Epidemiology/Biostats VHM 812/802 Course
Winter 2015, 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
  - mainly for visual assessment
  - not very useful for assessing the model
  - Stata -glm- command

- one per covariate pattern
  - goodness-of-fit tests
  - residual analysis
    - Pearson residuals (standardized)
    - Deviance residuals (not covered in this course)
  - leverage
  - influential observations
    - delta χ^2 ($\Delta\chi^2$)
    - delta-beta ($\Delta\beta$)
  - Stata -logit / logistic - command

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

Pearson residuals per covariate pattern

- Pearson residual
  - $r_j = \dfrac{y_j - m_j * p_j}{\sqrt{m_j * p_j * (1 - p_j)}}$
    - y_j = nbr. pos. outcomes in j^{th} covariate pattern
    - m_j = nbr. obs. in the j^{th} covariate pattern
    - p_j = predicted prob. for the j^{th} 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
  - Note: there are also Deviance residuals that can be computed

Goodness of fit tests

- Pearson 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

- Example

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

$\chi^2=8.22$

  - ★ no indication of lack of fit Pearson χ2 with df=5
  - ★ largest contribution is from cov. pattern with no replication (eg cov # 4)

- Pearson χ2 after logit command

  - ★ -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
```

- Hosmer-Lemeshow Test
  - few replicates per covariate pattern
  - group by:
    - percentiles of estimated probability
    - fixed points of estimated probability
  - compares predicted probabilities to observed probabilities in groups (g) of data
    - 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
```

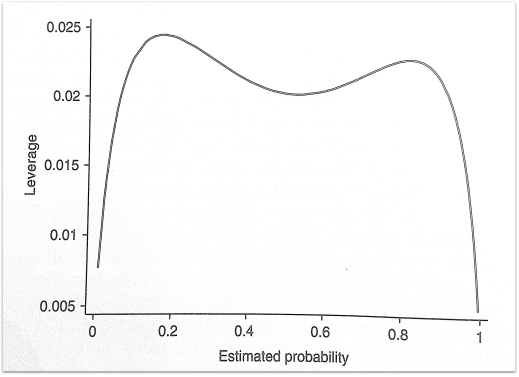
Overdispersion [L11a - L11b]

- 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.

Leverage

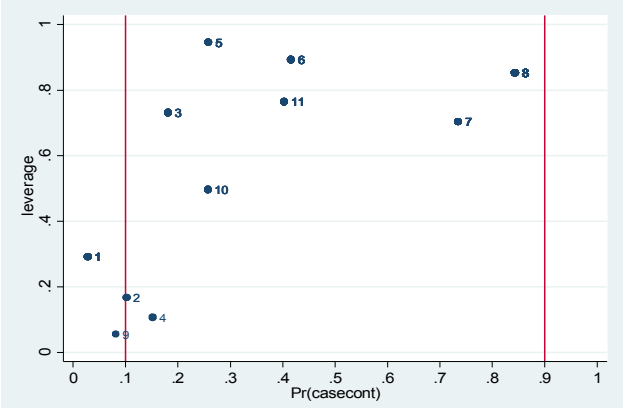
- potential impact of cov. pattern on the model
- extent to which the j^{th} cov. pattern is separated for the others in terms of the explanatory variables
- leverage depends on x's and predicted value

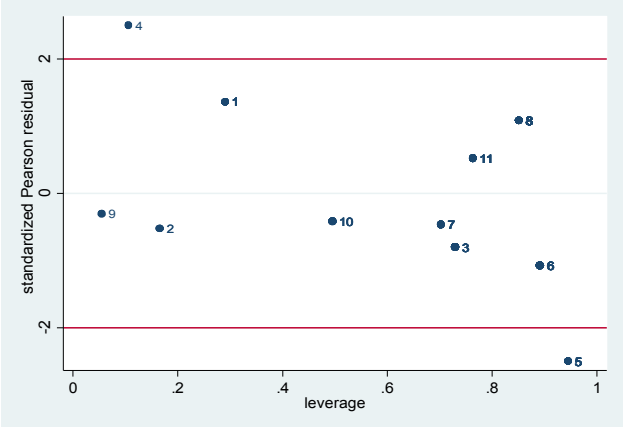
Predicted probabilities	Leverage
0.0 – 0.1	low
0.1 – 0.3	high
0.3 – 0.7	moderate
0.7 – 0.9	high
0.9 – 1.0	low



- If estimated probabilities >0.1 and <0.9 then leverage values can be interpreted as distance
  - look for large leverage values within this range
  - look for points that fall some distance from the rest of the data

- Example



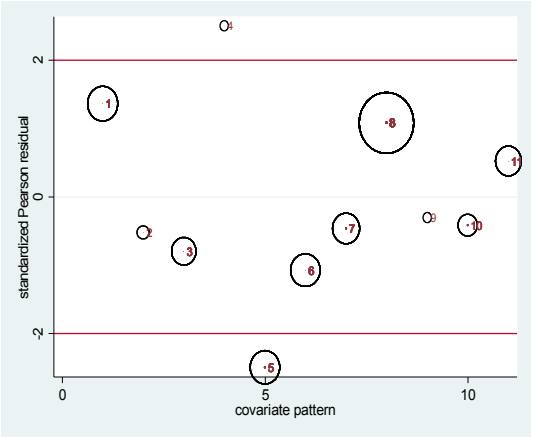


```
. l cov herds dcpct dneo dclox pv pear_std lev if  lev>0.1 & lev<0.9 , noobs

  +-------------------------------------------------------------------------+
  | cov   herds   dcpct   dneo   dclox         pv    pear_std        lev |
  |-------------------------------------------------------------------------|
  |   4       1      83      0       1    .152218    2.496497   .1063734 |
  |   2       2      75      0       0   .1024564   -.5232843   .1662458 |
  |   1      12       6      0       0   .0284367    1.358481   .2907239 |
  |  10       5      74      1       1   .2577896   -.4160336   .4957797 |
  |   7      10      69      1       0   .7353922   -.4654696   .7028956 |
  |-------------------------------------------------------------------------|
  |   3       8     100      0       0   .1817309   -.8012637   .7303169 |
  |  11       9     100      1       1   .4032532    .5182928     .76377 |
  |   8      38     100      1       0   .8439235    1.080066   .8515815 |
  |   6      11      15      1       0   .4160897   -1.073387   .8918833 |
  +-------------------------------------------------------------------------+
```

Residual analysis (covariate patterns)

- Pearson 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. Pearson residuals (per cov. pattern) vs cov. Pattern

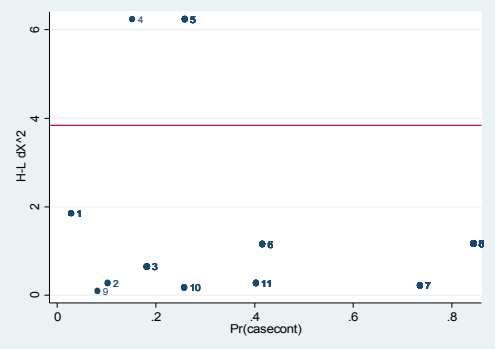


```
. list cov cnt dcpct dneo dclox avgcc pv pear_std if wcov==1 & abs(pear_std)>2
,noobs

  +-------------------------------------------------------------------------+
  | cov   cnt   dcpct   dneo   dclox      avgcc         pv    pear_std |
  |-------------------------------------------------------------------------|
  |   4     1      83     no     yes          1   .152218    2.496497 |
  |   5    11     100     no     yes  .1818182   .2588893   -2.496497 |
  +-------------------------------------------------------------------------+
```

Influential statistics

- delta χ^2 ($\Delta\chi^2$)
  - ★ effect of covariate pattern on Pearson χ^2
    - identifies patterns that do 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)
- Example - delta χ^2 ($\Delta\chi^2$) vs Pr(Pr)




```
. list cov herds dcpct dneo dclox pv dx2 lev if dx2>3.84, noobs table

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

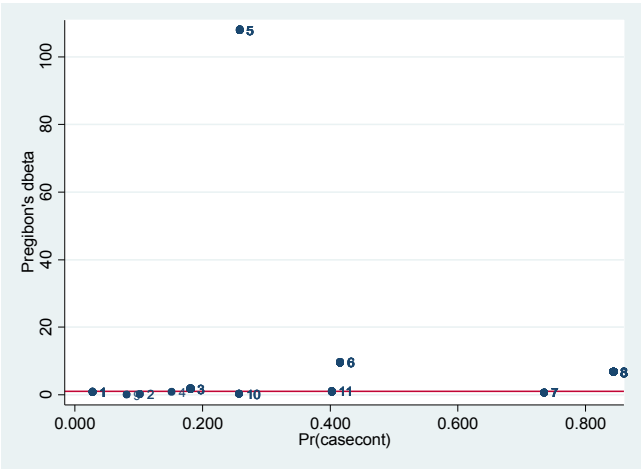
- delta-betas ($\Delta\beta$)
  - analogous to Cook's distance
  - measures influence of a cov. pattern on:
    - overall set of betas (Stata)
    - individual betas (SAS)
  - depends on leverages and # of observations (m_j) on the covariate pattern variable
    - Hosmer & Lemeshow suggest that values >1 might be influential
- Leverage, $\Delta\chi^2$ and $\Delta\beta$
  - 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

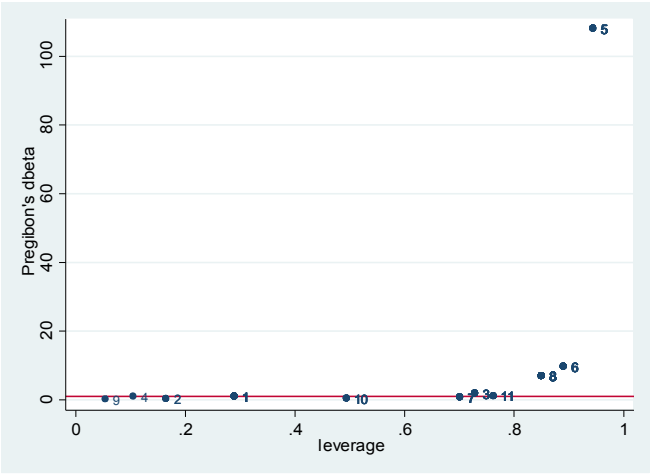
2 Hosmer and Lemeshow. Applied Log. Reg. 2nd Edition. pg-174-176

- Plots

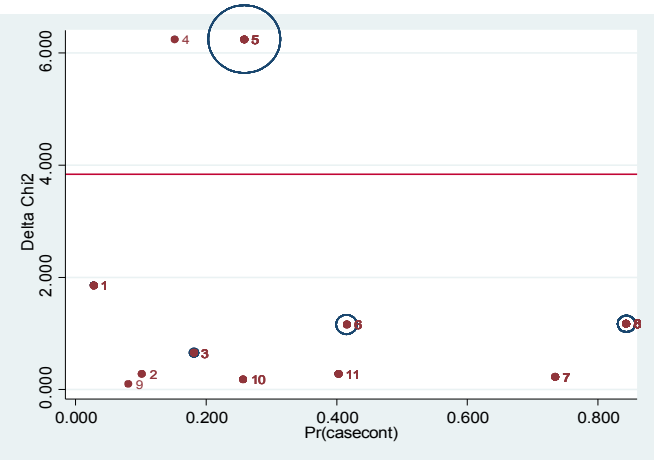
Δβ vs pred. prob.



Δβ vs leverage values



- $\Delta\chi^2$ vs predicted probabilities with size proportional to $\Delta\beta$



- influential observations

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

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

- 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]

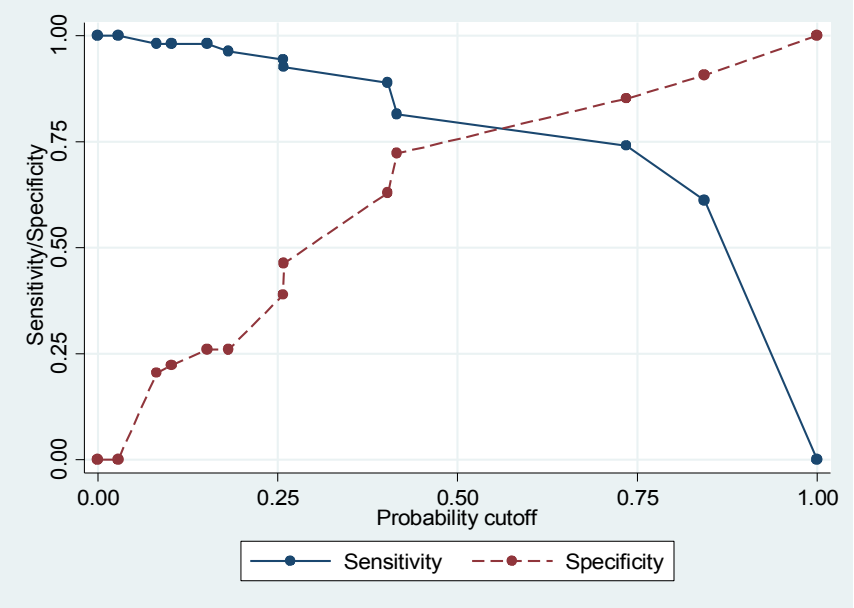
Sensitivity and Specificity

- 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.

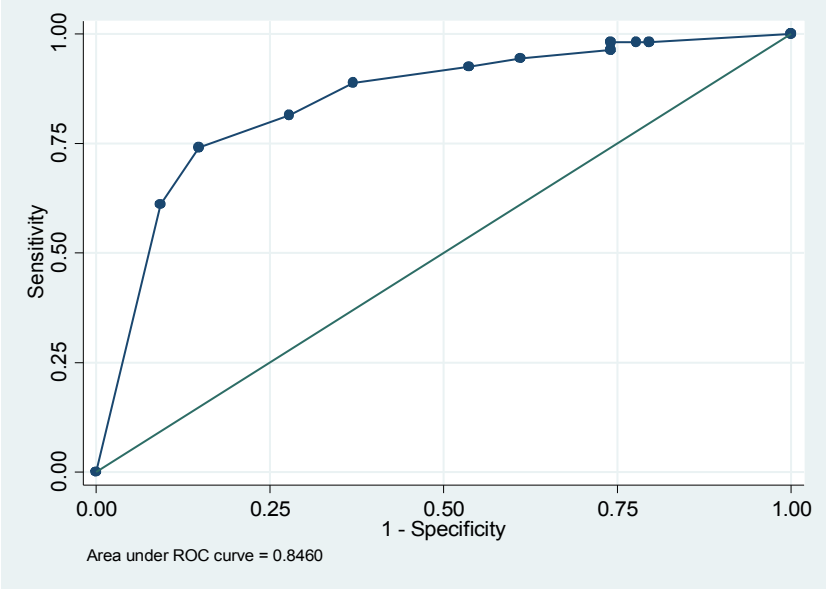


- ROC curve
  - Se vs 1-Sp
  - Area Under the Curve
    - proportion of the time that a subject with y=1 had $\hat{p}_1 > \hat{p}_0$
  - interpretation[1]

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 instance $\Delta\beta$)on the parameter estimates