Applied Logistic Regression

Week 7

1. Homework week 6: highlights
2. Assessing model calibration I
3. Assessing model calibration II
4. The Pearson chi-square statistic
5. The Hosmer-Lemeshow test
6. Statistical software for goodness of fit test I
7. Statistical software for goodness of fit test I
8. Assessing model discrimination (Area under the ROC curve) I
9. Assessing model discrimination (Area under the ROC curve) I
10. Homework

Stanley Lemeshow, Professor of Biostatistics

College of Public Health, The Ohio State University

THE OHIO STATE UNIVERSITY

Summary statistics may not be very specific about individual components

i.e.,

- a small value of one of these statistics does not rule out the possibility of some substantial deviation from fit for a few subjects.
- a large value for one of these statistics is a clear indication of a substantial problem with the model.

def: COVARIATE PATTERN - a single set of values for the covariates in a model

When developing models we assume that each subject is unique in their configuration of the covariates

- i.e., we assume # covariate patterns = *n*.

e.g.,

if AGE, RACE, SEX, WT were our variables, then the combination of these may well result in a unique set of values for each subject.

Once a final model is obtained there may be relatively few variables in the model, and the number of covariate patterns may be less than *n*.

e.g.,

if the final model contains only RACE and SEX, each coded at 2 levels, then there are only 4 possible covariate patterns.

The number of covariate patterns is not an issue in model development.

- The df for tests are based on the difference in the number of variables in competing models, not on the number of covariate patterns.
- They become an issue when assessing the fit of a model.

Suppose our fitted model contains *p* independent variables, $x_1, x_2, \ldots, x_p$. Let *J* denote the number of distinct values of $\underline{x}$ observed (i.e., covariate patterns).

- If some subjects have the same value of $\underline{x}$ then $J < n$.

Denote the number of subjects with $\underline{x} = \underline{x}_j$ by m_j, $j = 1,2,\ldots,J$.

Clearly, $\sum_{j=1}^{J} m_j = n$.

Let y_j denote the number of positive responses, $y = 1$, among the m_j subjects with $\underline{x} = \underline{x}_j$.

Then $\sum_{j=1}^{J} y_j = n_1 =$ total number of subjects with $y = 1$.

- The distribution of the goodness-of-fit statistics is obtained by letting *n* get large
- If *J*, the number of covariate patterns, also increases with *n*, then each value of m_j will tend to be small.
- Distributional results obtained under the condition that only *n* becomes large are said to be based on“n-asymptotics”.

If we fix $J < n$ and let n become large, then each value of m_j will tend to become large.

- Distributional results based on each m_j becoming large are said to be based on "m – asymptotics".

Initially we will assume that $J \approx n$ as in the case most frequently occurring.

- We expect this to be the case whenever we have some continuous covariates in the model.

Let us now review several of the available methods.

Let $\hat{\pi}_i = \dfrac{e^{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j}}{1 + e^{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j}}$ be computed for all individuals, $1 = 1, \ldots, n$.

Given the values $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n$, an informally used approach is to rank order these n values and establish“deciles of risk”.

i.e.,

1st decile contains the smallest $n/10$ values of $\hat{\pi}_i$

2nd decile contains the next smallest $n/10$ values of $\hat{\pi}_i$

$\vdots$

10th decile contains the largest $n/10$ values of $\hat{\pi}_i$

If $n/10$ is not an integer, then the 10 groups may have slightly different numbers.

Now, if the model holds then those who actually develop the outcome should have high values for $\hat{\pi}_i$. Similarly, those who don't develop the outcome should have low values for $\hat{\pi}_i$.

Procedures have been developed for comparing the observed number with the expected number in each decile.

i.e., for the j^{th} decile

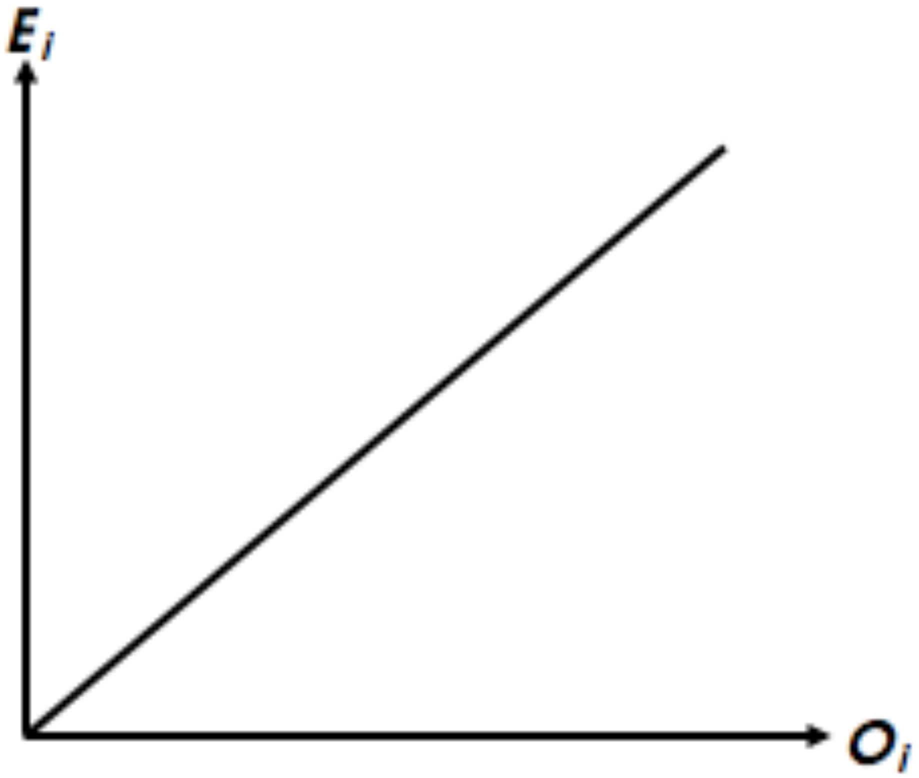
$$O_j = \sum_{i \in D_j} y_i$$

$$E_j = \sum_{i \in D_j} \hat{\pi}_i$$

where $j = 1, \ldots, 10$ and where D_j denotes the $n/10$ individuals in the j^{th} decile of risk.

Consider the pairs $(O_1, E_1), \ldots, (O_i, E_i), \ldots, (O_{10}, E_{10})$.

One method used has been to plot these pairs:



If the observed and expected correspond, then the 10 points should fall on a line with slope = 1, intercept = 0.

This is an eye-ball method as there is no test statistic associated with it.

Pearson Chi-Square Statistic

In linear regression we were concerned with residuals of the form

$$y_i - \hat{y}_i$$

In logistic regression fitted values are calculated for each covariate pattern, and depend on the estimated probability for that covariate pattern.

We denote the fitted value, $\hat{y}_i$, as

$$m_j \hat{\pi}_j = m_j \left\{ \frac{e^{\hat{g}(\underline{x}_j)}}{1 + e^{\hat{g}(\underline{x}_j)}} \right\}$$

where $\hat{g}\left(\underline{x}_j\right)$ is the estimated logit.

For a particular covariate pattern the Pearson residual is defined as

$$r\left(y_j\hat{\pi}_j\right) = \frac{\left(y_j - m_j\hat{\pi}_j\right)}{\sqrt{m_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)}}$$

The summary statistic based on these residuals is the Pearson chi-square statistic

$$X^2 = \sum_{j=1}^{J} r\left(y_j,\hat{\pi}_j\right)^2$$

and $X^2 \sim \chi^2\left(J-\left(p+1\right)\right)$ if the model holds.

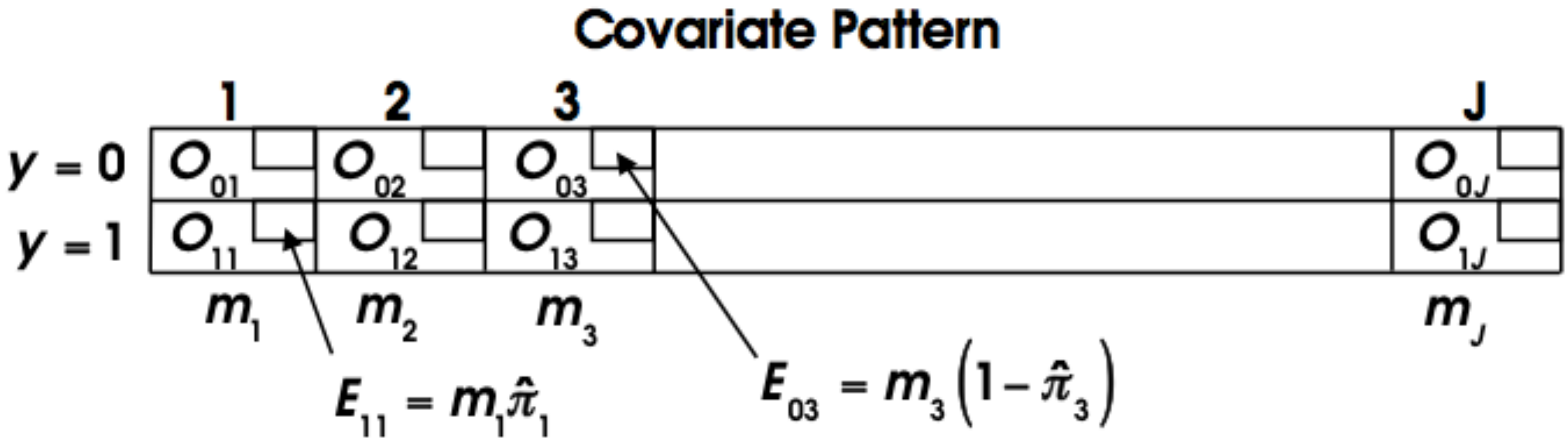
Problem: when $J \approx n$, the distribution is obtained under n - asymptotics and hence the number of parameters is increasing at the same rate as the sample size.

Hence, p - values calculated for χ^2 are incorrect when $J \approx n$.

Although the p-value may be slightly off, X^2 is an effective way to compare observed to expected frequencies for each covariate pattern.

This statistic is routinely produced by many software packages.

The Pearson Chi Square Statistic can be thought of as arising from the following $2 \times J$ table:



Covariate Pattern

	1	2	3		J
$y = 0$	O_{01}	O_{02}	O_{03}		O_{0J}
$y = 1$	O_{11}	O_{12}	O_{13}		O_{1J}
	m_1	m_2	m_3		m_J

$E_{11} = m_1\hat{\pi}_1$

$E_{03} = m_3\left(1 - \hat{\pi}_3\right)$

When chi-square tests are computed from a contingency table the *p*-values are correct under the hypothesis when the estimated expected values are “large” in each cell.

- This condition will hold under *m*-asymptotics.

In the previous table the expected values will always be quite small since the number of columns, *J*, ↑ as *n* ↑.

One way to avoid these difficulties under *n*-asymptotics is to group the data in such a way that *m*-asymptotics can be used.

e.g., we may collapse the columns into a fixed number of groups, *g*, and then calculate the observed and expected frequencies.

By fixing the number of columns, the estimated expected frequencies will become large as *n* becomes large.

- Thus *m*-asymptotics hold.

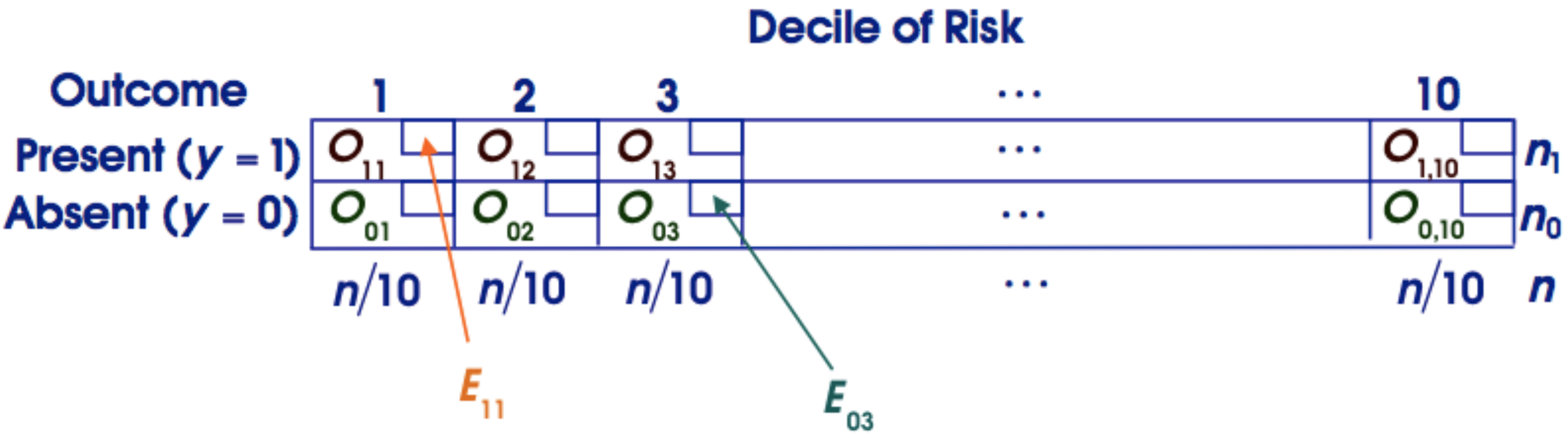
The Hosmer-Lemeshow Tests

Let us suppose that $J = n$. Two grouping strategies are proposed:

(1) Collapse the table based on percentiles of the estimated probabilities.

(2) Collapse the table based on fixed values of the estimated probabilities.

With method (1), use of $g = 10$ groups results in the first group containing the $n_1' = n/10$ subjects having the smallest estimated probabilities, and the last group containing the $n_{10}' = n/10$ subjects having the largest estimated probabilities.



	Decile of Risk					
Outcome	1	2	3	$\cdots$	10	
Present ($y = 1$)	O_{11}	O_{12}	O_{13}	$\cdots$	$O_{1,10}$	n_1
Absent ($y = 0$)	O_{01}	O_{02}	O_{03}	$\cdots$	$O_{0,10}$	n_0
	$n/10$	$n/10$	$n/10$	$\cdots$	$n/10$	n

Where

$$O_{1j} = \sum_{i \in D_j} y_i \qquad O_{0j} = \sum_{i \in D_j} \left(1 - y_i\right)$$

$$E_{1j} = \sum_{i \in D_j} \hat{\pi}_i \qquad E_{0j} = \sum_{i \in D_j} \left(1 - \hat{\pi}_i\right)$$

Then we compute

$$\hat{C} = \sum_{k=0}^{1} \sum_{j=1}^{10} \frac{\left(O_{kj} - E_{kj}\right)^2}{E_{kj}}$$

This is the Pearson chi-square statistic from the 2×g table of observed and expected frequencies.

If the 2nd grouping strategy is used, $g = 10$ groups results in cutpoints defined at the values $k/10,\ k = 1,2,\ldots,9$ and the groups contain all subjects with estimated probabilities between adjacent cutpoints.

e.g., 1st group = $0 = \hat{\pi}_i < .1$

2nd group = $.1 = \hat{\pi}_i < .2$

$\vdots$

10th group = $.9 = \hat{\pi}_i < 1.0$

Based on extensive simulations, it has been demonstrated that, when $J = n$ and the fitted logistic model is the correct model, the distribution of $\hat{C}$ is well approximated by $\chi^2(g-2)$.

The grouping method based on deciles of risk is preferable to the one based on fixed cutpoints in the sense of better adherence to the $\chi^2(g-2)$ distribution

- this is especially true when many of the estimated probabilities are small $(\text{i.e.}, <.02)$.

Assessing the fit of the model for the low birth weight data follows.

SYSTAT LOGIT

```
>model low=constant+lwt+race
>dc # smart=10
===============
DECILES OF RISK
===============

RECORDS PROCESSED: 189
SUM OF WEIGHTS = 189.00000

                             STATISTIC      P-VALUE          DOF
----------------------------------------------------------------
HOSMER-LEMESHOW         |      7.04419      0.53187      8.00000
PEARSON                 |    188.30343      0.41865    185.00000
DEVIANCE                |    223.25909      0.02869    185.00000
----------------------------------------------------------------

CAT.            0.16785      0.22266      0.25301      0.27064      0.29538
---------------------------------------------------------------------------
RESP OBS |      2.00000      4.00000      5.00000      4.00000      5.00000
     EXP |      2.19847      3.52513      4.66830      4.72146      4.48439
REF  OBS |     16.00000     14.00000     15.00000     14.00000     11.00000
     EXP |     15.80153     14.47487     15.33170     13.27854     11.51561
---------------------------------------------------------------------------
AV. PROB.       0.12214      0.19584      0.23341      0.26230      0.28027

CAT.            0.33324      0.36796      0.40774      0.47690      1.00000
---------------------------------------------------------------------------
RESP OBS |      7.00000      6.00000      4.00000     12.00000     10.00000
     EXP |      6.58226      6.28696      8.38347      8.30865      9.84091
REF  OBS |     14.00000     12.00000     18.00000      7.00000      9.00000
     EXP |     14.41774     11.71304     13.61653     10.69135      9.15909
---------------------------------------------------------------------------
AV. PROB.       0.31344      0.34928      0.38107      0.43730      0.51794
```

```
 >dc

===============
DECILES OF RISK
===============

RECORDS PROCESSED: 189
SUM OF WEIGHTS = 189.00000

                             STATISTIC      P-VALUE          DOF
----------------------------------------------------------------
HOSMER-LEMESHOW         |      2.34774      0.67209      4.00000
PEARSON                 |    188.30343      0.41865    185.00000
DEVIANCE                |    223.25909      0.02869    185.00000
----------------------------------------------------------------

CAT.            0.10000      0.20000      0.30000      0.40000      0.50000
---------------------------------------------------------------------------
RESP OBS |      0.00000      4.00000     19.00000     14.00000     16.00000
     EXP |      0.30527      3.51225     17.55645     17.45570     13.74300
REF  OBS |      4.00000     19.00000     50.00000     36.00000     15.00000
     EXP |      3.69473     19.48775     51.44355     32.54430     17.25700
---------------------------------------------------------------------------
AV. PROB.       0.07632      0.15271      0.25444      0.34911      0.44332

CAT.            0.60000      0.70000      0.80000      0.90000      1.00000
---------------------------------------------------------------------------
RESP OBS |      6.00000      0.00000      0.00000      0.00000      0.00000
     EXP |      6.42734      0.00000      0.00000      0.00000      0.00000
REF  OBS |      6.00000      0.00000      0.00000      0.00000      0.00000
     EXP |      5.57266      0.00000      0.00000      0.00000      0.00000
---------------------------------------------------------------------------
AV. PROB.       0.53561      0.00000      0.00000      0.00000      0.00000
 >quit

STOP
```

STATA

```
. xi:logit low lwt i.race
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood =  -111.7491
Iteration 2:   log likelihood = -111.62983
Iteration 3:   log likelihood = -111.62955

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(3)      =      11.41
                                                  Prob > chi2     =     0.0097
Log likelihood = -111.62955                       Pseudo R2       =     0.0486

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |  -.0152231   .0064393    -2.36   0.018    -.0278439   -.0026023
    _Irace_2 |   1.081066   .4880512     2.22   0.027     .1245034    2.037629
    _Irace_3 |   .4806033   .3566733     1.35   0.178    -.2184636     1.17967
       _cons |   .8057535   .8451625     0.95   0.340    -.8507345    2.462241
------------------------------------------------------------------------------
```

```
. lfit, group(10) table

Logistic model for low, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.1681 |     2 |   2.4 |    17 |  16.6 |    19 |
  |     2 | 0.2228 |     4 |   4.2 |    17 |  16.8 |    21 |
  |     3 | 0.2531 |     5 |   4.0 |    12 |  13.0 |    17 |
  |     4 | 0.2708 |     4 |   5.0 |    15 |  14.0 |    19 |
  |     5 | 0.2955 |     8 |   5.4 |    11 |  13.6 |    19 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.3334 |     6 |   6.1 |    13 |  12.9 |    19 |
  |     7 | 0.3681 |     6 |   8.2 |    17 |  14.8 |    23 |
  |     8 | 0.4078 |     3 |   5.8 |    12 |   9.2 |    15 |
  |     9 | 0.4770 |    12 |   8.9 |     8 |  11.1 |    20 |
  |    10 | 0.5975 |     9 |   8.9 |     8 |   8.1 |    17 |
  +--------------------------------------------------------+

       number of observations =       189
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =         7.61
                  Prob > chi2 =         0.4728

. lfit

Logistic model for low, goodness-of-fit test

       number of observations =       189
 number of covariate patterns =       109
            Pearson chi2(105) =       111.22
                  Prob > chi2 =         0.3204
```

```
. logit STA AGE CAN _ISYSGP_4 TYP LOCD

Iteration 0:   log likelihood = -100.08048
Iteration 1:   log likelihood = -70.385527
Iteration 2:   log likelihood = -67.395341
Iteration 3:   log likelihood = -66.763511
Iteration 4:   log likelihood = -66.758491
Iteration 5:   log likelihood = -66.758489

Logistic regression                               Number of obs   =        200
                                                  LR chi2(5)      =      66.64
                                                  Prob > chi2     =     0.0000
Log likelihood = -66.758489                       Pseudo R2       =     0.3330

------------------------------------------------------------------------------
         STA |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         AGE |    .040628   .0128617     3.16   0.002     .0154196    .0658364
         CAN |   2.078751   .8295749     2.51   0.012     .4528141    3.704688
   _ISYSGP_4 |   -1.51115   .7204683    -2.10   0.036    -2.923242   -.0990585
         TYP |   2.906679   .9257469     3.14   0.002     1.092248     4.72111
        LOCD |   3.965535   .9820316     4.04   0.000     2.040788    5.890281
       _cons |  -6.680532   1.320663    -5.06   0.000    -9.268984    -4.09208
------------------------------------------------------------------------------
```

```
. lfit, group(10) table

Logistic model for STA, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0105 |     0 |   0.1 |    20 |  19.9 |    20 |
  |     2 | 0.0290 |     0 |   0.4 |    20 |  19.6 |    20 |
  |     3 | 0.0492 |     2 |   1.0 |    21 |  22.0 |    23 |
  |     4 | 0.0666 |     0 |   1.0 |    17 |  16.0 |    17 |
  |     5 | 0.1083 |     2 |   1.8 |    19 |  19.2 |    21 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.1674 |     2 |   2.6 |    17 |  16.4 |    19 |
  |     7 | 0.2254 |     5 |   3.9 |    15 |  16.1 |    20 |
  |     8 | 0.3171 |     4 |   5.5 |    16 |  14.5 |    20 |
  |     9 | 0.4554 |     8 |   7.6 |    12 |  12.4 |    20 |
  |    10 | 0.9623 |    17 |  16.1 |     3 |   3.9 |    20 |
  +--------------------------------------------------------+

       number of observations =       200
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =         4.00
                  Prob > chi2 =         0.8570

. lfit

Logistic model for STA, goodness-of-fit test

       number of observations =       200
 number of covariate patterns =       135
             Pearson chi2(129) =        79.23
                  Prob > chi2 =         0.9998
```

Because the distribution of $\hat{C}$ depends on m - asymptotics, the appropriateness of the p - value will depend on the estimated expected frequencies being large enough to employ this theory.

If one is concerned about the magnitude of the expected frequencies, selected adjacent columns may be combined to increase the size of the expected frequencies. Unfortunately, when this is done the power of the test is reduced since the degrees of freedom are reduced.

When $\hat{C}$ is calculated from fewer than 6 groups, it will almost always indicate that the model fits. Thus, try to use with as many groups as possible.

WEEK 7: AREA UNDER THE ROC CURVE

Some researchers have proposed using the 2×2 classification table as a measure of fit.

This table is the result of cross-classifying $y\left(0,1\right)$ with

$$\bar{y} = \begin{cases} 0 & \text{if } \hat{\pi} < c \\ 1 & \text{if } \hat{\pi} \geq c \end{cases}$$

and c is often taken to = .5.

This table is, unfortunately, a measure of $\left|\hat{\beta}\right|$ not the correctness of the model. We know in the normal theory discriminant function situation that:

(1) The logistic model is the correct model for $\Pr\left(y = 1 \middle| \underline{x}\right)$

(2) Classification is a function of the separation of the 2 groups

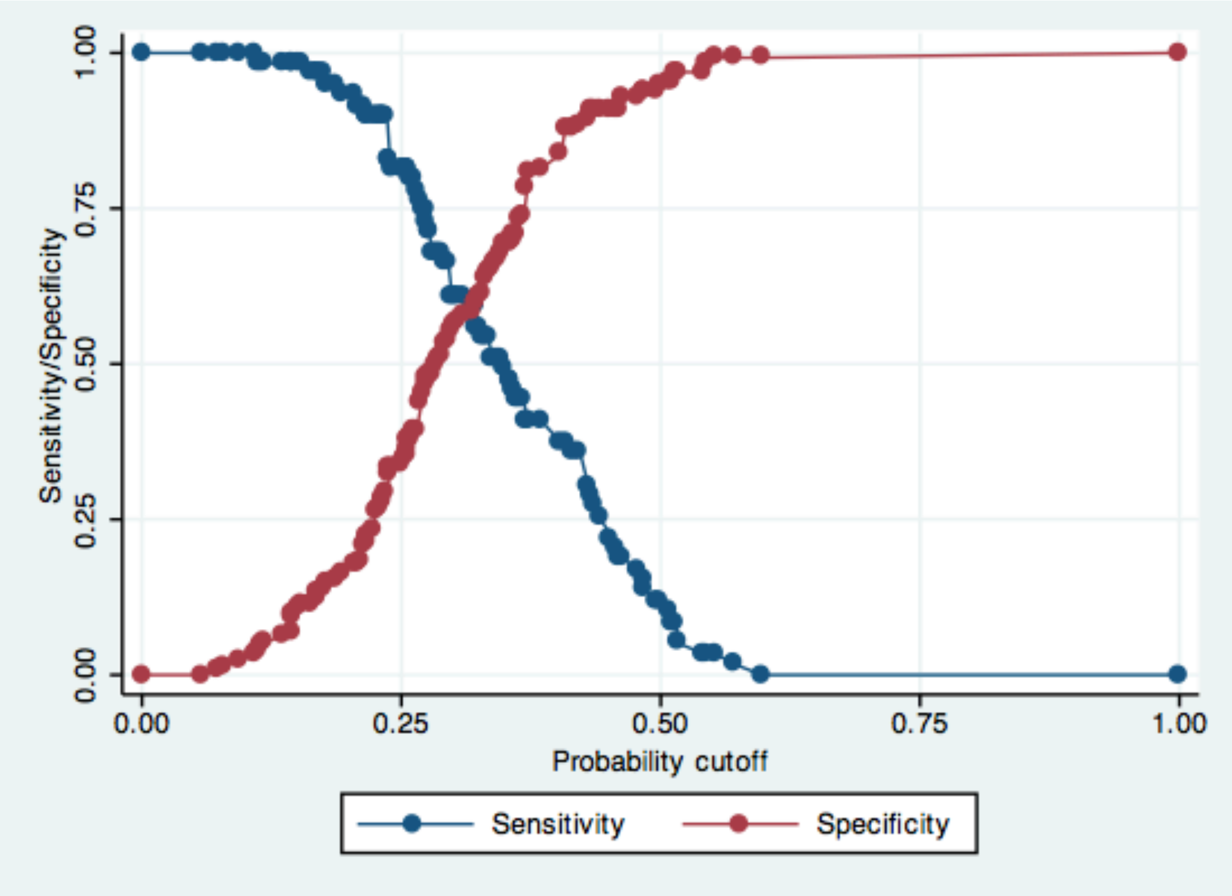
$$\Delta^2 = \left(\underline{\mu}_1 - \underline{\mu}_0\right)' \Sigma^{-1} \left(\underline{\mu}_1 - \underline{\mu}_0\right) = \underline{\beta}' \left(\underline{\mu}_1 - \underline{\mu}_0\right)$$

This is a function of $\underline{\beta}$ ... not the correctness of the model assumptions.

*Classification should not be used as a criterion for model adequacy unless it is a stated goal of the analysis.

STATA produces a graph of sensitivity and specificity versus probability cutoff:

```
.lsens
```



Area Under The ROC Curve
(a measure of *discrimination*)

Consider the model for estimating the probability of low birth weight.

- suppose we were interested in *predicting* the outcome for each patient.
  - One rule we might try is as follows:
    - predict baby will be low birth weight if $\Pr\left(\text{LOW}\right) \geq .50$
    - predict baby will be normal birth weight if $\Pr\left(\text{LOW}\right) < .50$

$\left(\text{Choice of .50 is traditional rather than optimal.}\right)$

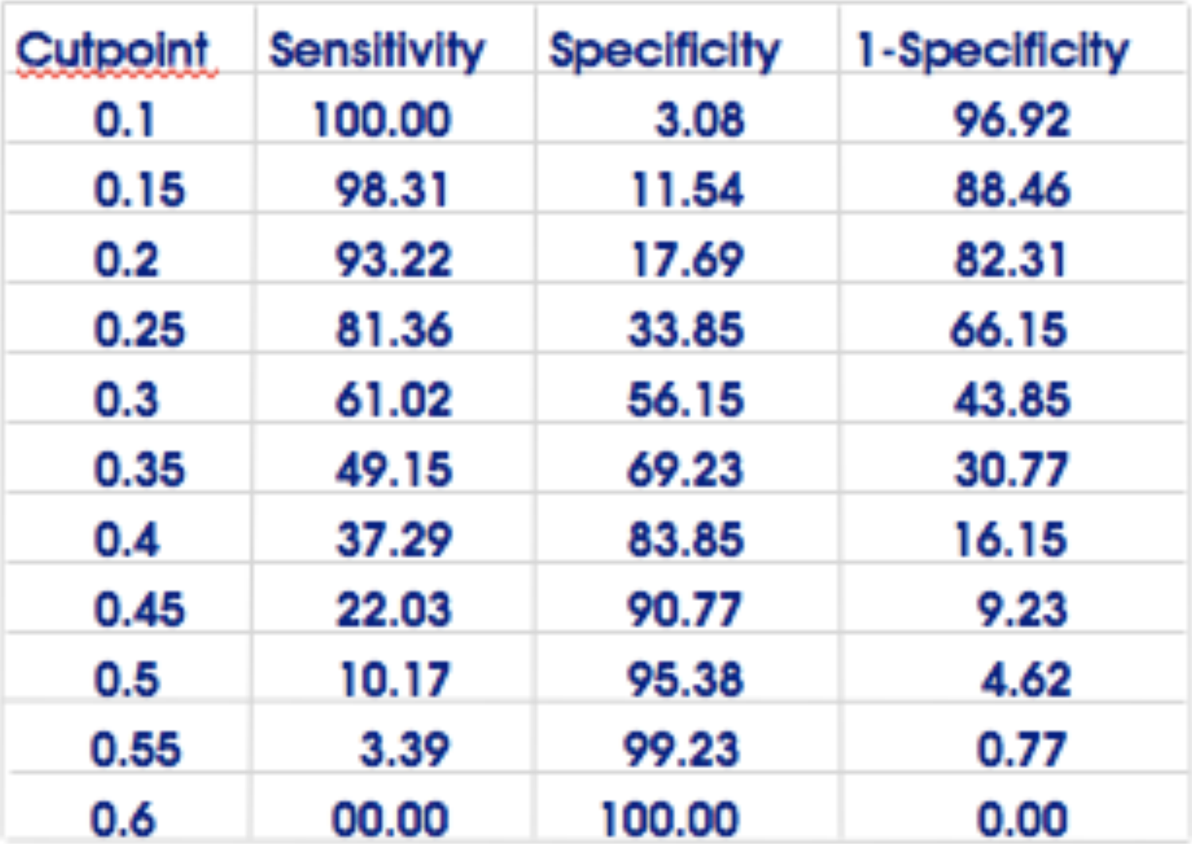
This would result in the following 2 × 2 classification table:

```
. lstat
Logistic model for low
              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         6             6  |         12
     -     |        53           124  |        177
-----------+--------------------------+-----------
   Total   |        59           130  |        189

Classified + if predicted Pr(D) >= .5
True D defined as low != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   10.17%
Specificity                     Pr( -|~D)   95.38%
Positive predictive value       Pr( D| +)   50.00%
Negative predictive value       Pr(~D| -)   70.06%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    4.62%
False - rate for true D         Pr( -| D)   89.83%
False + rate for classified +   Pr(~D| +)   50.00%
False - rate for classified -   Pr( D| -)   29.94%
--------------------------------------------------
Correctly classified                        68.78%
--------------------------------------------------
```

Suppose that, instead of a cutpoint of .5, .6 had been used:

```
. lstat, cutoff(.6)
Logistic model for low
              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         0             0  |          0
     -     |        59           130  |        189
-----------+--------------------------+-----------
   Total   |        59           130  |        189

Classified + if predicted Pr(D) >= .6
True D defined as low != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    0.00%
Specificity                     Pr( -|~D)  100.00%
Positive predictive value       Pr( D| +)       .%
Negative predictive value       Pr(~D| -)   68.78%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.00%
False - rate for true D         Pr( -| D)  100.00%
False + rate for classified +   Pr(~D| +)       .%
False - rate for classified -   Pr( D| -)   31.22%
--------------------------------------------------
Correctly classified                        68.78%
--------------------------------------------------
```

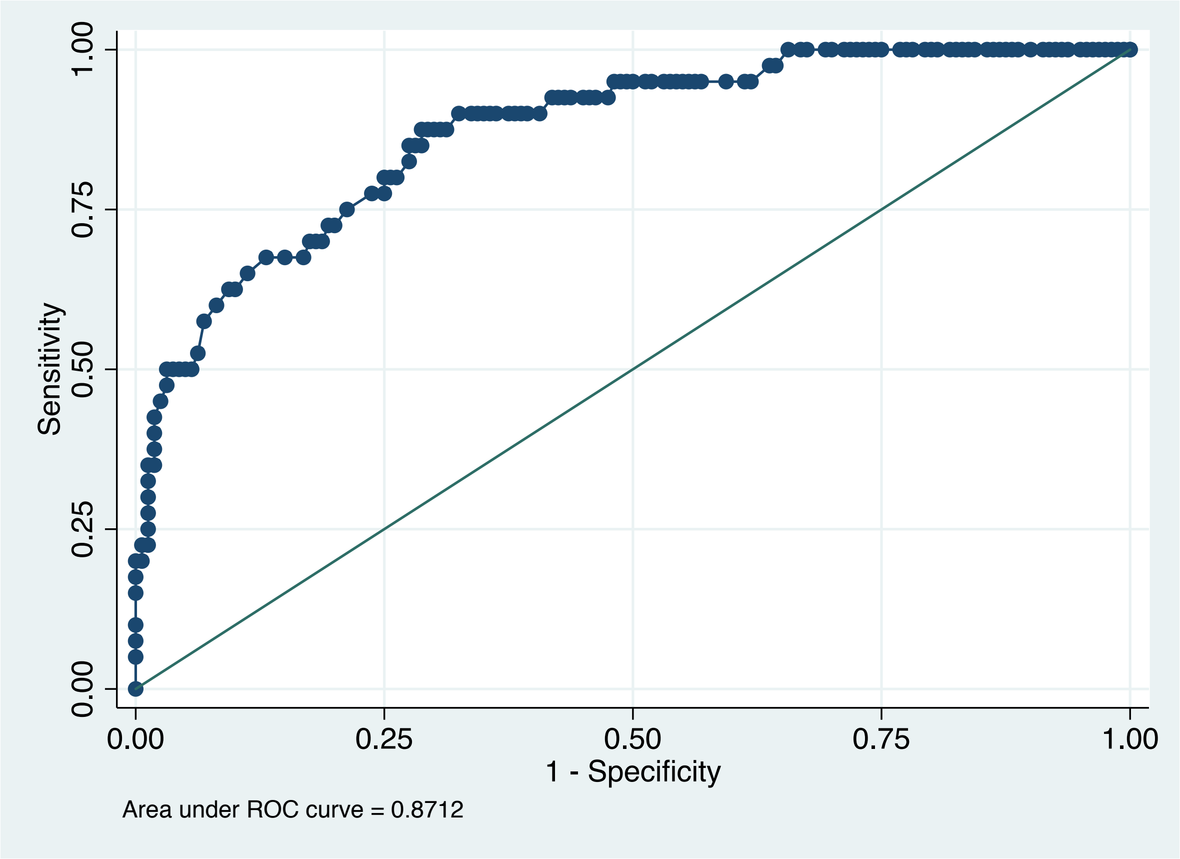
Summarizing results for all cutpoints between .1 and .6 in steps of .05, we have:



Cutpoint	Sensitivity	Specificity	1-Specificity
0.1	100.00	3.08	96.92
0.15	98.31	11.54	88.46
0.2	93.22	17.69	82.31
0.25	81.36	33.85	66.15
0.3	61.02	56.15	43.85
0.35	49.15	69.23	30.77
0.4	37.29	83.85	16.15
0.45	22.03	90.77	9.23
0.5	10.17	95.38	4.62
0.55	3.39	99.23	0.77
0.6	00.00	100.00	0.00

Plotting sensitivity vs (1 - specificity) we have the ROC Curve:

```
. lroc
Logistic estimates for LOW
Area under ROC curve = 0.6473
```



The area under the ROC curve is a measure of discrimination.

- It is a measure of the likelihood that a patient who has a low birth weight baby will have a higher Pr(LOW) than a patient with a normal birth weight baby.

As a general rule:

- $ROC = .5$: no discrimination $\left(\text{might as well just flip a coin}\right)$
- $ROC \geq .7$: considered acceptable discrimination
- $ROC \geq .8$: excellent discrimination
- $ROC \geq .9$: outstanding discrimination $\left(\text{very unusual}\right)$

note: a poorly fitting model (i.e., poorly calibrated as assessed by goodness-of-fit, Pearson X^2, etc) may still have good discrimination.

- e.g., simply add .25 to every probability in a good fitting logistic model.
  - the calibration will fall apart
  - the discrimination will not be affected at all

Another way to get the area under the ROC Curve is as follows:

- let n_1 = no. of patients who have low birthweight babies
- let n_2 = no. of patients who have normal birthweight babies
- create $n_1 \times n_2$ pairs

  i.e., each patient who gave birth to a low birthweight baby is paired with each patient who had a normal birthweight baby
- of these $n_1 \times n_2$ pairs, determine proportion of the time that the woman who had the low birthweight baby had the higher of the two probabilities

  - This proportion is the area under the ROC Curve

This can be done easily by running the nonparametric Mann-Whitney U Test:

```
. ranksum pihat, by(LOW)

Test: Equality of medians (Two-Sample Wilcoxon Rank-Sum)

 Sum of Ranks: 6735 (LOW == 1)
 Expected Sum: 5605

 z-statistic  3.24
 Prob > |z|   0.0012

. tabulate LOW
        LOW|      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        130       68.78       68.78
          1 |         59       31.22      100.00
------------+-----------------------------------
      Total |        189      100.00

. display 59*130+((59*60)/2)-6735
2705

. display 2705/(59*130)
.35267275
```

Wilcoxon Rank-sum test

Converts Wilcoxon Rank-sum test to Mann-Whitney U

$$U = mn + \frac{m(m+1)}{2} - T$$

$$m = \min\{n_1, n_2\}$$

$$n = \max\{n_1, n_2\}$$

Area Under ROC Curve

$$= 1 - .3527 = .6473$$

Suppose you need the confidence intervals for the area under the ROC curve. Stata 13 now has the option to perform nonparametric analysis of the ROC curve using bootstrap. To do this:

(1) Run the logistic regression

(2) Estimate the logit, and

(3) Run the nonparametric ROC regression command on the logit.

Logistic regression, generation of the logit, and lroc

```
. logit LOW LWT i.RACE, nolog

Logistic regression                             Number of obs     =        189
                                                LR chi2(3)        =      11.41
                                                Prob > chi2       =     0.0097
Log likelihood = -111.62955                     Pseudo R2         =     0.0486

------------------------------------------------------------------------------
         LOW |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         LWT |  -.0152231   .0064394    -2.36   0.018     -.027844   -.0026022
             |
        RACE |
          2  |   1.081066   .4880522     2.22   0.027     .1245015    2.037631
          3  |   .4806033   .3566737     1.35   0.178    -.2184644    1.179671
             |
       _cons |   .8057535   .8451667     0.95   0.340    -.8507428     2.46225
------------------------------------------------------------------------------

. predict logit, xb

. lroc, nograph all

Logistic model for LOW

number of observations =      189
area under ROC curve   =   0.6473
```

Nonparametric ROC regression of low birth weight on the predicted logit with a random seed, 500 replications, and correction for ties

- the nodots option suppresses some of the Stata output

```
. rocreg LOW logit,  bseed(04062012) breps(500)  tiecorrected nodots


Bootstrap results                                   Number of obs      =       189
                                                    Replications       =       500

Nonparametric ROC estimation

Control standardization: empirical, corrected for ties
ROC method             : empirical

Area under the ROC curve

Status    : LOW
Classifier: logit
------------------------------------------------------------------------
      |    Observed                 Bootstrap
  AUC |       Coef.       Bias      Std. Err.     [95% Conf. Interval]
------+-----------------------------------------------------------------
      |    .6473272    .002192     .0427415     .5635555    .731099   (N)
      |                                         .5724257   .7336735   (P)
      |                                         .5720991   .7292663 (BC)
```

Note:

(N) normal confidence interval; (P) percentile confidence interval

(BC) bias-corrected confidence interval