

## Lecture 17

1. Fitted probabilities from logistic regression
2. Hosmer-Lemeshow lack-of-fit test
3. Sensitivity, specificity, false positive, false negative
4. Percent correctly predicted from logistic regression
5. ROC curve

1

### Logistic regression example: hypertension 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 either by blood pressure or by prescribed medication to treat hypertension. 37% of sample had hypertension.

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

bmi\_class has 3 values:

normal ( $18 \leq \text{BMI} \leq 25$ ), overweight ( $25 < \text{BMI} \leq 30$ ), and obese ( $30 < \text{BMI}$ ).

2

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

3

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?

4

## Odds Ratio Estimates

Effect	Point	95% Wald	
	Estimate	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:

5

## Fitted probabilities from logistic regression

```
Proc Logistic descending data= NCHS ;  
  class bmi_class;  
  model hypertension = age male age*male bmi_class/  
  output out =B predicted =phat  
  lower =lowerCI upper =upperCI;
```

Logistic regression is on the log scale (log-odds),

so fitted probabilities  $\hat{p}_i$  are back-transformed from log scale.

Confidence interval instead of SE.

6

First 10 observations in output data B:

Obs	hypertension	phat	lowerCI	upperCI	age	male	bmi_ class
1	0	0.85626	0.77123	0.91325	85	0	3 Normal
2	0	0.61555	0.53666	0.68879	60	0	1 Obese
3	0	0.11866	0.08132	0.16997	32	1	2 Overwt
4	0	0.16304	0.11983	0.21798	39	1	2 Overwt
5	0	0.64705	0.56296	0.72292	64	1	1 Obese
6	1	.	.	.	74	1	
7	0	0.09279	0.06128	0.13813	40	0	2 Overwt
8	1	0.22907	0.17945	0.28760	47	1	2 Overwt
9	1	0.59168	0.49877	0.67846	69	0	2 Overwt
10	1	0.76614	0.68062	0.83432	75	1	1 Obese

Why is obs 6 missing  $\hat{p}_i$ ?

How good are the predictions?

7

### 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. Find the mean probability  $\bar{p}_i$  in each decile.
3. For each decile, calculate expected events as  $\bar{p}_i N_i$ , for  $i = 1, \dots, 10$ .
4. Calculate chi-square test using observed and expected events:

$$X^2 = \sum_{i=1}^{10} \frac{(\text{observed count} - \text{expected count})^2}{(\text{expected count})}$$

Null hypothesis: no lack of fit—expected count = observed count.

8

Request this test with the model option LACKFIT :

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

9

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

Conclusion?

### Another approach: correct predictions

		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 of responses correctly predicted?

Terminology from diagnostic testing for disease.

11

### 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

12

## 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:</i>	<i>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**

13

First 10 observations in output data B:

Obs	hypertension	phat	lowerCI	upperCI	age	male	bmi_ class
1	0	0.85626	0.77123	0.91325	85	0	3 Normal
2	0	0.61555	0.53666	0.68879	60	0	1 Obese
3	0	0.11866	0.08132	0.16997	32	1	2 Overwt
4	0	0.16304	0.11983	0.21798	39	1	2 Overwt
5	0	0.64705	0.56296	0.72292	64	1	1 Obese
6	1	.	.	.	74	1	
7	0	0.09279	0.06128	0.13813	40	0	2 Overwt
8	1	0.22907	0.17945	0.28760	47	1	2 Overwt
9	1	0.59168	0.49877	0.67846	69	0	2 Overwt
10	1	0.76614	0.68062	0.83432	75	1	1 Obese

When would we predict hypertension using the logistic regression model?

14

Try 3 different cut-off values: predict hypertension when  $\hat{p}_i > .45, .50, .55$

```
data C;
  set B;
  cutoff45 = (phat > 0.45);
  cutoff50 = (phat > 0.50);
  cutoff55 = (phat > 0.55);
  if phat=. then do;
    cutoff45=.; cutoff50=.; cutoff55=.; why is this needed?
  end;
```

```
Proc Freq data=C;
  tables (cutoff45 cutoff50 cutoff55)*hypertension / nopercnt ;
```

15

```
cutoff45
      hypertension(hypertension

Frequency|
Row Pct  |
Col Pct  |          0|          1| Total
-----+-----+-----+
          0 |    519 |    125 |    644
          |  80.59 |  19.41 |
          |  81.99 |  39.18 |
-----+-----+-----+
          1 |    114 |    194 |    308
          |  37.01 |  62.99 |
          |  18.01 |  60.82 |
-----+-----+-----+
Total          633      319      952
```

number correct?

sensitivity =

specificity =

16

cutoff50  
 hypertension(hypertension)

Frequency			
Row Pct			
Col Pct	0	1	Total
0	538	144	682
	78.89	21.11	
	84.99	45.14	
1	95	175	270
	35.19	64.81	
	15.01	54.86	
Total	633	319	952

number correct?

sensitivity =

specificity =

17

cutoff55  
 hypertension(hypertension)

Frequency			
Row Pct			
Col Pct	0	1	Total
0	562	159	721
	77.95	22.05	
	88.78	49.84	
1	71	160	231
	30.74	69.26	
	11.22	50.16	
Total	633	319	952

number correct?

sensitivity =

specificity =

18

<i>Cut-off</i>	<i>Correct</i>	<i>Sensitivity</i>	<i>Specificity</i>
$\hat{p}_i > .45$			
$\hat{p}_i > .50$			
$\hat{p}_i > .55$			

19

Proc Logistic will calculate these tables:

```
proc logistic descending data=a ;
  class bmi_class;
  model hypertension = age male bmi_class age*male /
    CTABLE PPROB =(.45 to .55 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)*

20

Classification Table

Prob Level	Correct		Incorrect		Correct	Percentages			
	Event	Non- Event	Event	Non- Event		Sensi- tivity	Speci- ficity	False POS	False NEG
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

21

Default Proc Logistic output below appears to be related to this question, but is not.

Association of Predicted Probabilities and Observed Responses

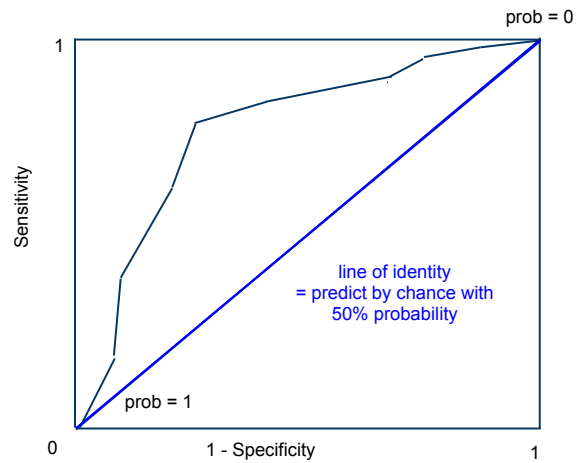
Percent Concordant	82.2	Somers' D	0.647
Percent Discordant	17.5	Gamma	0.649
Percent Tied	0.3	Tau-a	0.289
Pairs	201927	c	0.824

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

22

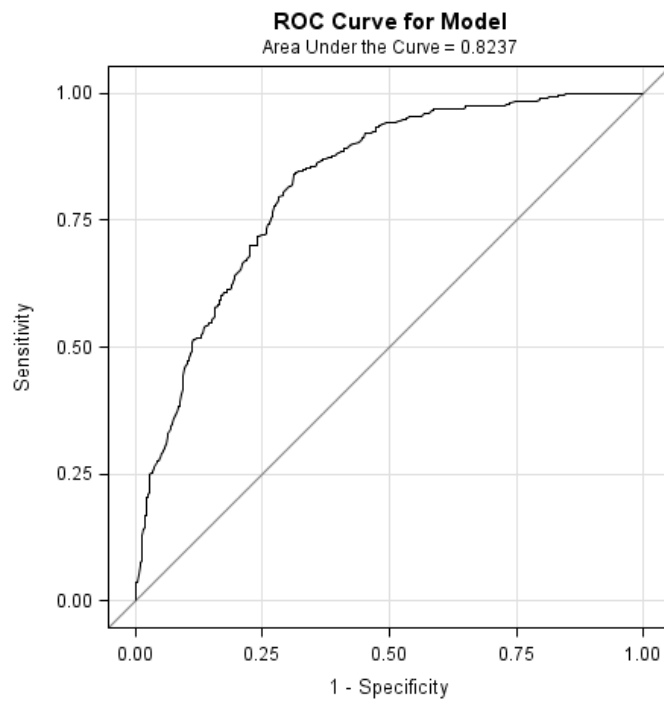
## ROC (Receiver Operating Characteristic) curve

For each  $\hat{p}$  cut-off from 0 to 1, plot sensitivity (vertical axis) against (1 – specificity) (horizontal axis)



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

23



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

24

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 /
  outroc=ROC;
run;
ODS graphics off;
```

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

25

### 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)

26

2. Method  $A$  is the **gold standard**, regarded as “truth.”

- $G$  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**

Treat gold standard classification  $A$  as the 0/1 response

Use logistic regression to predict response from method  $B$  measurement.

What % of responses correctly predicted?