

## Lecture 17

1. Comparing alternative measurements of the same quantity
2. Sensitivity, specificity, false positive, false negative
3. Percent correctly predicted from logistic regression
4. ROC curve
5. Subset selection
6. Over-dispersion
7. Hosmer-Lemeshow lack-of-fit test
8. Conditional logistic regression

1

### Comparing alternative measurements of the same quantity

Two methods,  $A$  and  $B$ , for measuring characteristic  $G$ . Sample of cases where both  $A$  and  $B$  were applied (so ordered pair data).

1. Both  $A$  and  $B$  are assumed to measure  $G$  with error. Neither is “correct answer.”
  - $G$  is continuous: do  $A$  and  $B$  differ in mean (not follow  $x = y$  line) or variance or both?  
**concordance correlation coefficient**, Bland-Altman plots
  - $G$  is binary: do  $A$  and  $B$  agree?  
**kappa** = agreement beyond what would be expected by chance,  
**positive agreement, negative agreement**

Good review: Sanchez, Binkowitz: *Journal of Biopharmaceutical Statistics*, 9(3), 417–438 (1999)

2

2. Method *A* is the **gold standard**, regarded as “correct answer.”

- *G* is continuous: does *B* approximate *A* closely enough?

**Calibration problem**

- *G* is binary: does *B* correctly classify cases according to *A*?

**Classification or discrimination problem**

Logistic regression can classify:

0/1 response is assumed correct = gold standard.

Develop regression equation to predict response.

What % of responses correctly predicted?

3

What percent of responses correctly predicted?

		Observed Response	
		<i>1</i>	<i>0</i>
<i>Predicted 1</i>	<i>A</i>	<i>B</i>	
<i>Predicted 0</i>	<i>C</i>	<i>D</i>	

What percent correct would indicate a successful model?

Terminology from diagnostic testing for disease.

4

## Sensitivity and Specificity

		True Disease Status	
		<i>Disease +</i>	<i>Disease –</i>
<i>Diagnostic Test: Positive</i>	<i>A</i>	<i>B</i>	
<i>Negative</i>	<i>C</i>	<i>D</i>	

true positives =

false positives =

true negatives =

false negatives =

**Sensitivity** = If disease present, chance that test is positive =  $A/(A + C)$

**Specificity** = If no disease, chance that test is negative =  $D/(B + D)$

Want both sensitivity and specificity as close to 100% as possible

5

### Trade-off between sensitivity and specificity

How do we make our diagnostic test more sensitive?

		True Disease Status	
		<i>Disease +</i>	<i>Disease –</i>
<i>Diagnostic Test: Positive</i>	<i>A</i>	<i>B</i>	
<i>Negative</i>	<i>C</i>	<i>D</i>	

If we lower the threshold for positive test, we increase *A* and what else?

What is the effect on specificity =  $D/(B + D)$ ?

**Trade-off: increases in sensitivity reduce specificity**

6

## High blood pressure in NHANES data

National Center for Health Statistics posted a tutorial dataset, hypertension.xls, of 1019 NHANES observations for people over age 20. High blood pressure (hypertension) was documented either by high values for blood pressure or by prescribed medication to treat hypertension.

Model hypertension on obesity, adjusted for gender and age.

bmi\_class has 3 values: normal ( $18 \leq \text{BMI} \leq 25$ ), overweight ( $25 < \text{BMI} \leq 30$ ), and obese ( $30 < \text{BMI}$ ).

```
Proc Logistic descending data= NCHS ;
  class bmi_class;
  model hypertension = age  male  age*male  bmi_class;
```

7

### The LOGISTIC Procedure

Model Information		
Data Set	WORK.A	
Response Variable	hypertension	hypertension
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	1019
Number of Observations Used	952

### Response Profile

Ordered Value	hypertension	Total Frequency
1	1	319
2	0	633

Probability modeled is hypertension=1.

NOTE: 67 observations were deleted due to missing values for the response or ex

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	120.9322	<.0001
male	1	13.3272	0.0003
bmi_class	2	28.9211	<.0001
age*male	1	13.4863	0.0002

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.6453	0.4917	131.8290	<.0001
age	1	0.0914	0.00831	120.9322	<.0001
male	1	2.2425	0.6143	13.3272	0.0003
bmi_class 1 Obese	1	0.6313	0.1175	28.8729	<.0001
bmi_class 2 Overwt	1	-0.2912	0.1140	6.5283	0.0106
age*male	1	-0.0386	0.0105	13.4863	0.0002

Will we get any odds ratios?

9

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
bmi_class 1 Obese vs 3 Normal	2.642	1.745 3.998
bmi_class 2 Overwt vs 3 Normal	1.050	0.702 1.571

Why only these odds ratios?

Interpretation:

How do we get this comparison of predicted and observed hypertension?

		Observed Response	
		1	0
Predicted	1	A	B
	0	C	D

What is a predicted value from the logistic regression model?

When would we predict hypertension using the logistic regression model?

11

Prob Level	Correct		Incorrect		Correct	Percentages			
	Event	Non-Event	Event	Non-Event		Sensitivity	Specificity	False POS	False NEG
0.400	214	494	139	105	74.4	67.1	78.0	39.4	17.5
0.410	207	497	136	112	73.9	64.9	78.5	39.7	18.4
0.420	204	507	126	115	74.7	63.9	80.1	38.2	18.5
0.430	201	509	124	118	74.6	63.0	80.4	38.2	18.8
0.440	196	514	119	123	74.6	61.4	81.2	37.8	19.3
0.450	194	515	118	125	74.5	60.8	81.4	37.8	19.5
0.460	191	522	111	128	74.9	59.9	82.5	36.8	19.7
0.470	189	527	106	130	75.2	59.2	83.3	35.9	19.8
0.480	181	530	103	138	74.7	56.7	83.7	36.3	20.7
0.490	175	534	99	144	74.5	54.9	84.4	36.1	21.2
0.500	173	538	95	146	74.7	54.2	85.0	35.4	21.3
0.510	167	546	87	152	74.9	52.4	86.3	34.3	21.8
0.520	166	551	82	153	75.3	52.0	87.0	33.1	21.7
0.530	165	554	79	154	75.5	51.7	87.5	32.4	21.8
0.540	162	556	77	157	75.4	50.8	87.8	32.2	22.0
0.550	159	562	71	160	75.7	49.8	88.8	30.9	22.2
0.560	157	562	71	162	75.5	49.2	88.8	31.1	22.4
0.570	152	563	70	167	75.1	47.6	88.9	31.5	22.9
0.580	149	566	67	170	75.1	46.7	89.4	31.0	23.1
0.590	141	571	62	178	74.8	44.2	90.2	30.5	23.8
0.600	134	573	60	185	74.3	42.0	90.5	30.9	24.4

Trade-off most obvious in Incorrect columns.

12

Table produced by

```
Proc Logistic descending data= NCHS ;  
  class bmi_class;  
  model hypertension = age  male  age*male  bmi_class /  
    / Ctable  PProb = (.4 to .6 by .01);
```

CTABLE gives classification table at range of predicted probability PPROB

Default is PProb = (0 to 1 by .02) *(start to end by stepsize)*

13

The output below appears to be related, but is not.

Association of Predicted Probabilities and Observed Responses

Percent Concordant	82.2	Somers' D	0.647
Percent Discordant	17.5	Gamma	0.649
Percent Tied	0.3	Tau-a	0.289
Pairs	201927	c	0.824

This does not give percent correctly classified or “concordant” between predicted and observed.

14

Graphical summary of classification table:

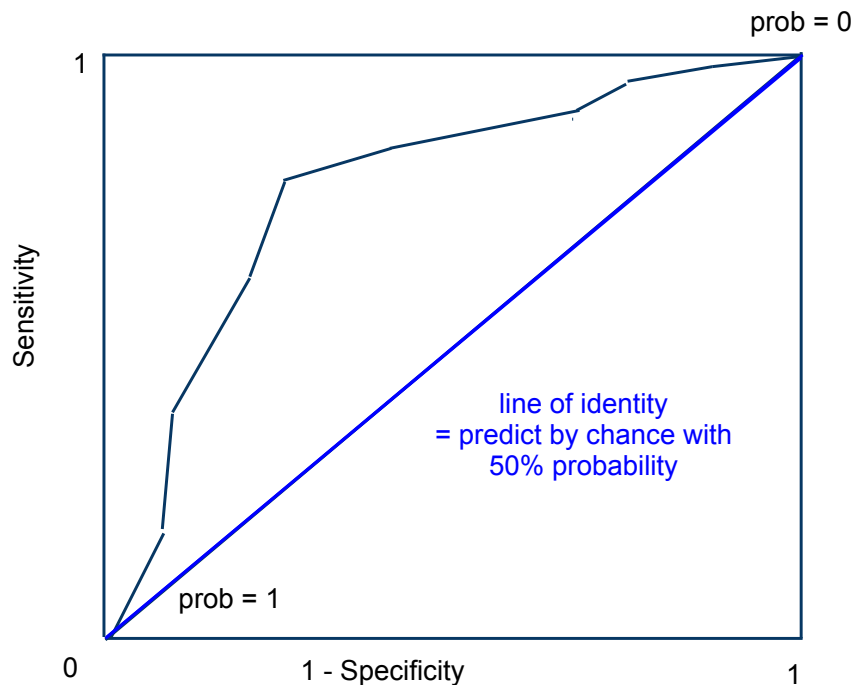
For each  $p$  from 0 to 1 in the classification table, plot sensitivity (vertical axis) against  $(1 - \text{specificity})$  (horizontal axis)

Classification Table

Prob Level	Correct		Incorrect		Correct	Percentages		False POS	False NEG
	Event	Non-Event	Event	Non-Event		Sensitivity	Specificity		
0.000	319	0	633	0	33.5	100.0	0.0	66.5	.
0.020	319	26	607	0	36.2	100.0	4.1	65.6	0.0
0.040	319	81	552	0	42.0	100.0	12.8	63.4	0.0
0.060	316	118	515	3	45.6	99.1	18.6	62.0	2.5
0.080	311	167	466	8	50.2	97.5	26.4	60.0	4.6
0.100	309	222	411	10	55.8	96.9	35.1	57.1	4.3
. . .									
0.880	12	631	2	307	67.5	3.8	99.7	14.3	32.7
0.900	11	632	1	308	67.5	3.4	99.8	8.3	32.8
0.920	4	632	1	315	66.8	1.3	99.8	20.0	33.3
0.940	0	633	0	319	66.5	0.0	100.0	.	33.5

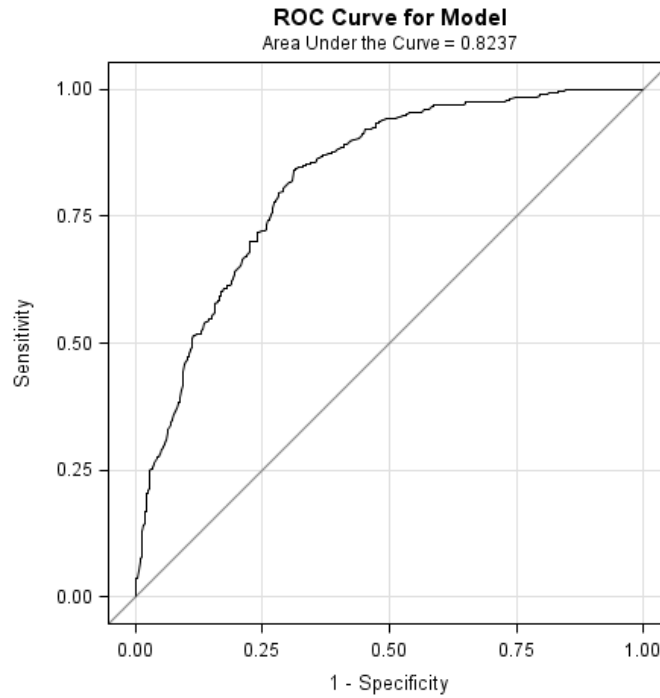
15

**ROC curve** (Receiver Operating Characteristic)



Want curve to approach upper left (0,1) corner.

16



Area under ROC curve (c in Concordance table) can be used to compare models. (Better than  $R^2$ ).

17

To get classification table and ROC curve, use ODS graphics:

```
ODS graphics on;
Proc Logistic descending data= NCHS ;
  class bmi_class;
  model hypertension = age male age*male bmi_class
    CTABLE PPROB =(.4 to .6 by .01) outroc=ROC;
run;
ODS graphics off;
```

Actually creates an output dataset which can be plotted in Proc Gplot, for more options.

18

## Automatic model reduction

Proc Logistic has options for forward and backward selection of subset models. It will also show the changes in the ROC curve.

```
ODS graphics on;
Proc Logistic descending data= NCHS ;
  class bmi_class;
  model hypertension = age male smoker bmi_class waist_circumf
    log_triglyceride pulse age*smoker age*bmi_class
    male*bmi_class age*male / selection=backward outroc=ROC;
run;
ODS graphics off;
```

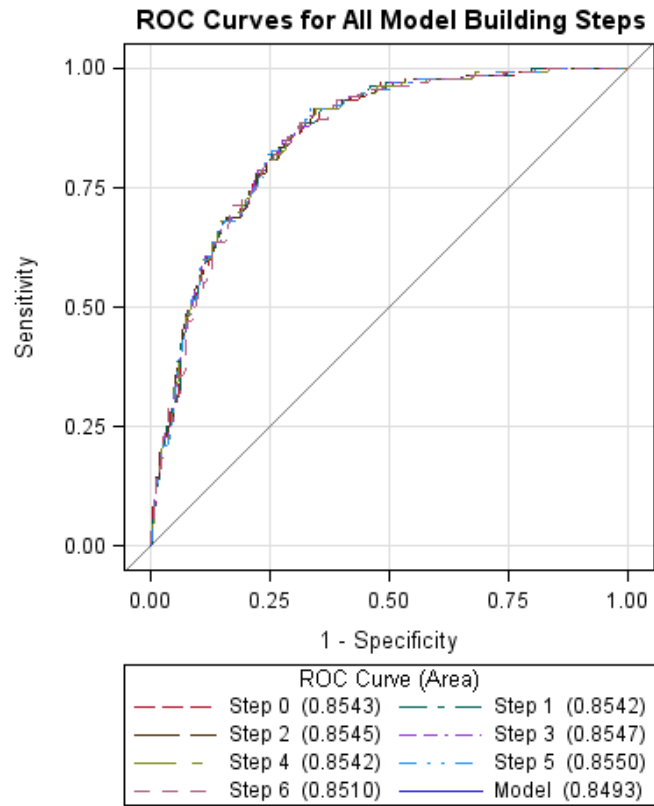
19

### Summary of Backward Elimination

Effect	Number	Wald	Variable
Step Removed	DF	In Chi-Square Pr > ChiSq	Label
1 waist_circumf	1	10 0.0278	0.8676 waist_circumf
2 age*bmi_class	2	9 0.3413	0.8431
3 log_triglyceride	1	8 0.0558	0.8133
4 age*smoker	1	7 0.2701	0.6033
5 smoker	1	6 0.1994	0.6552 smoker
6 male*bmi_class	2	5 3.1819	0.2037
7 pulse	1	4 1.4839	0.2232 pulse

The result is the first model we fitted.

20



21

Forward selection did very poorly:

No (additional) effects met the 0.05 significance level for entry into the model.

#### Summary of Forward Selection

Effect	Number	Score
Step Entered	DF	In Chi-Square Pr > ChiSq
1 age	1	124.4104 <.0001
2 bmi_class	2	9.5969 0.0082

Area under ROC curve  $c = .840$

#### Association of Predicted Probabilities and Observed Responses

Percent Concordant	83.8	Somers' D	0.679
Percent Discordant	15.9	Gamma	0.681
Percent Tied	0.3	Tau-a	0.287
Pairs	40260	<b>c</b>	<b>0.840</b>

22

## Over-dispersion

Recall that the variance of a Bernoulli (0/1) random variable  $Y$  is a function of its mean,  $\pi$ , the chance of an event:

$$\text{mean of } Y = \pi, \quad \text{variance of } Y = \pi(1 - \pi).$$

Observed binary data may have larger or smaller variance than prescribed: *over-dispersion* or *under-dispersion*, modelled by

$$\text{variance of } Y = \phi \cdot \pi(1 - \pi),$$

for **scale** factor  $\phi > 1$  or  $\phi < 1$ .

23

Procedure:

1. Estimate scale factor  $\phi$  from full model.
2. If  $\phi > 1.05$  (about) then rescale by  $\phi$ .

Generally, do not correct for **underdispersion** ( $\phi < 1$ ) because this will make SEs smaller.

Correcting for **overdispersion** will increase SEs.

There are 2 versions of  $\phi$ , with arguments for either choice.

Another issue is that the problem may be lack of fit not overdispersion.

24

Estimate scale factor  $\phi$  from full model:

```
Proc Logistic descending data= NCHS ;  
  class bmi_class;  
  model hypertension = age male smoker bmi_class waist_circumf  
    log_triglyceride pulse age*smoker age*bmi_class  
    male*bmi_class age*male / aggregate scale=N;
```

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	376.5747	422	0.8924	0.9452
Pearson	398.4633	422	0.9442	0.7889

Pearson and deviance-based estimates of  $\phi$  suggest under-dispersion

25

To rescale, change `scale=N` to `scale=P` (rescale by Pearson  $\phi$ )

or `scale=D` (rescale by deviance  $\phi$ )

or, if you plan to use a reduced model, use `scale = numeric value of  $\phi$`

Generally, do not correct for **underdispersion** because this will make SEs smaller.

Consider correcting for **overdispersion**  $\phi > (\text{approx})1.05$

26

## Lack of fit test: Hosmer-Lemeshow

Hosmer and Lemeshow (2000) proposed a test for lack of fit:

1. Divide the fitted probabilities into deciles (rank and divide into tenths).
2. Use the mean probability in a decile to calculate expected events.
3. Calculate chi-square test using observed and expected events, summed across deciles.

```
Proc Logistic descending data= NCHS ;  
  class bmi_class;  
  model hypertension = age male age*male bmi_class  
    / lackfit ; model option requests Hosmer-Lemeshow test
```

27

### Partition for the Hosmer and Lemeshow Test

Group	Total	hypertension = 1		hypertension = 0	
		Observed	Expected	Observed	Expected
1	96	0	2.68	96	93.32
2	98	8	6.71	90	91.29
3	96	6	10.61	90	85.39
4	95	17	15.88	78	79.12
5	96	19	22.08	77	73.92
6	95	45	30.93	50	64.07
7	96	43	41.08	53	54.92
8	92	51	50.31	41	41.69
9	96	55	63.37	41	32.63
10	92	75	75.35	17	16.65

### Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
18.8644	8	0.0156

28

## Matched-pair logistic regression

Clinical study: compare a new treatment for a skin disease against placebo.

Investigators enrolled 2 patients at each of 79 clinics and randomly assigned one to each treatment ( $T$  or  $P$ ). This is a **stratified randomization**, with clinic as stratum.

Covariates: baseline condition, age and gender.

At the end of the study, each patient was rated as improved or not improved.

Obs	center	treatment	gender	age	improve	baseline_ score
1	1	T	F	27	0	1
2	1	P	F	32	0	2
3	41	T	F	13	1	2
4	41	P	M	22	0	3
5	2	T	F	41	1	3

(Source: Ch 10 of Stokes, Davis, Koch: *Categorical Data Analysis Using the SAS System, 2nd ed.*)

29

This is a stratified randomization, with clinic as stratum.

If the response (improved) were continuous and normally distributed then we could try to fit:

```
Proc GLM;  
  class clinic treatment gender;  
  model improved = center baseline_score age gender treatment;
```

Issue: need to estimate 78 regression coefficients for the centers.

Only 2 observations per center to do this.

Worse with binary data:

difference of the responses within pairs is either  $-1$ ,  $0$ , or  $1$ .

diff	Frequency	Percent
-1	20	25.32
0	25	31.65
1	34	43.04

Where difference is zero, no information about treatment difference.

As a result, we cannot estimate 78 regression coefficients for the centers.

31

### Conditional logistic regression

Use only the pairs where the responses were different,  
don't estimate regression coefficients for strata.

```
Proc Logistic descending ;  
  class center treatment gender;  
  model improve = baseline_score treatment;  
  strata center;
```

The strata variable must be in the class statement.

32

The LOGISTIC Procedure  
 Conditional Analysis

Model Information

Data Set	WORK.STOKES_TRIAL
Response Variable	improve
Number of Response Levels	2
Number of Strata	79
Number of Uninformative Strata	25 <i>pairs with zero difference</i>
Frequency Uninformative	50
Model	binary logit
Optimization Technique	Newton-Raphson ridge

Number of Observations Read	158
Number of Observations Used	158

33

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
baseline_score	1	10.6106	0.0011
age	1	1.2253	0.2683
gender	1	0.9176	0.3381
treatment	1	3.8053	0.0511

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
baseline_score		1	1.0915	0.3351	10.6106	0.0011
age		1	0.0248	0.0224	1.2253	0.2683
gender	M	1	0.2656	0.2773	0.9176	0.3381
treatment	T	1	0.3512	0.1801	3.8053	0.0511

Center is not in the model, so no test for center differences.

### Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
baseline_score		2.979	1.545	5.745
age		1.025	0.981	1.071
gender	M vs F	1.701	0.574	5.043
treatment	T vs P	2.019	0.997	4.089

Adjusting for baseline condition, gender and age, those receiving the new treatment had twice the odds of improvement of those receiving placebo.

CLodds=PL not allowed with strata statement.