

Lecture 5a: Logistic regression diagnostics

Index	Page
Covariate patterns.....	2
Pearson residuals per covariate pattern.....	4
Goodness of fit tests.....	5
Overdispersion [L11a - L11b].....	7
Leverage.....	8
Residual analysis (covariate patterns).....	9
Influential statistics.....	10
Dealing with influential observations.....	13
Summary logistic regression diagnostics.....	14
Predictive ability of a logistic model.....	15

Today:

- ★ logistic regression: model building
- ★ logistic regression: model diagnostics

Tomorrow:

- ★ quiz #3

Friday:

- ★ lab: review model building ex and logistic regression exercises

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 predictor variables
 - ★ e.g. Sex = 1 and Breed = 1 -> covariate pattern = 1
 - ★ e.g. Sex = 1 and Breed = 2 -> covariate pattern = 2
- Example: dichotomous (categorical) predictors

Dneo	Dclox	Covariate Pattern	# Positives (cases)	N	Proportion
0	0	1	6	10	0.60
0	1	2	3	20	0.15
1	0	3	5	50	0.10
1	0	4	4	10	0.40

- Example: continuous predictors

Age (years)	Weight (kg)	Covariate Pattern	# Positives (cases)	N	Proportion
4.3	527.2	1	0	1	0.00
3.7	489.6	2	1	1	1.00
2.1	535.4	3	1	1	1.00
5.6	501.4	4	0	1	0.00

Residuals in logistic regression

- One per observation (based on Hilbe, 2009¹)
 - ★ (standard) residual and influential analyses
 - ★ 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 ($\Delta \beta$) (Cook's distance in normal regression)
 - ★ Stata command *-logit / logistic -*

● Example - residuals

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

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

Pearson residuals per covariate pattern

- Pearson residual

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

- 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)

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

- k = number of parameters in the model

- ★ contribution of each cov. pattern to the χ^2 statistic

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

- eg. `logit casecon i.dneo`

- `predict pear, rresidual`

- `predict stdpear, rstandard`

- Other residuals

- ★ deviance

- ★ Anscombe (more normal distributed - glm)

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 ~ χ^2 -statistic

- at least 1 expected count in each cell

- at least 80% expected values ≥ 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. ²
- 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 χ^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

- ★ 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
 - χ^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
```

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) that would be expected

- ★ apparent overdispersion – wrong model

- missing important predictors

- outliers

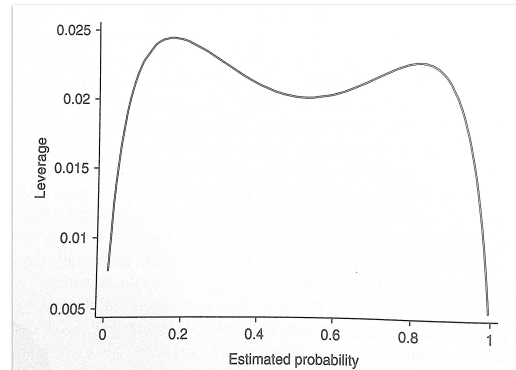
- ★ real overdispersion – usually due to clustering

- ★ too 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

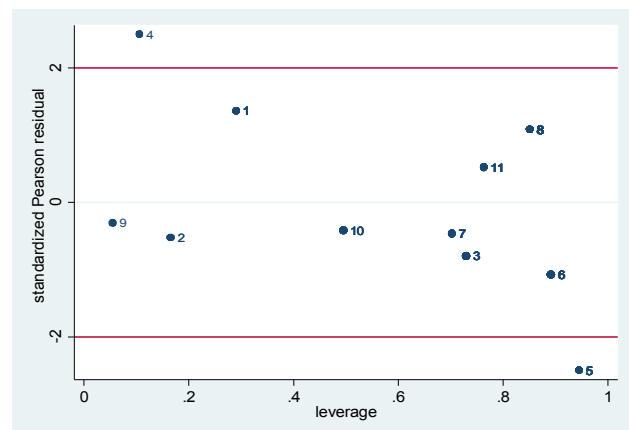
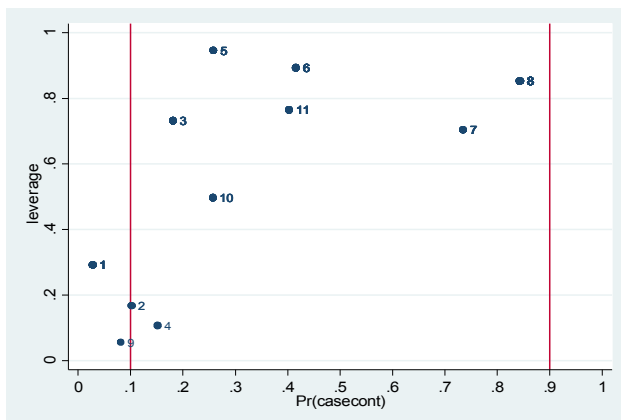
Predicted probabilities	Leverage
0.0 - 0.1	low
0.1 - 0.3	high
0.3 - 0.7	moderate
0.7 - 0.9	high
0.9 - 1.0	low



★ 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



```
. list cov cnt dcpct3 dneo dclox opr pv pear_std lev if pv>0.1 & pv<0.9 & wcov==1, noobs
```

cov	cnt	dcpct3	dneo	dclox	opr	pv	pear_std	lev
2	2	50	no	no	0.000	0.102	-0.523	0.166
3	8	100	no	no	0.125	0.182	-0.801	0.730
4	1	50	no	yes	1.000	0.152	2.496	0.106
5	11	100	no	yes	0.182	0.259	-2.496	0.945
6	11	0	yes	no	0.364	0.416	-1.073	0.892
7	10	50	yes	no	0.700	0.735	-0.465	0.703
8	38	100	yes	no	0.868	0.844	1.080	0.852
10	5	50	yes	yes	0.200	0.258	-0.416	0.496
11	9	100	yes	yes	0.444	0.403	0.518	0.764

Residual analysis (covariate patterns)

- Pearson residuals

- ★ standardized residuals

- ➔ 95% between -2 and +2

- ★ identify large negative and positive residuals

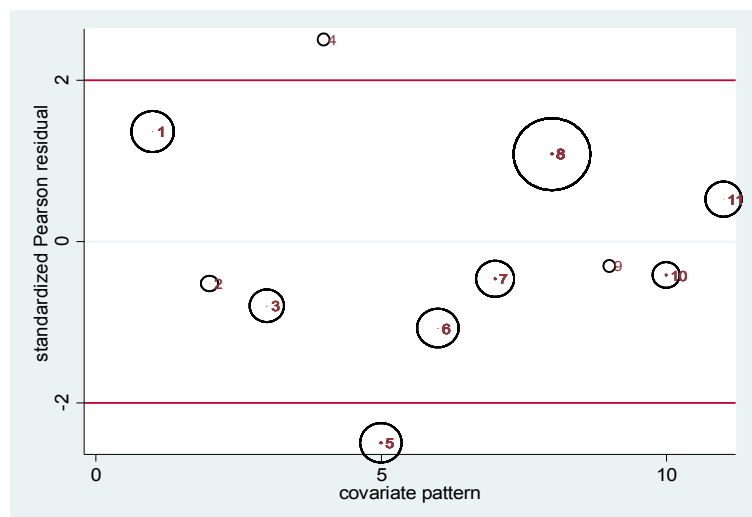
- ★ characteristics of observations

- ★ visual assessment

- ➔ some guidelines about outlying obs.

- Example

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

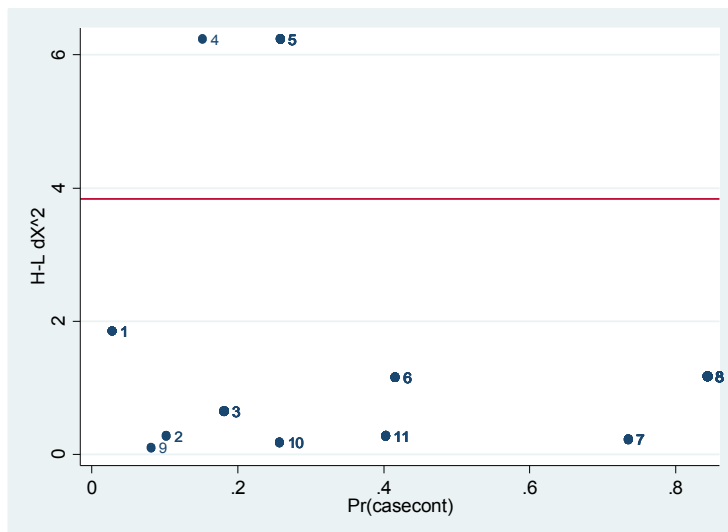


```
. list cov cnt dcpct3 dneo dclox opr pv pear_std if wcov==1 & abs(pear_std)>2, noobs
```

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

Influential statistics

- Delta χ^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 ≥ 3.84 (95th percentiles X^2 distribution with 1df)
- Example - delta χ^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-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 & Lemeshow² 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

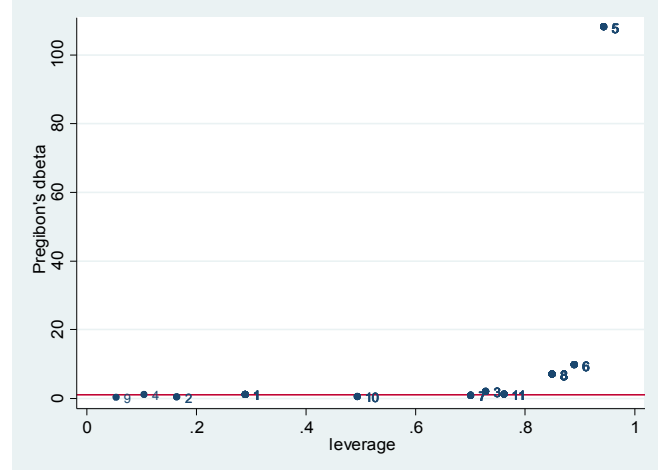
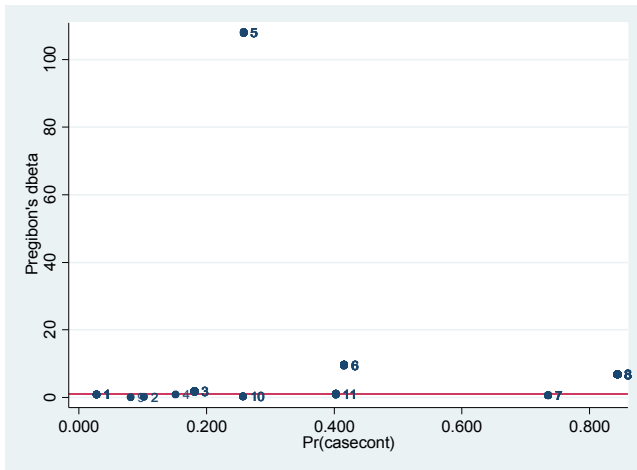
Predicted probabilities	Leverage	$\Delta\chi^2$	$\Delta\beta$
0.0 - 0.1	small	large or small	small
0.1 - 0.3	large	moderate	large
0.3 - 0.7	moderate	moderate	moderate
0.7 - 0.9	large	moderate	large
0.9 - 1.0	small	large or small	small

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

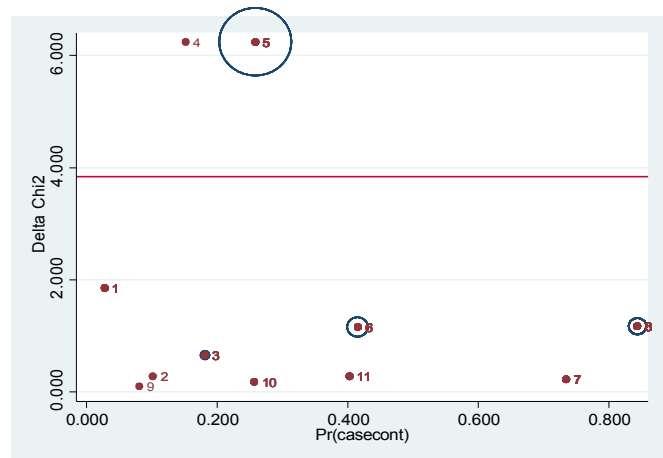
● Plots

$\Delta\beta$ vs pred. prob.

$\Delta\beta$ vs leverage values



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



★ 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 case and controls – 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^2 – 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) =

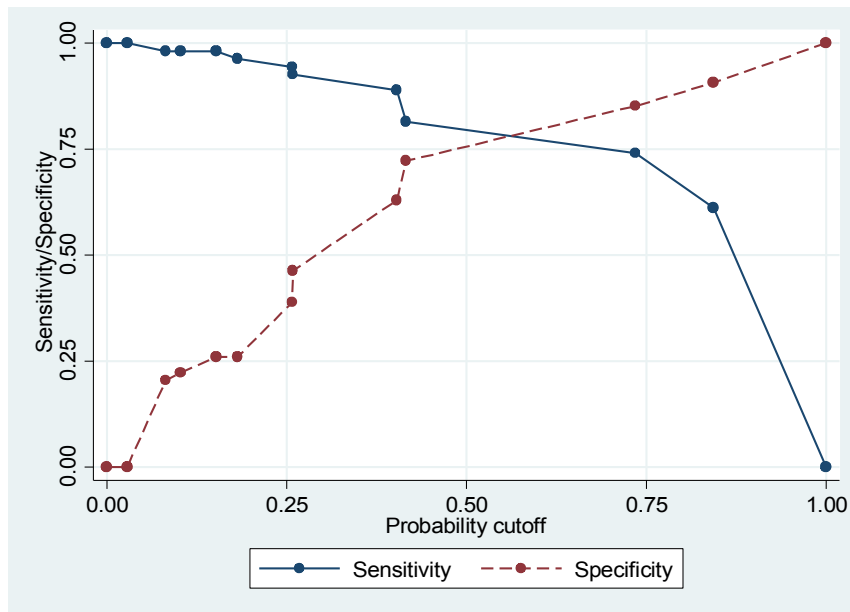
★ positive predictive value (PPV) =

★ negative predictive value (NPV) =

★ overall correctly classified =

● Two-graph ROC (Se-Sp plot)

★ effect of changing the cutpoint on Se and Sp.



```
. 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-
(>= 0)	100.00%	0.00%	50.00%	1.0000	
(>= .05)	98.15%	20.37%	59.26%	1.2326	0.0909
(>= .1)	98.15%	22.22%	60.19%	1.2619	0.0833
(>= .15)	98.15%	25.93%	62.04%	1.3250	0.0714
(>= .25)	94.44%	38.89%	66.67%	1.5455	0.1429
(>= .4)	88.89%	62.96%	75.93%	2.4000	0.1765
(>= .7)	74.07%	85.19%	79.63%	5.0000	0.3043
(>= .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)

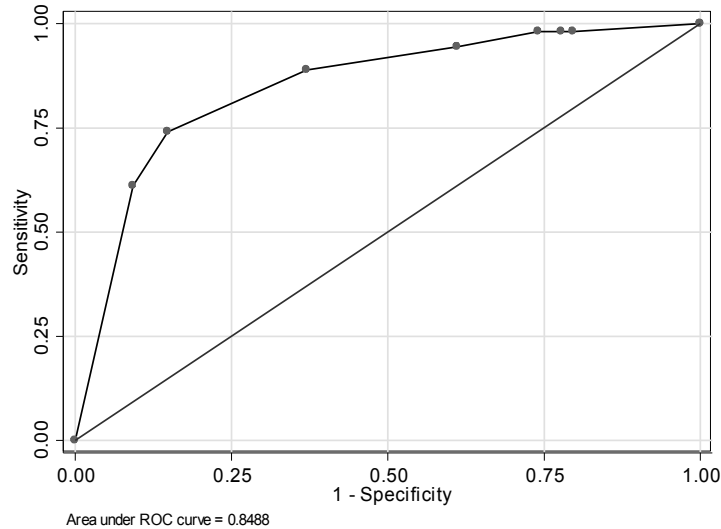
- ★ quantify Area Under the Curve (AUC)

- ★ AUC interpretation¹

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

- 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



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

- Concordant pairs (Minitab/SAS)

- ★ total # of pairs of obs. with different outcomes

- total pairs = $n_1 * n_0$ (eg 54 cases * 54 controls) = 2916

- concordant pairs = $\hat{p}_1 > \hat{p}_0$

- discordant pairs = $\hat{p}_1 < \hat{p}_0$

- tied pairs = $\hat{p}_1 = \hat{p}_0$

- ★ Area Under the Curve (AUC)

- $$\text{AUC} = \frac{\text{concordant pairs} + 0.5 * \text{Nbr. ties}}{\text{total pairs}}$$

- concordant pairs = 2322

- tied pairs = 291

- total pairs = 2916

- $\text{AUC} = (2322 + 0.5 * 291) / 2916 = 0.8462$

- ★ interpretation:

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