PBHS 32410 / STAT 22401 J. Dignam Winter 2025

Regression Diagnostics for Binary Outcome Models

- As in the case of ordinary least squares regression, post-model diagnostics can be used to
  - check model assumptions
  - check (indirectly) for omitted predictors, functional form of predictors
  - identify unusual data observations - that do not fit the model well
  - other aspects
- For binary regression models, one useful quantity is the *residual* - difference between observed and predicted. This quantity is most relevant for predicted proportions (recall all individual cases have a 0/1 outcome), but can be computed for each individual case

Regression Diagnostics for Binary Outcome Models

Before specific diagnostics, one overarching issue in binary outcome data is *sparseness* **or** *sparse tables*

– if for a given level of a discrete covariate, all responses are zero, then the odds ratio against a reference category is 0. If baseline category has zero events, then $OR = \infty$

– sparse tables can occur when many covariates are considered together, leading to a high-dimensional layout of the covariate combinations cross-classified with outcome.

– In multiway tables, main effects of each covariate may be estimated, but interaction effects will become not estimable. This may not be a problem in many analyses, since high-level interactions are difficult to interpret or verify

Zeros in Tables

The Donner Party data - females under 25

```
. cc dstat sexcode if  AGE < 25

                 |        sexcode         |             Proportion
                 |   Exposed   Unexposed  |      Total      exposed
-----------------+------------------------+------------------------
           Cases |         4           0  |          4       1.0000
        Controls |         3           6  |          9       0.3333
-----------------+------------------------+------------------------
           Total |         7           6  |         13       0.5385
                 |                        |
                 |      Point estimate    |    [95% conf. interval]
                 |------------------------+------------------------
      Odds ratio |                .       |    1.416577           . (Cornfield)
 Attr. frac. ex. |                .       |    .2940729           . (Cornfield)
 Attr. frac. pop |                .       |
                 +-------------------------------------------------
                               chi2(1) =     4.95  Pr>chi2 = 0.0261

 Note: Exact confidence levels not possible with zero count cells.
```

```
. logit dstat sexcode  if AGE < 25

note: sexcode != 1 predicts failure perfectly;
      sexcode omitted and 6 obs not used.

Iteration 0:  Log likelihood = -4.7803567
Iteration 1:  Log likelihood = -4.7803567

Logistic regression                                     Number of obs =      7
                                                        LR chi2(0)    =   0.00
                                                        Prob > chi2   =      .
Log likelihood = -4.7803567                             Pseudo R2     = 0.0000

------------------------------------------------------------------------------
       dstat | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     sexcode |          0  (omitted)
       _cons |   .2876821   .7637626     0.38   0.706    -1.209265    1.784629
------------------------------------------------------------------------------
```

Zeros in Tables

Note: main effects model in all subjects is estimable:

```
. logit dstat sexcode age25plus
. . .
------------------------------------------------------------------------------
       dstat | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     sexcode |   1.280376   .6992295     1.83   0.067    -.0900883    2.650841
   age25plus |   1.348034   .7356685     1.83   0.067    -.0938496    2.789918
       _cons |  -1.581687   .7804491    -2.03   0.043    -3.111339   -.0520348
------------------------------------------------------------------------------
```

	female	male
under 25	1.00	3.60
25+	3.85	13.85

Zeros in Tables

Note: main with interactions effects model not estimable:

```
. logit dstat sexcode age25plus agebysex
. . .
------------------------------------------------------------------------------
       dstat | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     sexcode |   17.63652   2388.794     0.01   0.994    -4664.313    4699.586
   age25plus |   34.60506   4777.587     0.01   0.994    -9329.294    9398.504
    agebysex |  -17.03307   2388.794    -0.01   0.994    -4698.983    4664.917
       _cons |  -17.34885   2388.794    -0.01   0.994    -4699.298    4664.601
------------------------------------------------------------------------------
```

Sub-tables do not all have OR estimates

Zeros in Tables

Reclassifying age into 25 or under and 26+:

```
. logit dstat sexcode age26plus

------------------------------------------------------------------------------
       dstat | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     sexcode |   1.355987   .6763103     2.00   0.045     .0304436    2.681531
   age26plus |   .5395536   .6381147     0.85   0.398    -.7111282    1.790235
       _cons |  -.9569176   .6412532    -1.49   0.136    -2.213751    .2999156
------------------------------------------------------------------------------
```

	female	male
under 25	1.00	3.88
25+	1.72	6.66

Zeros in Tables

Reclassifying age into $\leq$ 25 and 26+, interaction added:

```
. logit dstat sexcode age26plus age26bysex

------------------------------------------------------------------------------
       dstat | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     sexcode |    2.75684   1.226364     2.25   0.025     .3532114    5.160469
   age26plus |   2.233592   1.313846     1.70   0.089     -.341498    4.808682
  age26bysex |  -2.438387   1.531299    -1.59   0.111    -5.439678    .5629047
       _cons |   -1.94591   1.069045    -1.82   0.069      -4.0412    .1493795
------------------------------------------------------------------------------
```

	female	male
under 25	1.00	15.75
25+	9.33	12.83

Age effect within sex different here - present in women only

Zeros in Tables and Sparse Tables Generally

- How predictor variables are grouped or partitioned may need to be considered for estimable effects.
- Zero is a valid value for an odds ratio.
- Methods for *sparse contingency tables* is a large area in statistical methodology and application. These methods are needed when there are a lot of covariates (and combinations) relative to number of failures - for examle, genomic variant data

Logistic Regression - Back to Residuals

Two types of residuals:

– Pearson residuals are defined to be the standardized difference between the observed frequency (proportion) and the predicted frequency . These measure the relative deviations between the observed and fitted values.

$$r_i = \frac{y_i - \hat{u}_i}{\sqrt{\hat{u}_i(n_i - \hat{u}_i)/n_i}} \tag{1}$$

– Equals square root of the i^{th} component of the Pearson Chi-square statistic.

Logistic Regression Residuals

Two types of residuals (cont.):

– Deviance residuals are components of the deviance statistic, which measure the disagreement between the maxima of the observed and the fitted log likelihood functions.

$$d_i = \sqrt{2\{y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{\mu}_i})\}} \qquad (2)$$

– Equals square root of i^{th} component of the deviance statistic (with sign re-attached).

– Analogous to the raw residual in OLS regression, where the goal is to minimize the sum of squared residuals. Logistic regression estimation minimizes the sum of the deviance residuals (using maximum likelihood to solve)

Logistic Regression Residuals

- These quantities can be listed, plotted to identify unusual values
- In grouped data, can approximately normally distributed, zero-centered random variable (i.e., Z-statistic
- Thus, values greater than +/- 2 may be of interest (‘extreme’ values on the Z scale)

Ex/ Logistic Regression Residuals

– Outcome (survival) of root cuttings may be related to cutting time and length

Table 1: Survival rate of plum root-stock cuttings

Length	Planting	Surviving	Proportion
Short	Immediate	107	0.45
	In spring	31	0.13
Long	Immediate	156	0.65
	In spring	84	0.35

Ex/ Logistic Regression Residuals

```
. glm number length time, family(binomial total)
.
Generalized linear models                          No. of obs      =          4
Optimization     : ML                              Residual df     =          1
                                                   Scale parameter =          1
Deviance         =  2.293839315                    (1/df) Deviance =   2.293839
Pearson          =  2.270478953                    (1/df) Pearson  =   2.270479
Variance function: V(u) = u*(1-u/total)            [Binomial]
Link function    : g(u) = ln(u/(total-u))          [Logit]
                                                   AIC             =   7.758987
Log likelihood   = -12.51797425                    BIC             =    .907545
------------------------------------------------------------------------------
             |                 OIM
      number |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   1.017691     .14548     7.00   0.000     .7325559    1.302827
        time |  -1.427542   .1464624    -9.75   0.000    -1.714603   -1.140481
       _cons |  -.3039203   .1171647    -2.59   0.009     -.533559   -.0742816
------------------------------------------------------------------------------
```

Ex/ Logistic Regression Residuals

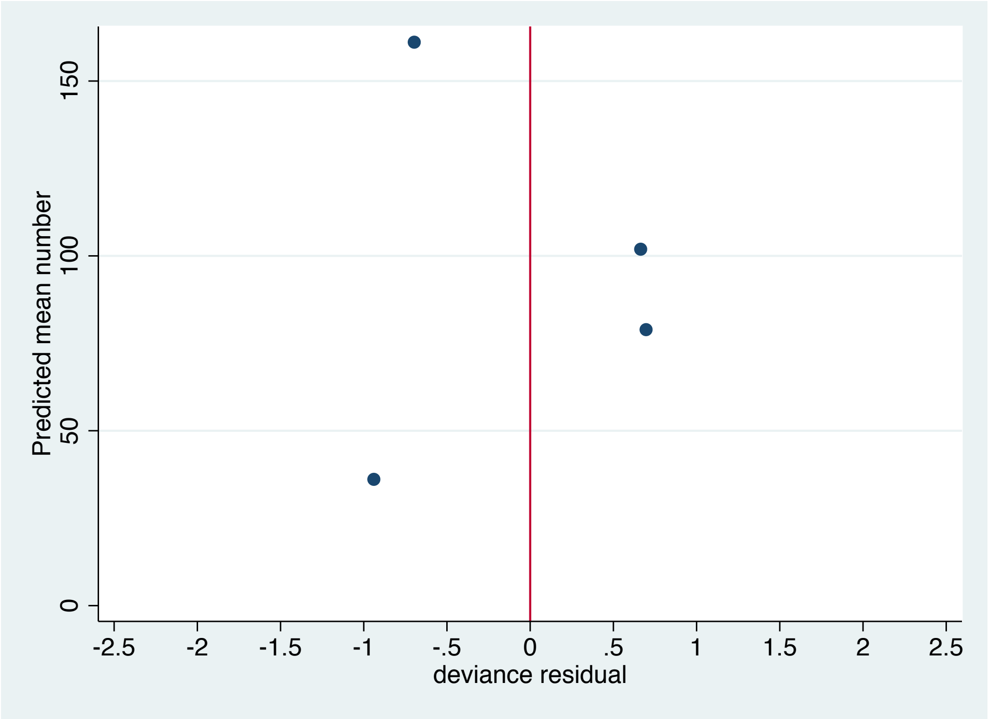
Model output and and residuals

```
. predict yhat
(option mu assumed; predicted mean number)
. predict dres, d
. predict pres, p
. list

     +------------------------------------------------------------------+
     | length   time   number   total        yhat         pres         dres |
     |------------------------------------------------------------------|
  1. |      0      0      107     240    101.9039     .6655197     .6641749 |
  2. |      0      1       31     240    36.09614    -.9202459    -.9392965 |
  3. |      1      0      156     240    161.0961    -.7002535    -.6965749 |
  4. |      1      1       84     240    78.90386     .7002535     .6965749 |
     +------------------------------------------------------------------+
```

Ex/ Logistic Regression Residuals

plot the residuals:

```
, twoway (scatter yhat dres), xlabel(-2.5(.5)2.5) xline(0)
```



Not very exciting - but these indicate reasonable fit. Residuals are more interesting for larger models and/or continuous predictors

Logistic Regression - another model

Example: Biomarkers in localized, high risk prostate cancer

- Men with localized prostate cancer deemed 'high risk' based on clinical and pathologic features may undergo radiation therapy and long-term androgen deprivation. For some, this is too aggressive an intervention, while for others, even this approach does not sufficiently decrease risk of eventual metastatic disease and death
- Using data from subset of clinical trial RTOG 9202, we investigated tumor biomarkers for relationship to distant metastasis event to identify and explain heterogeneity in distant mets risk (Pollack et al *Clin Cancer Res* 2014)

Logistic Regression -

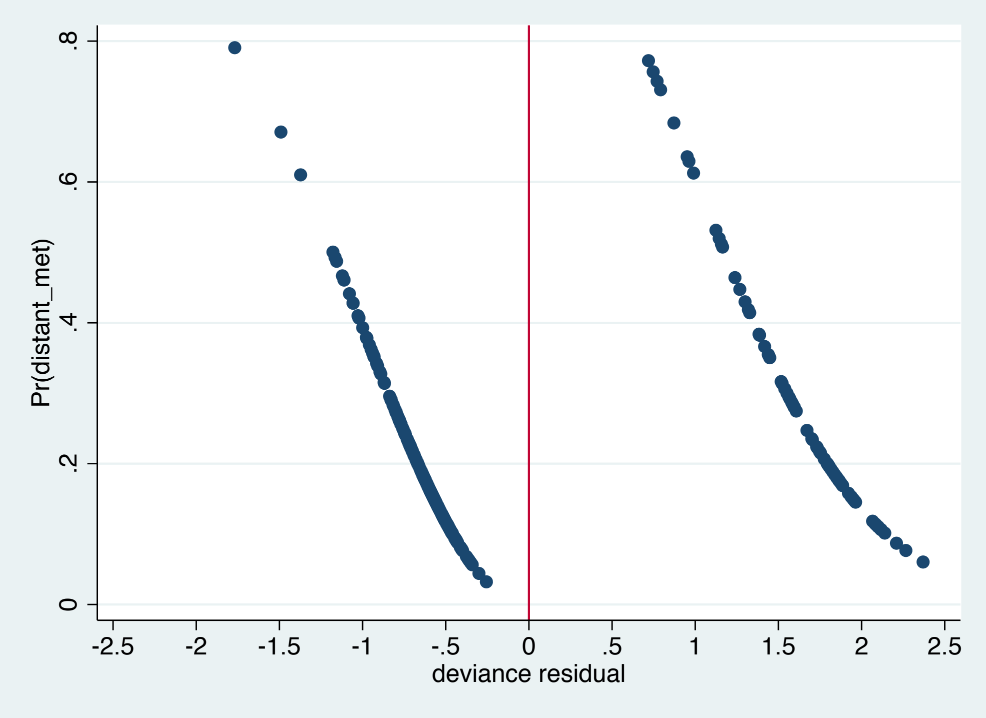
```
. tab distant_met if all_markers

distant_met |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,056       81.48       81.48
          1 |        240       18.52      100.00
------------+-----------------------------------
      Total |      1,296      100.00

. list distant_met ki67_acis10_index_percent p16_index_percent mdm2_intensity cox2_intensity
      distan~t   ki67_a~t   p16_~ent   mdm2_i~y   cox2_i~y
  2.         0        8.6        100        211        159
  7.         0        7.6       94.7        178        149
  8.         0        7.7       61.1        187        131
 16.         0        6.9       87.7        178        156
 21.         1       21.4       38.4        159        166
 23.         1       13.8       71.3        194        158
 25.         0         28       90.8        152        128
 35.         0       22.5       92.1        186        149
 36.         0       43.5       96.3        178        161
 38.         0        1.5       89.2        144        200
.
.
```

Logistic Regression - Goodness of Fit

```
. logistic distant_met ki67_acis10_index_percent p16_index_percent mdm2_intensity cox2_intensity

Logistic regression                               Number of obs     =        372
                                                  LR chi2(4)        =      34.05
                                                  Prob > chi2       =     0.0000
Log likelihood = -174.01005                       Pseudo R2         =     0.0891
------------------------------------------------------------------------------------------
              distant_met | Odds Ratio   Std. Err.    z   P>|z|  [95% Conf. Interval]
--------------------------+---------------------------------------------------------------
ki67_acis10_index_percent | 1.064318   .0163019    4.07  0.000  1.032842    1.096754
        p16_index_percent | .9867829   .0056341   -2.33  0.020  .9758019    .9978876
           mdm2_intensity | 1.004605   .0032227    1.43  0.152  .9983081    1.010941
           cox2_intensity | 1.012377   .0069322    1.80  0.072  .9988811    1.026056
                    _cons | .0296984    .028565   -3.66  0.000  .0045083    .1956366
-------------------------------------------------------------------------------------------
. estat gof, group(10)

Logistic model for distant_met, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =       372
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =        11.79
                  Prob > chi2 =         0.1610
```

Logistic Regression - Examining Residuals

```
 predict yhat
(option pr assumed; Pr(distant_met))
(1149 missing values generated)

. predict dres, dev
(1,149 missing values generated)

. twoway (scatter yhat dres), xlabel(-2.5(.5)2.5) xline(0)
```



- Residuals

```
. list cn distant_met yhat dres if dres > 2.0 & dres ~=., noobs clean

       cn   distan~t       yhat       dres
      101          1   .1182092   2.066543
      673          1   .0869857   2.209983
      826          1   .0602835   2.370104
      881          1   .0766417   2.266545
      996          1   .1068749   2.114756
     1006          1   .1013688   2.139621
     1098          1   .1148568   2.080418
     1250          1   .1131486   2.087608
     1337          1   .1060305   2.118504
     1486          1    .110245   2.100024
     1495          1   .1104176   2.099279
```

These cases had low predicted probability of distant mets yet had distant mets - may be worth further examination

- Logistic Model - add more predictors

We add some important clinical/pathologic predictors to the model

```
. logistic distant_met ki67_acis10_index_percent p16_index_percent mdm2_intensity
             cox2_intensity psahigh age gleason_d2
Logistic regression                                Number of obs     =        343
                                                   LR chi2(7)        =      65.42
                                                   Prob > chi2       =     0.0000
Log likelihood = -136.68406                        Pseudo R2         =     0.1931
------------------------------------------------------------------------------------------
              distant_met | Odds Ratio Std. Err.   z    P>|z|    [95% Conf. Interval]
--------------------------+---------------------------------------------------------------
ki67_acis10_index_percent | 1.05912   .0180011    3.38   0.001     1.02442   1.09499
        p16_index_percent |  .98398   .0060536   -2.62   0.009     .972189   .995919
           mdm2_intensity | 1.00355    .003444    1.03   0.301     .996829    1.01033
           cox2_intensity | 1.022     .0081364    2.73   0.006     1.00617   1.03807
                  psahigh | 3.31139   2.132884    1.86   0.063     .9370218   11.70236
                      age |  .957837   .0215753   -1.91   0.056     .9164707   1.001072
               gleason_d2 | 5.11480   1.648314    5.06   0.000     2.719684   9.619226
                    _cons | .109938   .1999009   -1.21   0.225     .0031146   3.880518
------------------------------------------------------------------------------------------
```

These factors contribute based on beta values, also can possibly drop mdm2 intensity as a predictor

Final Model

```
.  logistic distant_met ki67_acis10_index_percent p16_index_percent cox2_intensity
                psahigh age gleason_d2

Logistic regression                             Number of obs     =        392
                                                LR chi2(6)        =      75.42
                                                Prob > chi2       =     0.0000
Log likelihood = -149.23974                     Pseudo R2         =     0.2017
-------------------------------------------------------------------------------------------
              distant_met | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------------+----------------------------------------------------------------
ki67_acis10_index_percent | 1.063578   .0173017    3.79   0.000     1.030202    1.098035
        p16_index_percent | .9859311   .0055475   -2.52   0.012      .975118    .9968641
           cox2_intensity |  1.02084   .0072459    2.91   0.004     1.006737    1.035141
                  psahigh | 3.971405   2.331307    2.35   0.019     1.256809    12.54929
                      age | .9510925   .0206622   -2.31   0.021     .9114453    .9924643
               gleason_d2 | 6.013003   1.854401    5.82   0.000     3.285353    11.00527
                    _cons | .2621869   .4488802   -0.78   0.434     .0091475    7.514857
-------------------------------------------------------------------------------------------
```

Logistic Regression - Variability Explained

In logistic regression, several analogues to the R^2 have been suggested. One simple one is

$$R^2_{pseud} = 1 - \frac{\log L(\hat{\beta})}{\log \hat{L}_0}$$

where $\log L(\hat{\beta})$ is the log likelihood for the current model and $\log \hat{L}_0$ is the null model.

- additionally, an R^2 adjusted for the number of parameters may be used
- Generally, these measures are not as reliable as fits measures as in the linear regression setting
- For the prostate model, the value was about .20, or 20% of the variation in risk of distant mets is explained by covariates.

Logistic Regression - Prediction

Several other post-model outputs relating to prediction are available. One of these uses assessment of classification performance:

- Cross-classifying to compare agreement between predicted and observed outcomes under some assignment rule such as - 'any case with $> .50$ predicted probability of being an event (1) will be classified as an event' - Use familiar measures of sensitivity, specificity, PPV, NPV to summarize rule
- The receiver operating characteristic (ROC) curve - plot of sensitivity vs 1 - specificity - evaluation of prediction rule over all possible cut-points of probability. This can be shown to equal to probability of correctly determining among two random cases, which will be an event.

Results can look encouraging, but true test of model performance must be assessed on *independent* data - not use to build the model

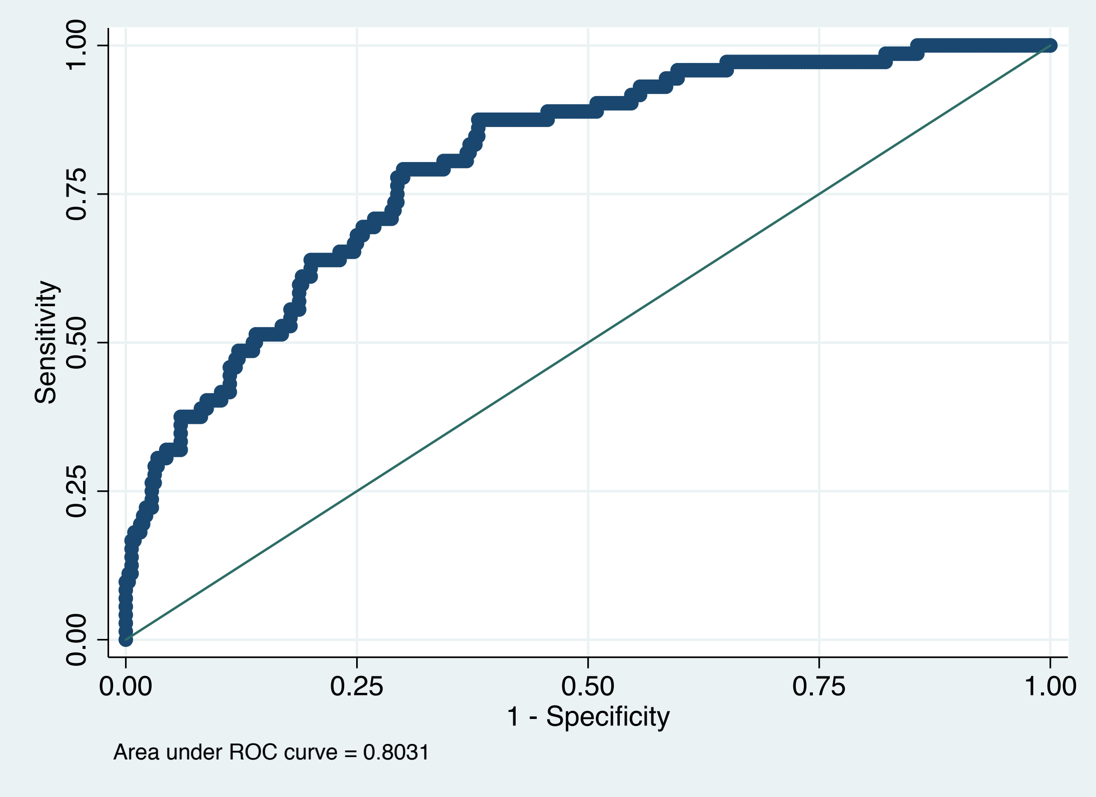
Logistic Regression - Prediction

```
. estat classification

 Logistic model for distant_met
               -------- True --------
 Classified |         D            ~D  |      Total
 -----------+--------------------------+-----------
     +      |        19             9  |         28
     -      |        53           311  |        364
 -----------+--------------------------+-----------
   Total    |        72           320  |        392
 Classified + if predicted Pr(D) >= .5   -True D defined as distant_met != 0
 --------------------------------------------------
 Sensitivity                     Pr( +| D)   26.39%
 Specificity                     Pr( -|~D)   97.19%
 Positive predictive value       Pr( D| +)   67.86%
 Negative predictive value       Pr(~D| -)   85.44%
 --------------------------------------------------
 False + rate for true ~D        Pr( +|~D)    2.81%
 False - rate for true D         Pr( -| D)   73.61%
 False + rate for classified +   Pr(~D| +)   32.14%
 False - rate for classified -   Pr( D| -)   14.56%
 --------------------------------------------------
 Correctly classified                        84.18%
```

Logistic Regression - Prediction

Curve Based on 6-predictor model for distant mets:

```
. lroc
Logistic model for distant_met
number of observations =      392
area under ROC curve   =   0.8031
```



Changing the cutpoint will change the performance:

```
. estat classification, cutoff(.35)
Logistic model for distant_met
              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        33            36  |         69
     -     |        39           284  |        323
-----------+--------------------------+-----------
   Total   |        72           320  |        392
Classified + if predicted Pr(D) >= .35  -True D defined as distant_met != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   45.83%
Specificity                     Pr( -|~D)   88.75%
Positive predictive value       Pr( D| +)   47.83%
Negative predictive value       Pr(~D| -)   87.93%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   11.25%
False - rate for true D         Pr( -| D)   54.17%
False + rate for classified +   Pr(~D| +)   52.17%
False - rate for classified -   Pr( D| -)   12.07%
--------------------------------------------------
Correctly classified                        80.87%
```

Model Assessment

- Diagnostic tools follow those of linear regression models, but may be a bit more difficult to interpret. These are nonetheless useful.
- There are many more diagnostics covered in other texts, including identifying influential observations, outliers, assumptions about the distribution (ie link function), etc.
- Post-model classification tools assess the utility of the model in prediction. Developing such models further requires independent validation.