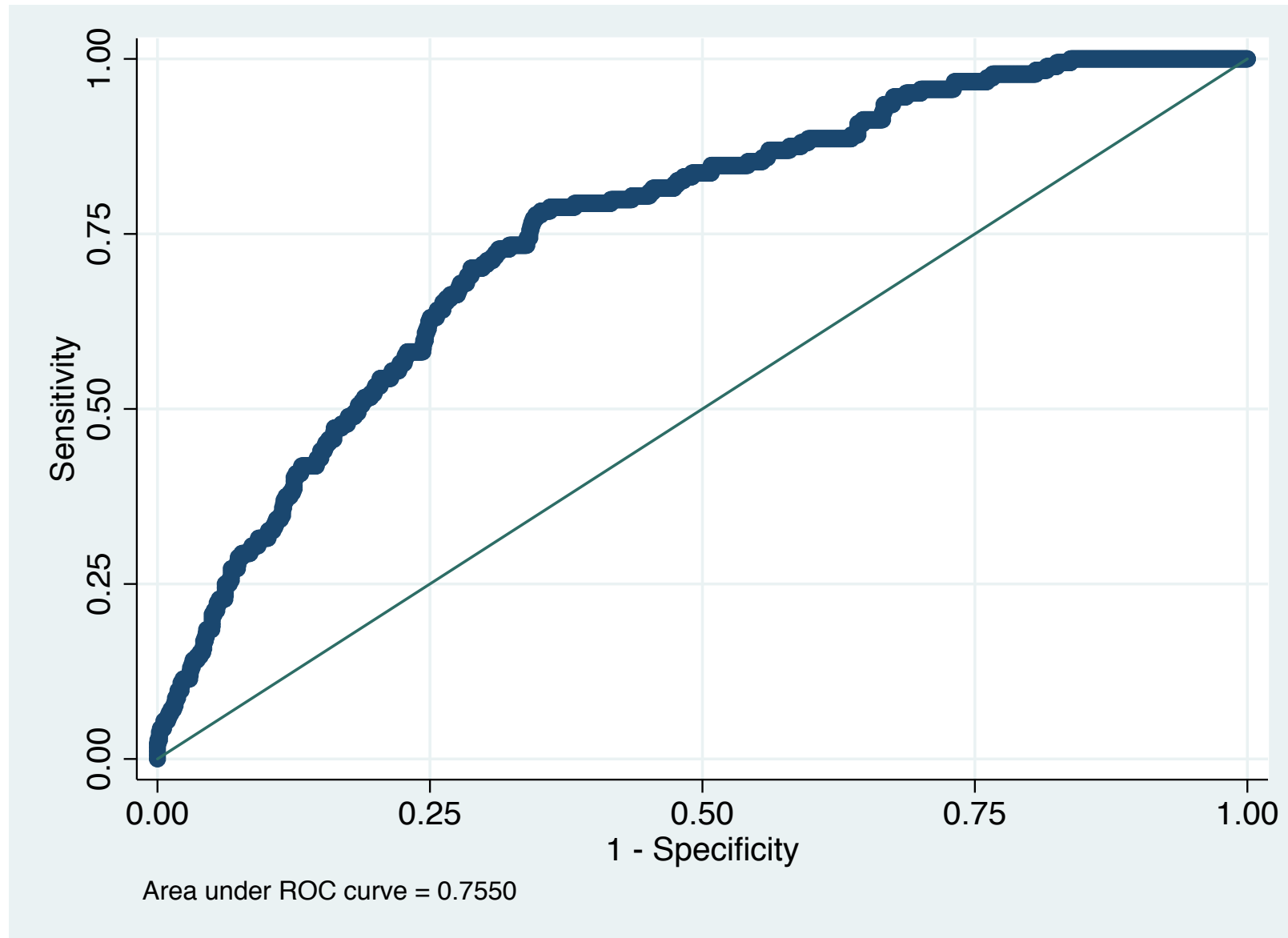
Formula for P :

$$P = \frac{\exp(\beta_0 + \beta_1 \text{age_10} + \beta_2 \text{chol_50} + \beta_3 \text{sbp_50} + \beta_4 \text{bmi_10} + \beta_5 \text{dibpat} + \beta_6 \text{smoke} + \beta_7 \text{bmichol} + \beta_8 \text{bmisbp})}{1 + \exp(\beta_0 + \beta_1 \text{age_10} + \beta_2 \text{chol_50} + \beta_3 \text{sbp_50} + \beta_4 \text{bmi_10} + \beta_5 \text{dibpat} + \beta_6 \text{smoke} + \beta_7 \text{bmichol} + \beta_8 \text{bmisbp})}$$

Prediction for binary outcomes

- Evaluate overall prediction performance in the group used to develop the model

```
. lroc if group==1
```

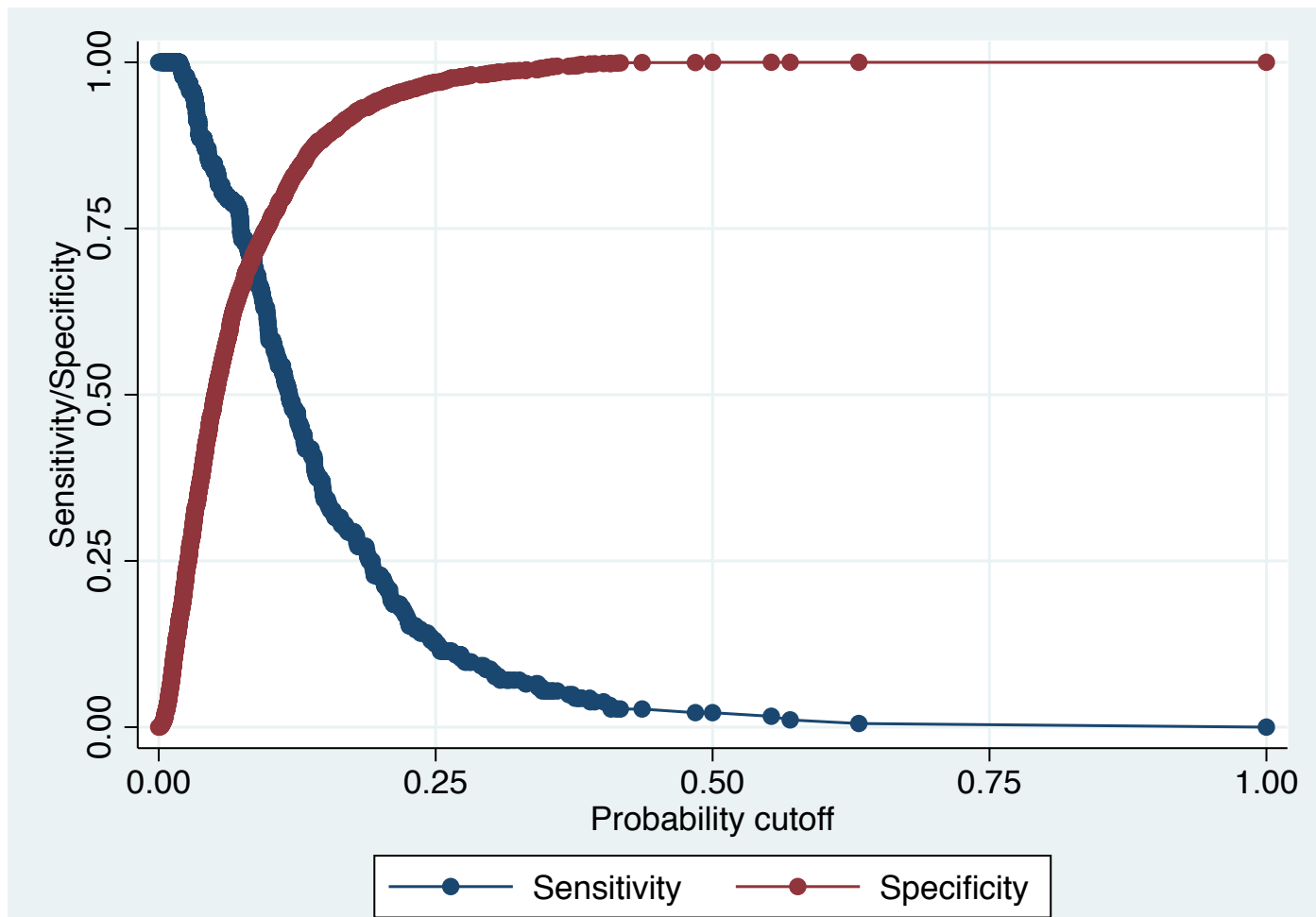


- Interpretation: C statistic = 0.7550 - represents (optimistic) classification performance in the development sample

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*)

Reference: 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

- Classification performance in development data for a particular cut-off for P :

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

Logistic model for chd69

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

Classified + if predicted $\text{Pr}(D) \geq .085$

True D defined as chd69 != 0

Sensitivity	$\text{Pr}(+ D)$	70.11%
Specificity	$\text{Pr}(- \sim D)$	70.72%
Positive predictive value	$\text{Pr}(D +)$	16.88%
Negative predictive value	$\text{Pr}(\sim D -)$	96.54%

False + rate for true ~D	$\text{Pr}(+ \sim D)$	29.28%
False - rate for true D	$\text{Pr}(- D)$	29.89%
False + rate for classified +	$\text{Pr}(\sim D +)$	83.12%
False - rate for classified -	$\text{Pr}(D -)$	3.46%

Prediction for binary outcomes

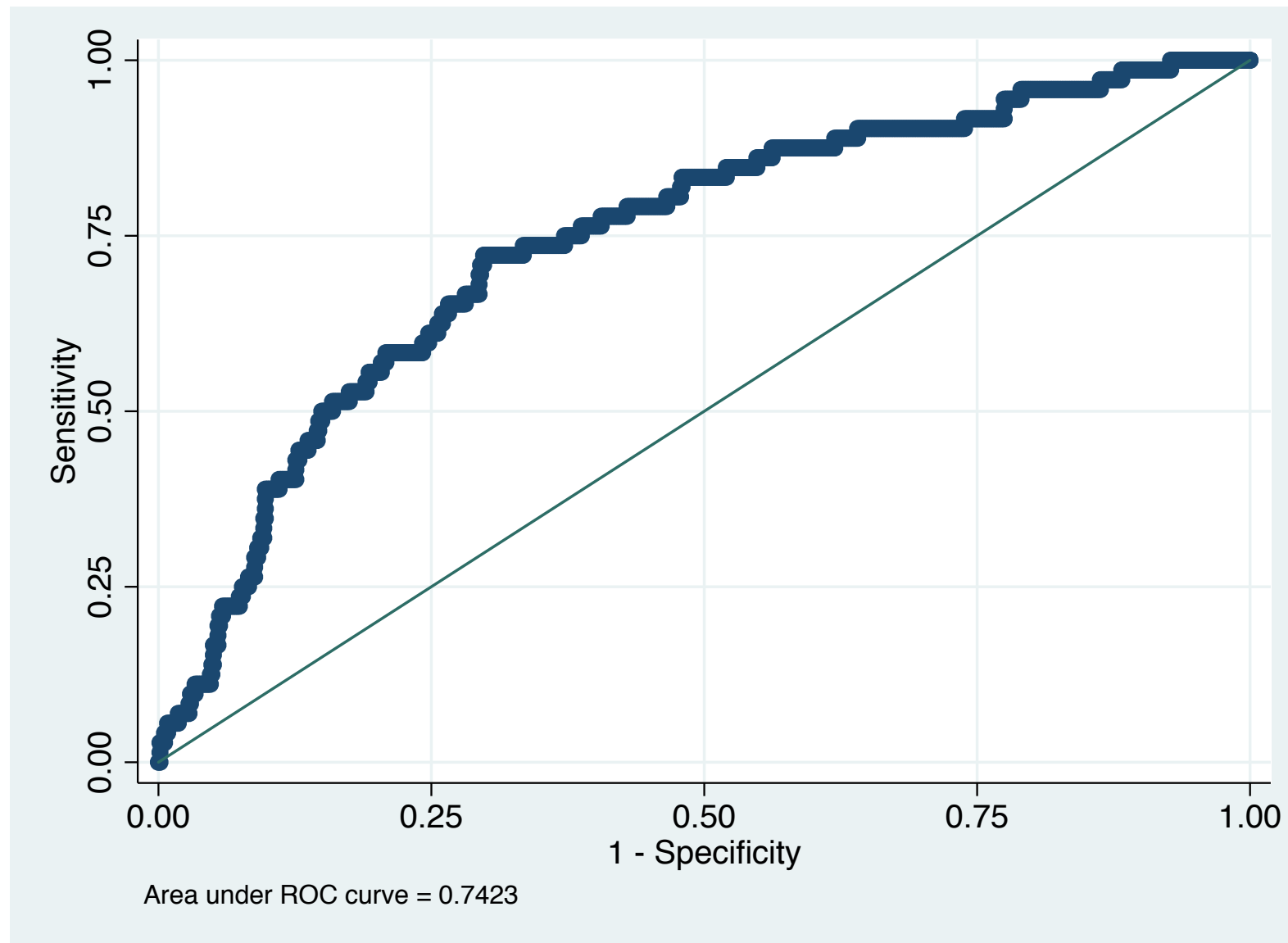
Split sample prediction example:

- **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 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

- 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

Split sample prediction example:

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

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

Logistic model for chd69

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

Classified + if predicted $\Pr(D) \geq .085$

True D defined as chd69 != 0

Sensitivity	$\Pr(+ D)$	72.22%
Specificity	$\Pr(- \sim D)$	70.07%
Positive predictive value	$\Pr(D +)$	19.55%
Negative predictive value	$\Pr(\sim D -)$	96.16%

False + rate for true ~D	$\Pr(+ \sim D)$	29.93%
False - rate for true D	$\Pr(- D)$	27.78%
False + rate for classified +	$\Pr(\sim D +)$	80.45%
False - rate for classified -	$\Pr(D -)$	3.84%

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 performance on left-out data
 - estimate optimism-corrected performance using cross-validated measure
- **Bootstrapping:**
 - in each of many (e.g. 500) bootstrap samples:
 - fit model to bootstrap sample, obtain 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
```

Obs	ROC Area	Std. Err.	-Asymptotic Normal-- [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 (i.e. optimism-corrected) area under ROC curve
. roctab chd69 cv_fitted
```

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

Prediction for binary outcomes

Example: bootstrap optimism for WCGS model

```
* 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 3140 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 3140 with-replacement
    .      bsample 3140 ,weight(bwt), in 1/3140
    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
.00697017
. dis 0.7536-O
.74662983
```

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"

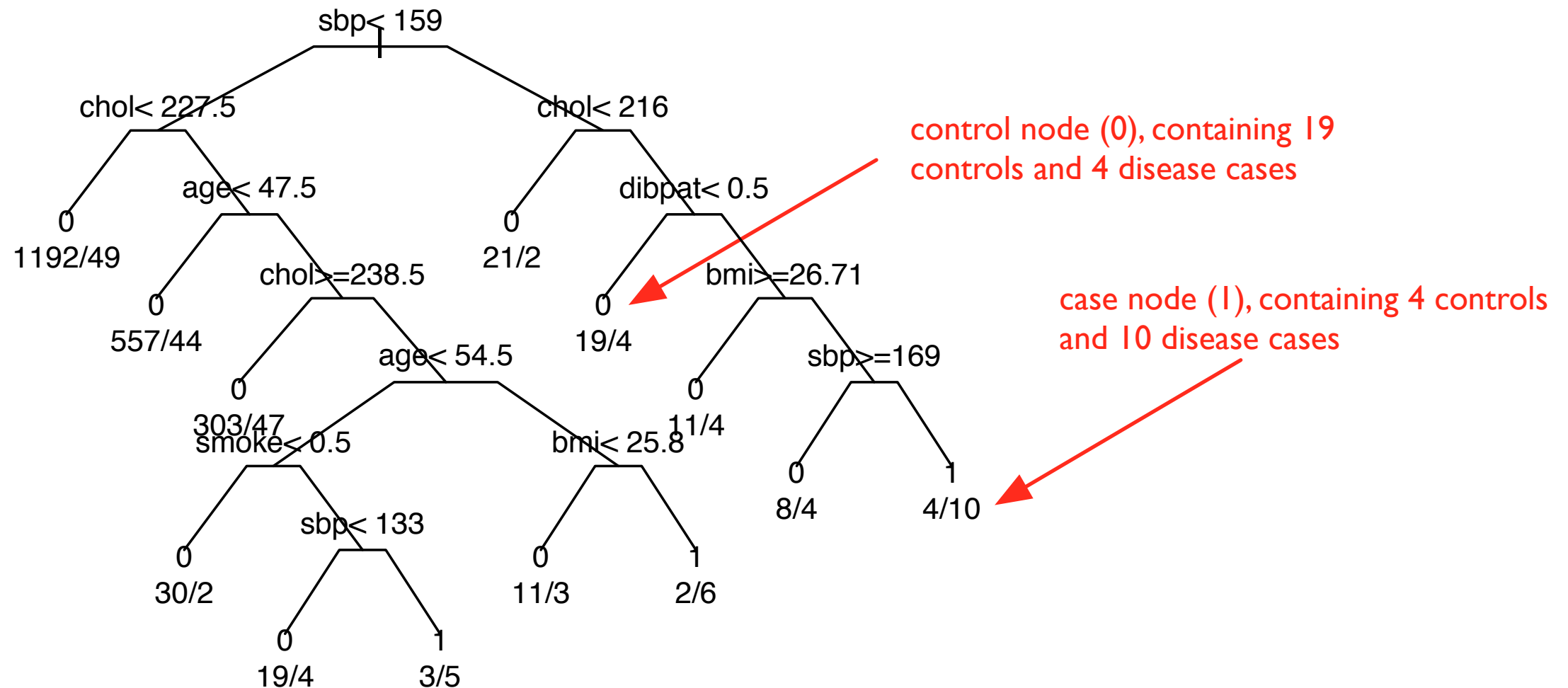
Advantages:

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

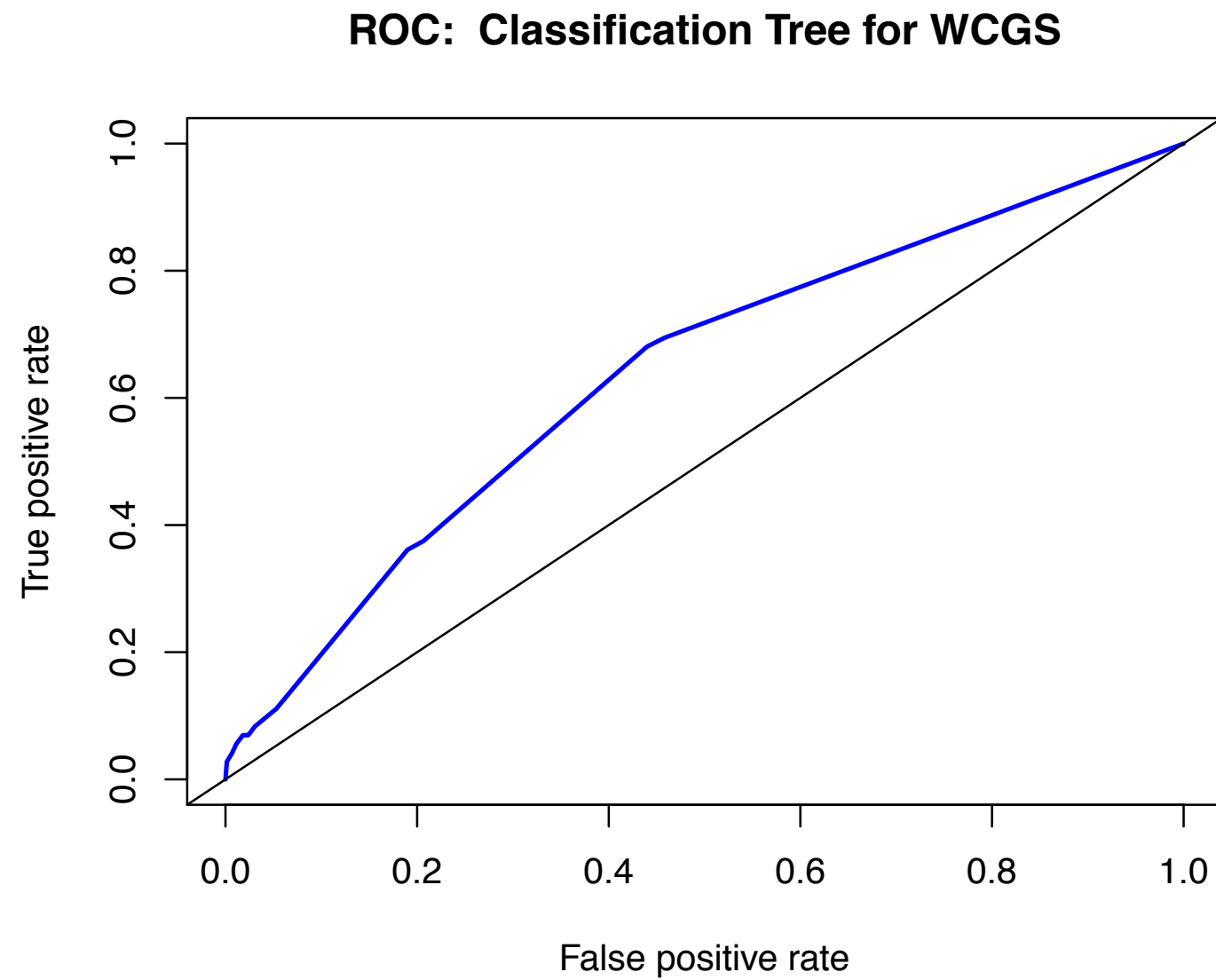
Prediction for binary outcomes

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)
- Default is to weight sensitivity and specificity as equally important



WCGS classification tree example: AUC = 0.65



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 causal models & DAGs)
 - change in coefficient criteria for logistic models affected by non-collapsibility of *ORs*
(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 can be used, but has limitations discussed previously,
 - Recently developed shrinkage techniques (e.g. LASSO regression) are useful

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
 - 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. Including such predictors results in decreased precision for effect estimates in logistic models.
 - can be important in randomized trials
 - methods based on estimating average causal effects of treatment (e.g. using inverse weighting via prop. scores or margins) may be preferred
- Logistic models for dependent outcomes will be covered in Biostat 209

Reference: 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.

Pooled logistic regression for grouped event time outcomes

- Application of the logistic regression model to survival (event time) outcomes
- For studies where observations of event times are limited to regular intervals (e.g. visit intervals in prospective studies)
- Data structured in “long” form, with one observation per person-visit, binary indicator of event occurrence (at most once / individual)
- Coefficients approximate log relative hazards from Cox proportional hazards regression when outcomes are rare
- Advantages over Cox regression:
 - Can be applied to when info about event times limited to visit intervals
 - Accepts weights - used to fit *marginal structural models* for causal inference accounting for time-dependent confounding variables
- Disadvantages:
 - requires explicit model for the time dependence of outcome risk
 - often requires modeling assumptions about event occurrences and predictor values within observation intervals

Pooled logistic regression for grouped event time outcomes

- Time dependence of outcome risk is modeled by the **intercept** part of the model:

$$\log\left(\frac{P}{1-P}\right) = \beta_0(t) + \beta_1 x_1 + \dots + \beta_p x_p$$

- $\beta_0(t)$ represents the log outcome odds at time t , among individuals with all predictors set to zero
 - No time dependence: constant $\beta_0(t) = \beta_0$,
 - including time (t) as an untransformed predictor equivalent to assuming linearity of log outcome odds with increasing time
 - flexible dependence can be modeled by a cubic spline for t
- The pooled logistic is a **proportional odds model**
 - Approximates the **proportional hazards model** for rare events
 - Used for control of predictors acting as time-dependent confounder/mediators in causal inference (via incorporation of inverse probability weights) (Biostat 215)

References:

1. D'Agostino RB, Lee ML, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. Stat Med. 1990;9:1501-15.
2. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology. 2000;11:561-70.

Pooled logistic regression for grouped event time outcomes

- **Example:** incident HSV-2 infection in the MIRA study

listing of data for two individuals (one with incident HSV-2):

```
. list id hsv2 mos agecat stihx newparts if id==28 | id==27
```

	id	hsv2	mos	agecat	stihx	newparts	
202.	27	0	3	1	0	0	
203.	27	0	6	1	0	0	
204.	27	0	9	1	0	0	
205.	27	0	12	1	0	0	
206.	27	0	15	1	0	0	
207.	27	0	18	1	0	0	
208.	27	0	21	1	0	0	
209.	27	0	24	1	0	0	
210.	28	0	3	1	1	0	
211.	28	0	6	1	1	0	
212.	28	0	9	1	1	0	
213.	28	0	12	1	1	1	
214.	28	0	15	1	1	0	
215.	28	1	18	1	1	0	

time varying predictor

individuals with the outcome stop contributing data after interval of outcome occurrence

mos: months on study, **hsv2:** indicator of incident HSV-2 infection,
agecat: age at baseline, **stihx:** baseline STI history,
newparts: recent sexual partners (Y/N)

Pooled logistic regression for grouped event time outcomes

- **Example:** incident HSV-2 infection in the MIRA study

- **generate spline basis to allow for potentially nonlinear effects of time**

- `mkspline spl = mos, cubic nknots(3)`

- `logistic hsv2 spl* i.agecat i.stihx newparts`

Logistic regression

Number of obs = 6069

LR chi2(6) = 33.69

Prob > chi2 = 0.0000

Log likelihood = -509.18109

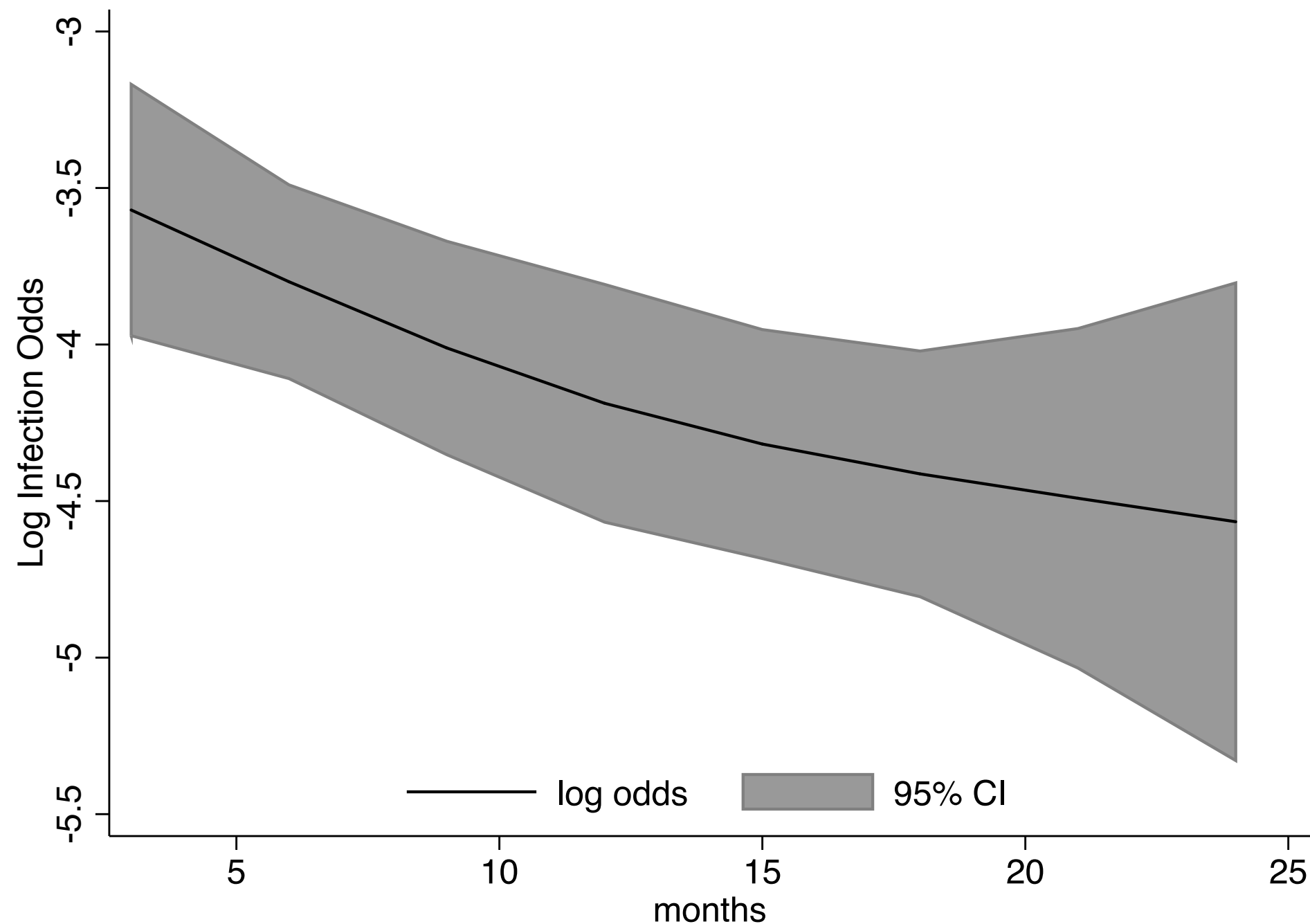
Pseudo R2 = 0.0320

hsv2	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
spl1	.9256021	.0379368	-1.89	0.059	.8541555	1.003025
spl2	1.035493	.0556198	0.65	0.516	.9320218	1.15045
agecat						
2	.6090384	.1369677	-2.20	0.027	.3919372	.946396
3	.5522162	.2213153	-1.48	0.138	.2517489	1.211297
1.stihx	1.92962	.4938946	2.57	0.010	1.168431	3.186695
newparts	1.826313	.4372661	2.52	0.012	1.142288	2.919945
_cons	.0354939	.0105938	-11.19	0.000	.0197741	.0637103

odds ratios for `spl*` predictors represent time variation (in months) in baseline HSV-2 risk; interest usually focuses on the other predictors in the model

Pooled logistic regression for grouped event time outcomes

- **Example:** Baseline log infection odds for HSV-2 example on previous slide (plotted using `xb1c` as in lab 8)



Linearity of log odds with months of follow-up appears reasonable

Cox regression for HSV-2 example

```
. stset mos hsv2, id(id)
. stcox i.agecat i.stihx newparts, efron
```

```
      failure _d:  hsv2
analysis time _t:  mos
           id:   id
```

Cox regression -- Efron method for ties

```
No. of subjects =          1000          Number of obs   =          6069
No. of failures =           104
Time at risk    =          18207

Log likelihood   =   -691.38351          LR chi2(4)       =          20.55
                                          Prob > chi2      =          0.0004
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
agecat						
2	.6131285	.1365595	-2.20	0.028	.3962479	.9487156
3	.5554717	.2209092	-1.48	0.139	.2547663	1.211105
1.stihx	1.918896	.4819664	2.59	0.009	1.172888	3.139399
newparts	1.793361	.4225623	2.48	0.013	1.130063	2.845987

- Results quite similar to pooled logistic regression results
- Modeling of the effect of time is *not required* for the Cox proportional hazards model (courtesy of the *proportional hazards assumption* and *partial likelihood*)
- Cox regression model doesn't accept weights (can't be used to control for time dependent confounder/mediators)