

## Lecture 5a: Logistic regression diagnostics (VER/MER 16.12)

<b>Index</b>	<b>Page</b>
Covariate patterns.....	2
Pearson residuals per covariate pattern.....	4
Goodness of fit tests.....	5
Residual analysis (covariate patterns).....	7
Leverage (h).....	8
Influential statistics .....	9
Dealing with influential observations.....	11
Summary logistic regression diagnostics.....	12
Predictive ability of a logistic model.....	13
Overdispersion [L11a - L11b].....	16

### Today:

- ★ logistic regression: model diagnostics
- ★ quiz #3

### Home work for Tuesday:

- ★ logistic regression exercises 2 and 3

### Dataset: nocardia.dta

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

## Covariate patterns

- Unique combination of values of predictor variables
- Categorical predictors
  - ★ dcpct -> converted to categorical (dcpct3= 0-49, 50-99, >=100)
  - ★ 11 covariate patterns

Dcpct_3	Dneo	Dclox	Covariate Pattern	Nbr obs
0	no	no	1	12
0	yes	no	2	11
0	Yes	Yes	3	1
50	no	no	4	2
50	no	yes	5	1
50	yes	no	6	10
50	yes	yes	7	5
100	no	no	8	8

- Continuous predictors
  - ★ production and SCC
  - ★ 99 covariate patterns

Prod (l/day)	SCC ('000s)	Covariate Pattern	Nbr obs
11.4	182	1	1
12.1	277	2	1
.	.	.	1
34.5	61	99	1

## Residuals in logistic regression

- One per observation (based on Hilbe, 2009<sup>1</sup>)
  - ★ (standard) residual and influential analyses
  - ★ mainly for visual assessment
  - ★ not very useful for assessing the model
  - ★ Stata command *-glm-*
    - ➔ Pearson
    - ➔ Anscombe (more normal distributed )
  
- One per covariate pattern
  - ★ goodness-of-fit tests
  - ★ residual analysis
    - ➔ Pearson residuals (standardized)
    - ➔ deviance residuals (not covered in this course)
  - ★ Stata command *-logit / logistic -*

### ● Example - residuals

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

<sup>1</sup> Hilbe J. Logistic Reg. Models. CRC Press: Boca Raton 2009

## Pearson residuals per covariate pattern

- Pearson residual (standardized)

- ★ 
$$r_j = \frac{y_j - m_j * p_j}{\sqrt{m_j * p_j * (1 - p_j)}}$$

- $y_j$  = nbr. pos. outcomes in  $j^{\text{th}}$  covariate pattern

- $m_j$  = nbr. obs. in the  $j^{\text{th}}$  covariate pattern

- $p_j$  = predicted prob. for the  $j^{\text{th}}$  covariate pattern

- ★ standardized residuals

- $$r_{sj} = \frac{r_j}{\sqrt{(1 - h_j)}} \quad ; \text{ makes variance to 1}$$

- $h_j$  = leverage (next slides)

- ★ 
$$\sum_{j=1}^J (r_j)^2 = \text{Pearson } \chi^2 \text{ statistic} \sim \chi^2 \text{ with } (J - k) \text{ df.}$$

- $k$  = number of parameters in the model

- ★ contribution of each cov. pattern to the  $\chi^2$  statistic

- ★ Stata command – *predict* after *logit/logistic* -

- eg. `logit casecon i.dneo`

- `predict pear, rresidual`

- `predict stdpear, rstandard`

## Goodness of fit tests

- Pearson  $\chi^2$  (1 per cov. Pattern) (see L3a-p9)
  - ★  $\chi^2$  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
      - at least 1 expected count in each cell
      - at least 80% expected values  $\geq 5$  counts
  - ★ indicates the fit of the model (Ho = model fits the data)

### ● Example: nocardia

cov pat and Case		Nbr herds	Nbr pos	Pred. pr.	Pear. Res.	Pear. Res. <sup>2</sup>
- Control						
1	no yes	12	1	0.028	1.144	1.308949
2	no yes	2	0	0.102	-0.478	.2283039
3	no yes	8	1	0.182	-0.416	.1731429
4	no yes	1	1	0.152	2.360	5.569528
5	no yes	11	2	0.259	-0.584	.3405482
6	no yes	11	4	0.416	-0.353	.1245677
7	no yes	10	7	0.735	-0.254	.0643712
8	no yes	38	33	0.844	0.416	.1731366
9	no yes	1	0	0.082	-0.298	.0890559
10	no yes	5	1	0.258	-0.295	.0872724
11	no yes	9	4	0.403	0.252	.0634578

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

- Pearson  $\chi^2$  after logit command

- ★ p-value is meaningful if enough replicates

- ★ Stata command (after logit)

- ➔ *estat gof*

```
. 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

- ★ the only useful test when there are few replicates per covariate pattern

- ★ group by:

- ➔ percentiles of predicted probabilities

- ➔ fixed points of predicted probabilities

- ★ compares predicted probabilities to observed probabilities in groups (g) of data

- ➔  $\chi^2$  with g-2 df, where “g” = number of groups

- ➔ low power if < 6 groups

- ★ Stata command (after logit)

- ➔ *estat gof, group(#) table*

## ● 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

```

## Residual analysis (covariate patterns)

### ● Pearson residuals

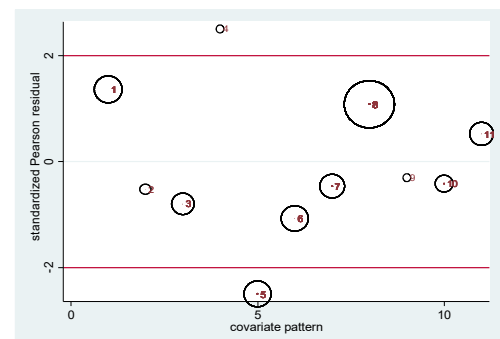
★ identify large negative and positive residuals

★ visual assessment

➔ some guidelines about outlying obs.

### ● Example

★ stdz. Pearson residuals (per cov. pattern) vs cov. pattern



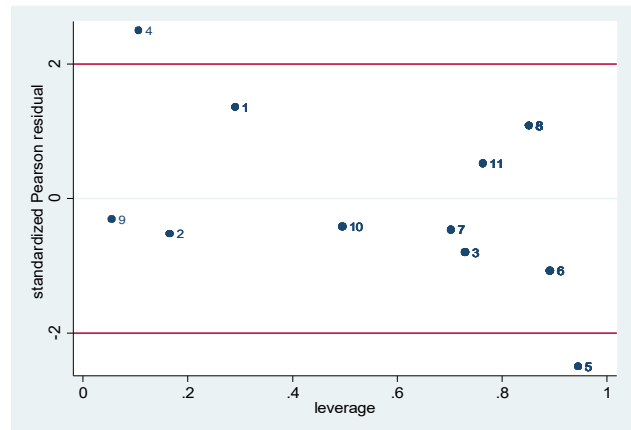
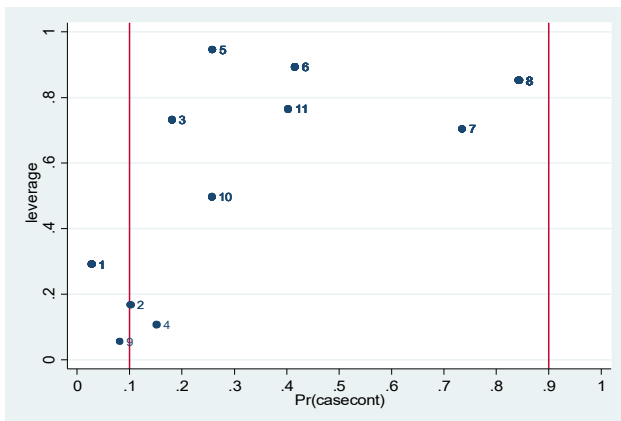
```
. list cov cnt dcpct3 dneo dclox opr pv pear_std if wcov==1 & abs(pear_std)>2, noobs
```

cov	cnt	dcpct3	dneo	dclox	opr	pv	pear_std
4	1	50	no	yes	1.000	0.152	2.496
5	11	100	no	yes	0.182	0.259	-2.496

## Leverage ( $h$ )

- Potential impact of covariate. pattern on the model
- Extent to which the  $j^{\text{th}}$  cov. pattern is separated for the others in terms of the explanatory variables
- One value per covariate pattern (Stata command `-logit / logistic -`)
- Leverage depends on x's and predicted probabilities
- ★ if predicted probabilities  $<0.1$  and  $>0.9$  then leverage values will be low
- ★ if predicted 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

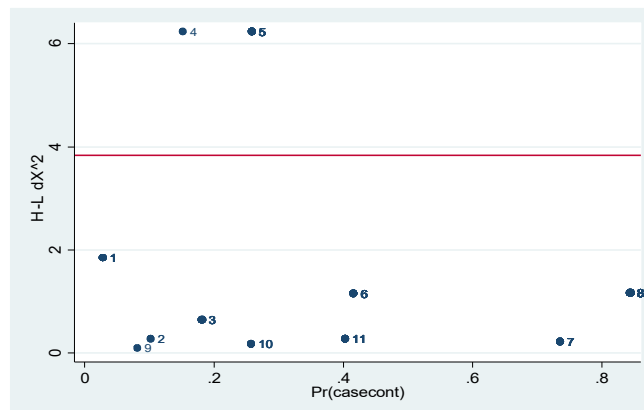
### ● Example



cov	cnt	dcpct3	dneo	dclox	opr	pv	pear_std	lev
4	1	50	no	yes	1.000	0.152	2.496	0.106
...								
11	9	100	yes	yes	0.444	0.403	0.518	0.764
8	38	100	yes	no	0.868	0.844	1.080	0.852
6	11	0	yes	no	0.364	0.416	-1.073	0.892
5	11	100	no	yes	0.182	0.259	-2.496	0.945

## Influential statistics

- ★ delta  $\chi^2$  (  $\Delta \chi^2$  )
- ★ delta-beta (  $\Delta\beta$  ) (Cook's distance in normal regression)
- ★ one per covariate pattern (Stata command *-logit / logistic -* )
- Delta  $\chi^2$  (  $\Delta \chi^2$  ) (or  $r_{sj}^2$ )
  - ★ effect of a covariate pattern on fit of the model
    - ➔ identifies patterns that do not fit well (outliers)
  - ★ plot delta values vs predicted probabilities
  - ★ plot delta values vs leverage
  - ★ delta-values  $\geq 3.84$  (95<sup>th</sup> percentiles  $X^2$  distribution with 1df)
- Example - delta  $\chi^2$  (  $\Delta \chi^2$  ) vs Pr(Pr)



```
. list cov cnt dcpct3 dneo dclox pv dx2 lev pear if dx2>3.84 & wcov==1, noobs
```

cov	cnt	dcpct3	dneo	dclox	pv	dx2	lev	pear
5	11	100	no	yes	.2588893	6.232499	.9453593	-0.584
4	1	50	no	yes	.152218	6.232499	.1063734	2.360

- Delta-beta (  $\Delta\beta$  )

- ★ 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<sup>2</sup> suggest that values >1 might be influential

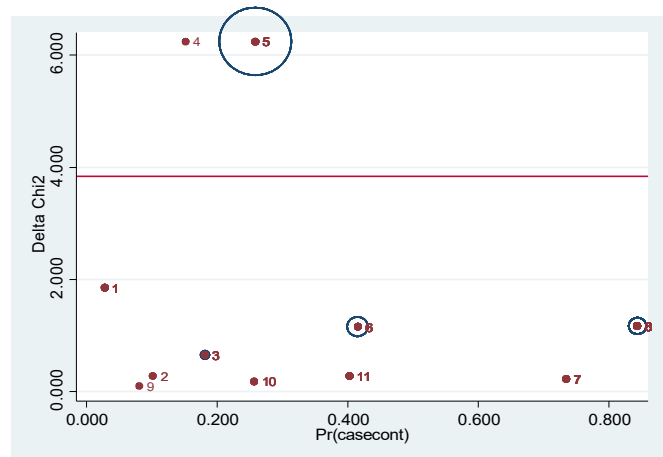
- Delta  $\chi^2$  and Delta-beta

- ★ values will depend on the predicted probabilities (similar to leverage)<sup>2</sup>

- Plots

- ★  $\Delta\beta$  vs pred. prob. and  $\Delta\beta$  vs leverage values

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



---

2 Hosmer and Lemeshow. Applied Log. Reg. 2<sup>nd</sup> Edition. pg-174-176

★ influential observations

```
. l cov cnt dcpct dneo dclox opr pv lev dx2 db if db > abs(1) & wcov==1, noobs
```

	cov	cnt	dcpct	dneo	dclox	opr	pv	lev	dx2	db
	3	8	100	no	no	0.125	0.182	0.730	0.642	1.739
	8	38	100	yes	no	0.868	0.844	0.852	1.167	6.693
	6	11	5	yes	no	0.364	0.416	0.892	1.152	9.504
	5	11	100	no	yes	0.182	0.259	0.945	6.232	107.831

★ 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?

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

Variable	final	wocov5	wocov6	wocov8
dneo				
yes	3.192***	3.248***	3.639***	2.518*
dclox				
yes	0.453	17.322	0.808	0.705
dneo#dclox				
yes#yes	-2.533*	-19.403	-3.018*	-2.053
dcpct3				
50	1.361	1.087	0.120	1.173
100	2.027**	2.132**	0.813	1.168
_cons	-3.531***	-3.581***	-2.672*	-2.972**
N	108	97	97	70

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

- ★ cp 5 – only cov. with dneo=0 and dclox=1 – part of the interaction
- ★ cp 6 – dneo=1 and dclox=0 with 0 dcpct
- ★ cp 8 – largest cov. pattern
- ★ cp 4 – largest contribution to the deviance and Pearson X<sup>2</sup> – however no influence (delta-beta = 0.74)

## 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

- ➔ eg. few replicates?

- ➔ increase the nbr of obs. per cov. pattern (or reduce the nbr. of cov. patterns)

- grouping (biological, frequency categories)

- remove continuous predictors

- eg. those that are borderline sig.

- ➔ however... you will be working with a different model – not your final model!

- ★ identify high influence cov. patterns (for instance  $\Delta\beta$ ) on the parameter estimates

## Predictive ability of a logistic model

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

★ note: not a true probability if derived from a case-control study [see 13b]

## Sensitivity and Specificity

- Predict D+ if  $p \geq 0.5$

★ choose other cutpoint

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

★ sensitivity (Se) =

★ specificity (Sp) =

★ positive predictive value (PPV) =

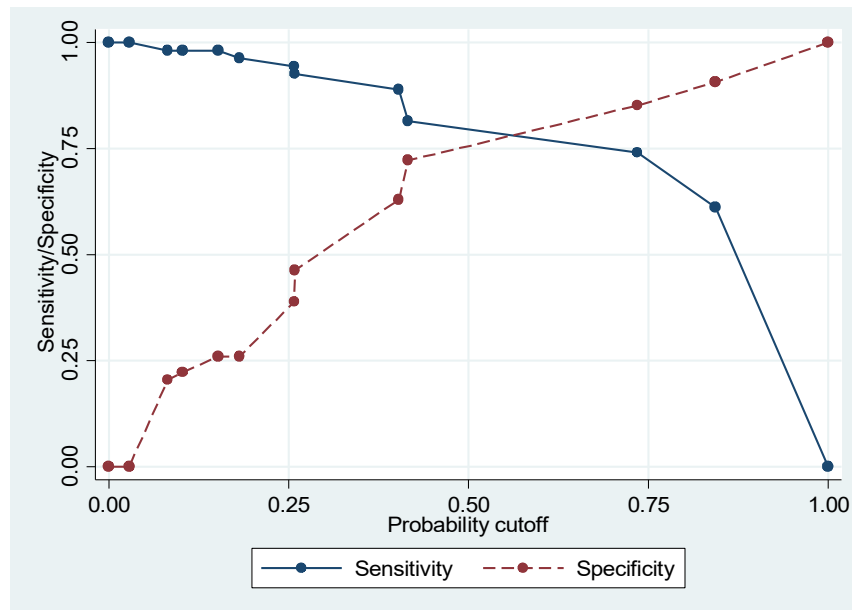
★ negative predictive value (NPV) =

★ overall correctly classified =

## Two-graph ROC (Se-Sp plot)

★ effect of changing the cut-point on Se and Sp.

➔ Se / Sp values corresponds the values  $\geq$  than the cutoff point where data is available (eg. for a cut-off  $\geq 0.5$ , the next value is 0.74)



```
. egen pv_cat=cut(pv), at(0(.05)1)
. roctab casecont pv_cat, graph sum detail scheme(s1mono)
```

Detailed report of sensitivity and specificity

Cutpoint	Sensitivity	Specificity	Correctly Classified	LR+	LR-
( $\geq 0$ )	100.00%	0.00%	50.00%	1.0000	
( $\geq .05$ )	98.15%	20.37%	59.26%	1.2326	0.0909
( $\geq .1$ )	98.15%	22.22%	60.19%	1.2619	0.0833
( $\geq .15$ )	98.15%	25.93%	62.04%	1.3250	0.0714
( $\geq .25$ )	94.44%	38.89%	66.67%	1.5455	0.1429
( $\geq .4$ )	88.89%	62.96%	75.93%	2.4000	0.1765
( $\geq .7$ )	74.07%	85.19%	79.63%	5.0000	0.3043
( $\geq .8$ )	61.11%	90.74%	75.93%	6.6000	0.4286
( $> .8$ )	0.00%	100.00%	50.00%		1.0000

- ROC curve

- ★ Se vs 1-Sp

- ★ assess discriminatory power of the model

- ➔ eg. probability that cases (or controls) are correctly classified by the model (at a given cutpoint)

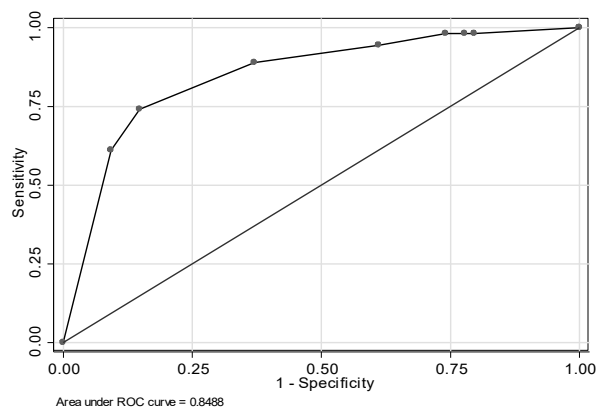
- ★ quantify Area Under the Curve (AUC)

- ★ AUC interpretation<sup>1</sup>

AUC	Interpretation
0.5	No discrimination (better flip a coin!)
0.5 - 0.7	Poor discrimination
0.7 - 0.8	Acceptable discrimination
>0.8	Excellent discrimination

- AUC = final model AUC= 0.85 (see commands do-file)

Obs	ROC Area	Std. Err.	-Asymptotic Normal-- [95% Conf. Interval]	
108	0.8488	0.0370	0.77621	0.92132



- ★ probability that a randomly selected individual is correctly classified is 85%

<sup>1</sup> Hosmer Lemeshow. Applied logistic regression (pg177)

## Overdispersion [L11a - L11b]

- Assumption  $y_i \sim$  binomial distribution
  - ★ mean =  $n_i * p_i$
  - ★ variance =  $n_i * p_i * (1 - p_i)$
- Overdispersion = the data are more dispersed (larger variance) than would be expected
  - ★ apparent overdispersion – wrong model
    - ➔ missing important predictors
    - ➔ outliers
  - ★ real overdispersion – usually due to clustering
  - ★ too small S.E.