Epidemiology/Biostats VHM 812/802 Course
Winter 2014, Atlantic Veterinary College, PEI
Javier Sanchez

Lecture 5a: Logistic regression diagnostics

Index **Page**

Friday

- logistic regression exercises 16.2 and 16.3
- conditional logistic regression (and exact logistic regression) or
- last chance to work on exercises!!

Dataset= nocardia.dta

- all the examples based on VER Ex. 16.5 (model with dcpct3, dneo, dclox and dneo*dclox)

Covariate patterns

- covariate pattern
  - unique combination of values predictor variables

Binomial Data $X_1 = 0/1$ $X_2 = 0/1$					
X_1	X_2	Cov. Patter	# pos	n	propn.
0	0	1	6	10	.6
0	1	2	3	20	.15
1	0	3	5	50	.1
1	1	4	4	10	.4

Binary Data X_1 = age (in yrs. to 1 decimal) X_2 = wt in kg. (to 1 decimal)					
X_1	X_2	Cov. Patter	# pos	n	propn.
4.3	527.2	1	0	1	0
3.7	489.6	2	1	1	1
2.1	535.4	3	1	1	1
5.6	501.4	4	0	1	0
		....			

- example

				Residual	
Obs.	Cov. pattern	Disease	Pred. Value	1 per Obs.	1 per Cov. Pat.
1					
2					

Residuals in logistic regression

- one per observation (based on Hilbe, 2009[1])
  - (standard) residual analysis
    - Pearson and Deviance residuals (and standardized forms)
  - mainly for visual assessment
  - not very useful for assessing the model
  - Stata -glm- command

- one per covariate pattern
  - goodness-of-fit tests
    - Person and Deviance residuals
  - residual analysis
    - Pearson and Deviance residuals (standardized)
  - leverage
  - influential observations
    - delta χ^2 ($\Delta\chi^2$)
    - delta Deviance (ΔD)
    - leverage
  - Stata -logit / logistic - command

1Hilbe J. Logistic Reg. Models. CRC Press: Boca Raton 2009

Pearson and Deviance residuals per covariate pattern

- Pearson residual
  - $r_j = \frac{y_j - m_j * p_j}{\sqrt{m_j * p_j * (1 - p_j)}}$
    - yj = nbr. pos. outcomes in jth covariate pattern
    - mj = nbr. obs. in the jth covariate pattern
    - pj = predicted prob. for the jth covariate pattern
  - standardized residuals (same as before)
  - $\sum_{j=1}^{J} (r_j)^2 = \text{Pearson}\,\chi^2\, statistic \sim \chi^2\, with\, (J-k)\, df.$
    - k = number of parameters in the model
  - contribution of each cov. pattern to the χ^2 statistic
  - Stata – logit/logistic- command
- Deviance residual
  - contribution of each cov. pattern to the deviance (d = -2*Log likelihood)
    - standardized residuals (same as before)

$$\sum_{j=1}^{J} (d_j)^2 = deviance\, X^2 \sim \chi^2\, distribution\, with\, (J-k)\, df$$

Goodness of fit tests

- Pearson χ^2 and Deviance χ^2 (1 per cov. pattern)
  - χ^2 distributions (J-k) df
    - J = # covariate patterns
    - k = # of parameters in model
  - only if enough number of replications per group (eg cov. patterns)
    - ~ guidelines ~ X^2-statistic
      - more than 1 expected counts in each cell
      - at least 80% expected values > 5 counts
  - indicates the fit of the model
    - Ho = model fits the data

- Pearson $\chi 2$
  - post estimation command (eg. after logit)
  - -estat gof- command –

```
. estat gof

Logistic model for casecont, goodness-of-fit test

       number of observations =       108
 number of covariate patterns =        11
             Pearson chi2(5) =         8.22
                 Prob > chi2 =         0.1444
```

- Deviance $\chi 2$
  - likelihood ratio test
    - model with J covariate patterns (full model) vs final model (reduced model) [see L4a-10]

- Example

```
. logit casecon i.cov, asis nolog

Logistic regression                                     Number of obs   =        108
                                                        LR chi2(10)     =      52.78
                                                        Prob > chi2     =     0.0000
Log likelihood =  -48.47167                             Pseudo R2       =     0.3525

------------------------------------------------------------------------------
    casecon |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
         cov |
          2  |  -13.21623   1737.975    -0.01   0.994    -3419.585    3393.153
          3  |   .4518912   1.494591     0.30   0.762    -2.477453    3.381236
  .....
         10  |   1.011361   1.530048     0.66   0.509    -1.987478      4.0102
         11  |   2.174692   1.241325     1.75   0.080    -.2582602    4.607645
             |
       _cons |   -2.39787   1.044455    -2.30   0.022    -4.444964   -.3507759
------------------------------------------------------------------------------

. estimates store full
. logit casecon dneo##dclox i.dcpct3, nolog

Logistic regression                                     Number of obs   =        108
                                                        LR chi2(5)      =      46.46
                                                        Prob > chi2     =     0.0000
Log likelihood = -51.632242                             Pseudo R2       =     0.3103
------------------------------------------------------------------------------
    casecon |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        dneo |
        yes  |    3.19238   .8361783     3.82   0.000       1.5535    4.831259
       dclox |
        yes  |   .4529145   1.026657     0.44   0.659    -1.559296    2.465125
  dneo#dclox |
    yes#yes  |  -2.532558   1.207714    -2.10   0.036    -4.899634   -.1654829
      dcpct3 |
         50  |   1.361002    .819178     1.66   0.097    -.2445579    2.966561
        100  |   2.026562   .6855237     2.96   0.003     .6829604    3.370164
       _cons |  -3.531226   .9364287    -3.77   0.000    -5.366593    -1.69586
------------------------------------------------------------------------------

. estimates store red
. lrtest full red
Likelihood-ratio test                                 LR chi2(5)  =      6.32
(Assumption: red nested in full)                      Prob > chi2 =    0.2762
```

- Example

covariate pattern and Case - Control		mean(cnt)	mean(pv)	mean(pear)	mean(dev)
1	no	12	0.028	1.144	.9331458
	yes	.			
2	no	2	0.102	-0.478	-.6575516
3	no	8	0.182	-0.416	-.4359247
	yes				
4	yes	1	0.152	2.360	1.940331
5	no	11	0.259	-0.584	-.6060105
	yes				
6	no	11	0.416	-0.353	-.355453
	yes				
7	no	10	0.735	-0.254	-.2503426
	yes				
8	no	38	0.844	0.416	.4256602
	yes				
9	no	1	0.082	-0.298	-.4130647
10	no	5	0.258	-0.295	-.3036099
	yes				
11	no	9	0.403	0.252	.2506801
	yes				

$\chi^2 = 8.22$ and Deviance = 6.32

- no indication of lack of fit Pearson X2 and Deviance X2 non-significant with df=5
- largest contribution is from cov. pattern with no replication (eg cov # 4)

- Hosmer-Lemeshow Test
  - compares predicted probabilities to observed probabilities in groups of data
  - group by:
    - percentiles of estimated probability
    - fixed points of estimated probability
      - Example: 5 obs. per group, 1st group

Group	Obs #	Y	Pred. prob.
1	1	0	0.05
1	2	0	0.05
1	3	0	0.05
1	4	1	0.05
1	5	0	0.05
total	5	1	0.05

  - χ^2 with g-2 df
  - low power if < 6 groups
- Example

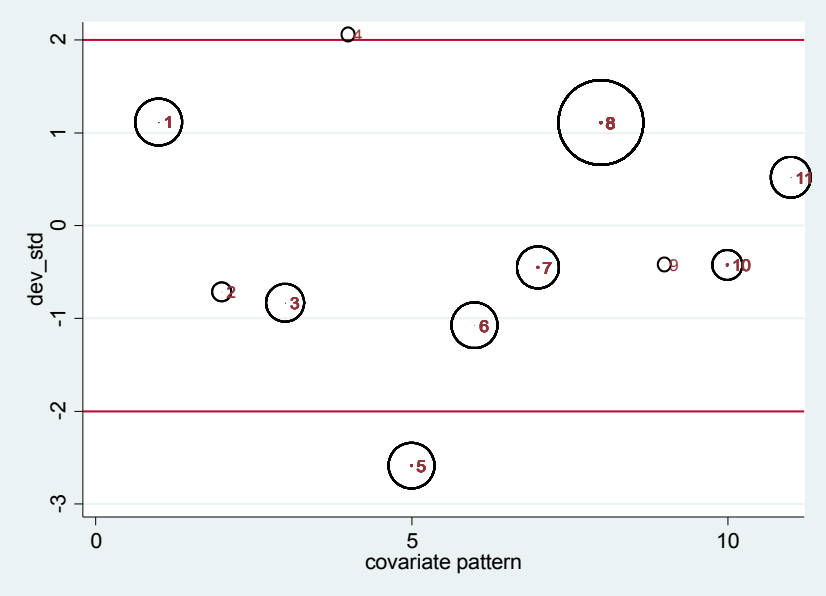
```
. estat gof, g(10) table

Logistic model for casecont, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  (There are only 7 distinct quantiles because of ties)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0284 |     1 |   0.3 |    11 |  11.7 |    12 |
  |     2 | 0.1817 |     2 |   1.9 |    10 |  10.1 |    12 |
  |     3 | 0.2589 |     3 |   4.1 |    13 |  11.9 |    16 |
  |     4 | 0.4033 |     4 |   3.6 |     5 |   5.4 |     9 |
  |     5 | 0.4161 |     4 |   4.6 |     7 |   6.4 |    11 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.7354 |     7 |   7.4 |     3 |   2.6 |    10 |
  |    10 | 0.8439 |    33 |  32.1 |     5 |   5.9 |    38 |
  +--------------------------------------------------------+

       number of observations =       108
             number of groups =         7
    Hosmer-Lemeshow chi2(5) =         2.16
                Prob > chi2 =         0.8262
```

Overdispersion [l11a - l11b]

- Assumption y_i ~ binomial distribution
  - mean = $n_i * p_i$
  - variance = $n_i * p_i * (1 - p_i)$
- overdispersion = the data are more dispersed (larger variance) that would be expected
  - apparent overdispersion – wrong model
    - missing important predictors
    - outliers
  - real overdispersion – usually due to clustering
  - biased estimates and small S.E.

Residual analysis (covariate patterns)

- Pearson and Deviance residuals
  - standardized residuals
    - 95% between -2 and +2
  - identify large negative and positive residuals
  - characteristics of observations
  - visual assessment
    - some guidelines about outlying obs.
- Example
  - stdz. deviance residuals (per cov. pattern) vs cov. pattern



cov	cnt	dcpct	dneo	dclox	avgcc	pv	dev_std	pear_std
5	11	100	no	yes	.1818182	.2588893	-2.592519	-2.496497
4	1	83	no	yes	1	.152218	2.052569	2.496497

Leverage

- potential impact of cov. pattern on the model
- extent to which the j^{th} cov. pattern is separated for the others in terms of the explanatory variables
- leverage depends on x's and predicted value
  - extreme value of predictors will have:

Predicted probabilities	Leverage
0.0 – 0.1	low
0.1 – 0.3	high
0.3 – 0.7	moderate
0.7 – 0.9	high
0.9 – 1.0	low

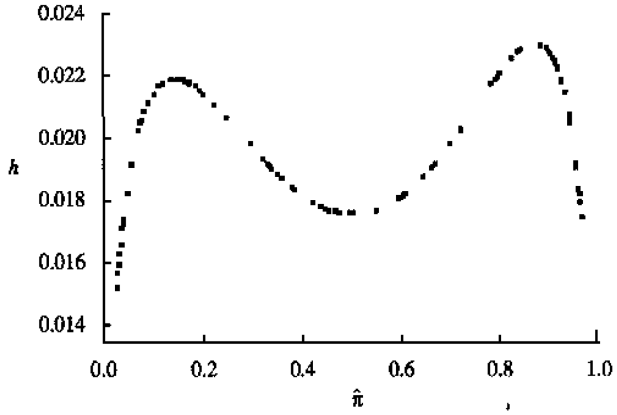


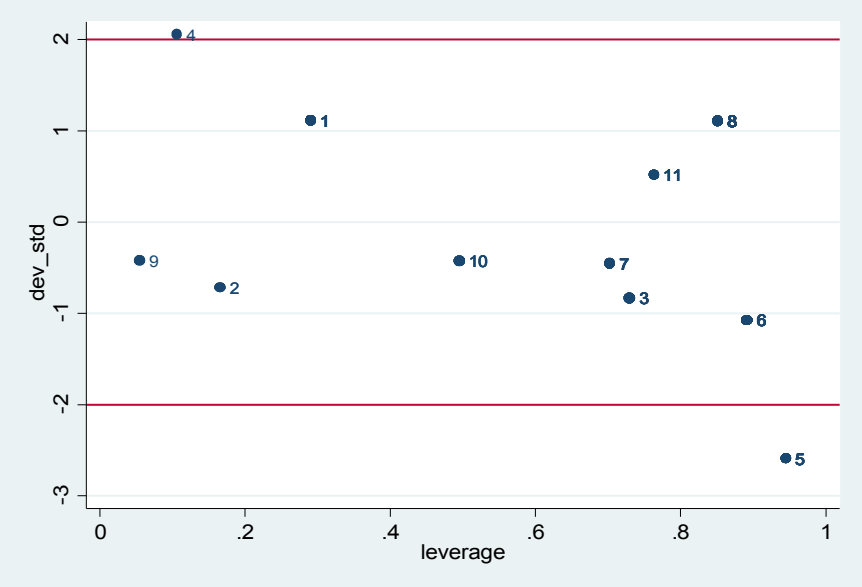
Figure 5.1 Plot of Leverage (h) Versus ($\hat{\pi}$).

- Example

```
.l cov herds dcpct dneo dclox pv dev_std lev if  lev >0.5 | abs(dev_std)>2, noobs
```

cov	herds	dcpct	dneo	dclox	pv	dev_std	lev
4	1	83	0	1	.152218	2.052569	.1063734
7	10	69	1	0	.7353922	-.4592829	.7028956
3	8	100	0	0	.1817309	-.8394304	.7303169
11	9	100	1	1	.4032532	.5157657	.76377
8	38	100	1	0	.8439235	1.10489	.8515815
6	11	15	1	0	.4160897	-1.081025	.8918833
5	11	100	0	1	.2588893	-2.592519	.9453593

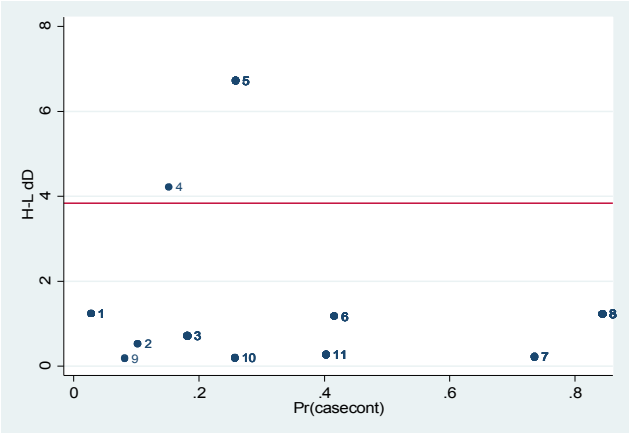
- standardized residuals vs leverage values

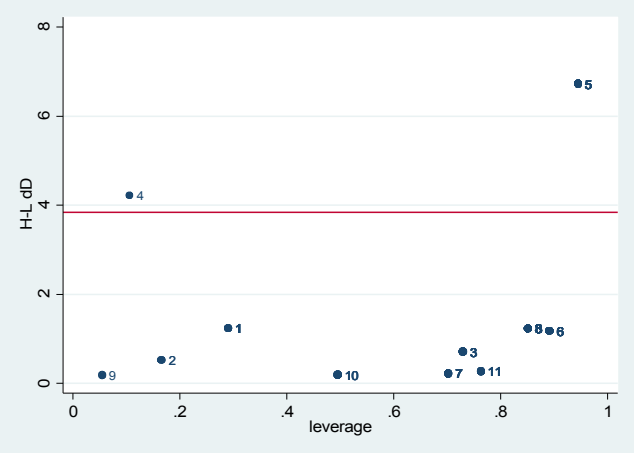


Influential statistics

- delta χ^2 ($\Delta\chi^2$) and delta deviance (ΔD)
  - effect of covariate pattern on Pearson χ^2 and Deviance χ^2 statistics
    - identifies patterns that are not fit well (outliers)
    - plot delta values vs predicted probabilities
    - plot delta values vs leverage
    - delta-values >=3.84 (95^{th} percentiles X^2 distribution with 1df)
    - recommended to use both $\Delta\chi^2$ and ΔD
      - generally $\Delta\chi^2$ is much greater than ΔD

- Example






```
. l cov herds dcpct dneo dclox pv dev_std dx2 ddev lev if ddev > 3.84 | dx2>3.84,
noobs table

  +-------------------------------------------------------------------------------+
  | cov   herds   dcpct   dneo   dclox      pv   dev_std     dx2     ddev     lev |
  |-------------------------------------------------------------------------------|
  |   4       1      83      0       1   0.152     2.053   6.232    4.213   0.106 |
  |   5      11     100      0       1   0.259    -2.593   6.232    6.721   0.945 |
  +-------------------------------------------------------------------------------+
```

- delta-betas ($\Delta\beta$)
  - analogous to Cook's distance
  - measures influence of a cov. pattern on:
    - overall set of betas (Stata)
    - individual betas (SAS)
  - depends on leverages and # of observations (m_j) on the covariate pattern variable
    - Hosmer & Lemeshow suggest that values >1 might be influential

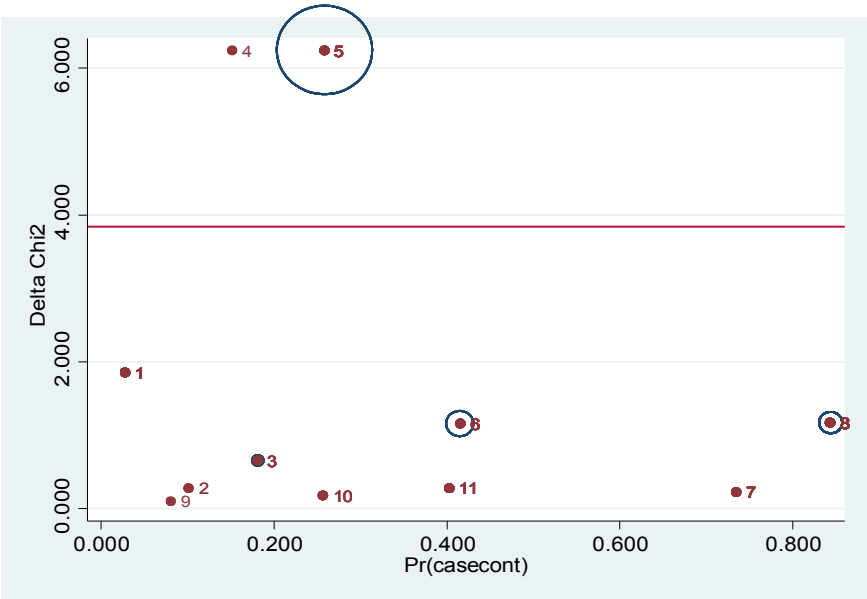
- deltas-betas and delta χ^2
  - values will depend on the predicted probabilities (similar to leverage)[2]

Predicted probabilities	Leverage	$\Delta\chi^2$	$\Delta\beta$
0.0 – 0.1	small	large or small	small
0.1 – 0.3	large	moderate	large
0.3 – 0.7	moderate	moderate	moderate
0.7 – 0.9	large	moderate	large
0.9 – 1.0	small	large or small	small

  - plots
    - $\Delta\beta$ vs predicted probabilities
    - $\Delta\beta$ vs leverage values
    - ΔD vs predicted probabilities with size proportional to $\Delta\beta$

2 Hosmer and Lemeshow. Applied Log. Reg. 2nd Edition. pg-174-176

- Example



- ★ influential observations

```
. l cov herds dcpct dneo dclox pv lev ddev dx2 db if db > abs(1), noobs table

  +-------------------------------------------------------------------------------+
  | cov   herds   dcpct   dneo   dclox      pv     lev    ddev     dx2         db |
  |-------------------------------------------------------------------------------|
  |   3       8     100      0       0   0.182   0.730   0.705   0.642   1.738636 |
  |   8      38     100      1       0   0.844   0.852   1.221   1.167    6.69328 |
  |   6      11      15      1       0   0.416   0.892   1.169   1.152   9.504465 |
  |   5      11     100      0       1   0.259   0.945   6.721   6.232   107.8308 |
  +-------------------------------------------------------------------------------+
```

- leverage not too informative (because all the predictors are categorical)
- dx2 and delta-dev similar to Pearson and deviance residuals
- delta-betas extreme for cov. 5 (same as VER), 6 (not in VER) and 8 (noted in VER to be due to a large group size)

Dealing with influential observations

- identify points with large residuals or large leverage values
- evaluate their covariate patterns – why are they outliers?
- delete from model and re-fit the model
  - does it change very much?

Variable	final	wocov5	wocov6	wocov8
dneo				
yes	3.192***	3.248***	3.639***	2.518*
dclox				
yes	0.453	-2.081**	0.808	0.705
dneo#dclox				
no#yes	(base)	(empty)	(base)	(base)
dneo#dclox				
yes#yes	-2.533*	(omitted)	-3.018*	-2.053
dcpct3				
50	1.361	1.087	0.120	1.173
100	2.027**	2.133**	0.813	1.168
_cons	-3.531***	-3.581***	-2.672*	-2.972**
N	108	96	97	70

legend: * p<.05; ** p<.01; *** p<.001

- cov pattern 5 – only cov. with dneo=0 and dclox=1 case and controls – part of the interaction
- cov pattern 6 – dneo=1 and dclox=0 with 0 dcpct
- cov pattern 8 – largest cov. pattern
- cov pattern 4 – largest contribution to the deviance and Pearson X2 – however no influence (delta-beta = 0.74)

Predictive ability of a logistic model

$$\ln\left(\frac{p}{1\text{-}p}\right)=\beta_0+\beta_1*X_1=X\beta \qquad p=\frac{e^{X*\beta}}{1+e^{-X*\beta}}$$

- note: not a true probability if derived from a case-control study [see l3b]

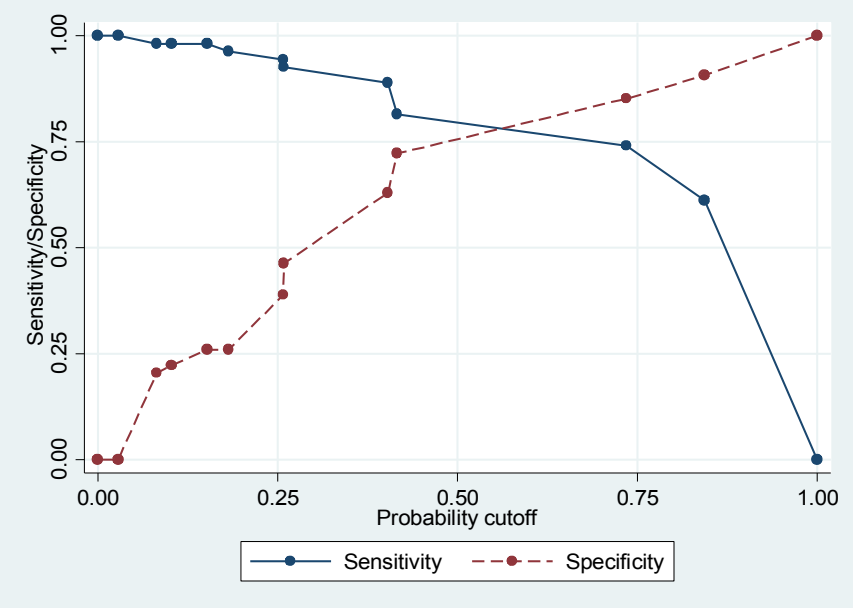
Sensitivity and Specificity

- predict D+ if p >= 0.5
  - choose other cutpoint

Classified (predicted status)	true D+	true D-	Total
T+ = p(D+) >= 0.5	40	8	48
T- = p(D+) < 0.5	14	46	60
Total	54	54	108

- sensitivity (Se) =
- specificity (Sp) =
- positive predictive value (PPV) =
- negative predictive value (NPV) =
- overall correctly classified =

- two-graph ROC (Se-Sp plot)
  - effect of changing the cutpoint on Se and Sp.

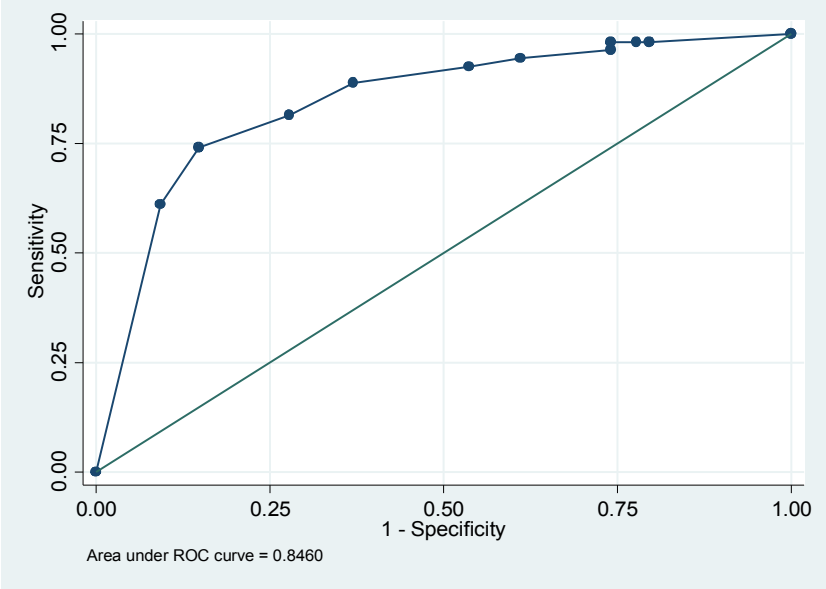


- ROC curve
  - Se vs 1-Sp
  - Area Under the Curve
    - proportion of the time that a subject with y=1 had $\hat{p}_1 > \hat{p}_0$
  - interpretation[1]

AUC	Interpretation
0.5	no discrimination (better flip a coin!)
0.5 - 0.7	good
0.7 - 0.8	very good
>0.9	excellent

1 Adapted from Hosmer Lemeshow. Aplied logistic regression (pg162)

- final model AUC= 0.8460



- Concordant pairs (Minitab/SAS)
  - ☆ total # of pairs of obs. with different outcomes
    - total pairs = n1 * n0 (eg 54 cases * 54 controls) = 2916
    - concordant pairs = (c): $\hat{p}_1 > \hat{p}_0$
    - discordant pairs = (d): $\hat{p}_1 < \hat{p}_0$
    - tied pairs = $\hat{p}_1 = \hat{p}_0$
  - ☆ Area Under the Curve (AUC)
    - AUC = $\dfrac{\text{concordant pairs} + 0.5 * \text{Nbr. ties}}{\text{total pairs}}$
      - concordant pairs = 2322
      - tied pairs = 291
      - total pairs = 2916
      - AUC = (2322+0.5*291)/2916 = 0.8462

Summary logistic regression diagnostics

- covariate pattern residuals
  - goodness-of-fit tests
    - inadequacies in the modelling of the predictors in the model
      - e.g non-linearity or missing interactions
    - can't detect missing predictors or clustering
  - outlying observations (cov. patterns)
- diagnostics
  - consequences of the current model
  - identify high influence cov. patterns (for $\Delta\beta$)on the parameter estimates

Stata code

```
** VHM 812 - Winter 2014
* Evaluating a logistic model

* change working directory and open a log file
cd c:\vhm812\data
capture log close
log using log_reg_dx, text replace
set more off

* open the Nocardia dataset
use nocardia.dta, clear
sum dcpct
tab dcpct
egen dcpct3=cut(dcpct), at(0,50,100,1000)
tab dcpct3

* residuals one per covariate pattern
* fitting a logistic model
logit casecont dneo##dclox i.dcpct3

* examining the covariate patterns
predict cov, num
predict pv, p
sort cov
* generate a count of the number of obs. in each cov. pattern
quietly by cov: gen cnt=_N
br cov cnt dcpct dneo dclox pv casecont

* examining Pearson residuals
predict pear, res
format pv pear  %5.3f
sort pear
summ pear
list cov cnt dcpct dneo dclox pv casecont pear  if abs(pear)>2, noobs sep(4)

* examining Deviance residuals
predict dev, dev /*one per covariate pattern*/
summ dev
sort cov
list cov cnt dcpct dneo dclox pv casecont dev pear  if abs(pear)>1, noobs sep(4)

* Pearson goodness-of-fit tests
**stata post estimation command
estat gof

* Deviance goodness-of-fit tests
logit casecon i.cov, asis nolog
estimates store full
logit casecon dneo##dclox i.dcpct3, nolog
estimates store red
lrtest full red

* Hosmer - Lemeshow Test
estat gof, g(10) table

*summary table - page 8
preserve
collapse pv cnt pear pear_sq dev dev_sq, by(cov casecont)
sort cov casecont
```

```
foreach var in pv cnt pear dev{
by cov:replace `var'=. if _n>1
}
table casecont, by(cov)  c(mean cnt mean pv mean pear mean dev)
restore

* Evaluating Important Observations in a Logistic Model
* fitting a logistic model
logit casecont i.dneo##dclox i.dcpct3
* evaluating outliers
capture drop cov pv
capture drop cnt
capture drop lev
capture drop dev_res
capture drop dev_std
capture drop pear_std
predict cov, num
quietly bysort cov: gen cnt=_N
predict pv, p
predict pear_std, rstandard
predict lev, hat
predict dev_res, deviance
gen dev_std=dev_res/sqrt(1-lev)

sort dev_std
twoway (scatter dev_std cov [aweight=cnt], msymbol(Oh) mlcolor(black)
mlwidth(medium)) scatter dev_std cov, msize(vtiny) mlabel(cov)), legend(off) yline(
-2 2)
bysort cov: gen wcov=_n
bysort cov: egen avgcc=mean(casecont)
list cov cnt dcpct dneo dclox avgcc pv dev_std pear_std if wcov==1 & abs(dev_std)>2
,noobs

* identifying highest leverage points
sort lev
summ lev, d
preserve
collapse (count) herds=casecont (mean) dcpct dneo dclox pv dev_std lev, by(cov)
sort lev
l cov herds dcpct dneo dclox pv dev_std lev if  lev >0.5 | abs(dev_std)>2, noobs
restore

* graph of stand. resid. vs leverage
scatter dev_std  lev , mlabel(cov) yline(-2 2)

* evaluating delta chisq and delta dev
predict dx2, dx2
predict ddev, ddev
scatter ddev pv, mlabel(cov) yline(3.84) /*delta deviance*/
scatter ddev lev, mlabel(cov) yline(3.84) /*delta deviance*/
scatter dx2 pv, mlabel(cov) yline(3.84) /*delta chi2*/

foreach var in pv dev_std db dx2 ddev lev {
format `var' %5.3f
}
preserve
collapse (count) herds=casecont (mean) dcpct dneo dclox pv dev_std dx2 ddev lev ,
by(cov)
sort ddev
l cov herds dcpct dneo dclox pv dev_std dx2 ddev lev if ddev > 3.84 | dx2>3.84,
noobs table
```

```
restore

* evaluating delta betas
predict db, dbeta
sort db
summ db, d
scatter db pv, ml(cov) yline(1)
scatter dx2 pv [aweight=db], msymbol(Oh)  || scatter dx2 pv, ml(cov) yline(3.84)
legend(off) ///
      ytitle("Delta Chi2")


preserve
collapse (count) herds=casecont (mean) dcpct dneo dclox pv lev ddev dx2 db, by(cov)
sort db
l cov herds dcpct dneo dclox pv lev ddev dx2 db if db > abs(1), noobs table
restore



* dropping the highest db covariate pattern and refitting the model
logit casecont dneo##dclox i.dcpct3 if cov~=5
* no interaction cov 5 only cov with dneo=no and dclox=yes with cases and controls
* the other cov with this patterns is 4 but only has one case.
table casecont dneo dclox
table casecont dneo dclox if cov~=5

* refitting and comparing the models
logit casecont dneo##dclox i.dcpct3
estimate store final
* without cov pattern 5
logit casecont dneo##dclox i.dcpct3 if cov~=5
estimates store wocov5
* without cov pattern 6
logit casecont dneo##dclox i.dcpct3 if cov~=6
estimates store wocov6
* without cov pattern 8
logit casecont dneo##dclox i.dcpct3 if cov~=8
estimates store wocov8

estimates table final wocov5 wocov6 wocov8 , b(%5.3f)  stats(N) star( .05 .01 .001)

*Predictive ability of the model
* sensitivity and specificty of logistic model
logit casecont dneo##dclox i.dcpct3
estat class
* two graph ROC
lsens, lpattern(solid dash)
* changing the cutpoint
estat class, cut(0.25)
* standard ROC graph
lroc
```