Index of Lecture 5a: Logistic regression model validation

Practical Information

News/Schedule:

- VHM 802 logistic regression assignment posted, due February 17,
- last (shared) regression session on Friday:
  - catch-up on lecture material, and time for questions,
  - third logistic regression exercise (VER 16.3).

Today's session:

- second logistic regression exercise (VER 16.2): review/discussion,
- catch-up from previous lecture (almost all of it!), likehood-based inference and prediction in multiple logistic regression,
- new material on logistic regression model validation:[1]
  - residuals and diagnostics: some major differences to linear models,
  - goodness-of-fit tests (additions to 3aL−9),
  - predictive ability for logistic regression models,
  - reliability — implementation for non-normal models.

[1] VER Section 16.12, except Section 16.12.4 on overdispersion (postponed → Session 10a).

Covariate Patterns

In any multivariable (e.g., linear or logistic) model, covariate patterns[2] are the unique combinations of predictor values.

Why are covariate patterns important? — observations within a covariate pattern are i.i.d. (or replications),

- can be grouped into an aggregate outcome (e.g., binary → binomial outcome), which can facilitate model fitting and model diagnostics,
- in linear models, their variance has minimal assumptions (→ error variance).

Example: Nocordia data and model with predictors `dcpct`, `dneo` and `dclox`: a total of 30 covariate patterns, some shown in table below:

Pattern (j)	Positive	Negative	Total (m_j)	`dcpct`	`dneo`	`dclox`
1	0	7	7	0	no	no
2	0	1	1	1	no	no
3	1	0	1	1	yes	no
...	...	...	...	...	...	...
27	1	7	8	100	no	no
28	2	9	11	100	no	yes
29	33	5	38	100	yes	no
30	4	5	9	100	yes	yes

[2] The term "covariate pattern" is standard, but it could equally be called "predictor pattern".

Pearson Residuals

Fact: residuals (and diagnostics) for logistic regression models are available in two forms:

- per covariate pattern (available in Stata after `logit`),
- per individual observation.[3]

Pearson residuals: closest analog of residuals in linear models, defined as "observed minus expected" normalized by "expected" across both outcomes,

$$\text{covariate pattern } j \;:\; r_j = \frac{\text{pos}_j - m_j\hat{p}_j}{\sqrt{m_j\hat{p}_j(1-\hat{p}_j)}},$$

$$\text{individual observation } i \;:\; r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}.$$

Example: covariate pattern #9 in Nocardia model (VER 16.6)

Pattern	Outcome	Total	Predictors			Prop.	Pred.	Pearson residual	
j	y_i	m_j	dcpct	dneo	dclox	pos_j/m_j	$\hat{p}_j$	r_j	r_i
9	0	2	20	yes	no	0.5	0.465	0.10	-0.93
9	1	2	20	yes	no	0.5	0.465	0.10	1.07

- covariate pattern residual ≈ 0 — seems reasonable,
- individual residuals opposite and away from 0 — seems less relevant.

[3] Only available in Stata when the logistic model is fitted as a generalised linear model, `glm` command.

Goodness-of-Fit Tests (continued)

Synthesis: multiple tests exist, and in different versions ⇒ guidance on interpretation and choice of test needed:

- aim: overall assessment of adequacy of model equation (linear, logistic or GLM scale),
- test power is variable and strongly dependent on sample size (small n ∼ little power, large n ∼ (too) much power),
- Pearson chi-square test needs replication (similar guidelines as for Pearson X^2 for two-way tables; VHM 801), and is often useless when based on individual residuals,
- Hosmer-Lemeshow test is the only valid test in absence of reasonable replication.

Pearson chi-square test for model with k parameters (incl. intercept): [4] [5]

$$X^2 = \sum_j r_j^2 \quad \sim \chi^2(J-k), \quad \text{where } j \sim \text{covariate patterns (a total of } J).$$

Deviance/likelihood-ratio test of model with k parameters (incl. intercept) against "saturated model" with one parameter per covariate pattern, $D \sim \chi^2(J-k)$. [6]

Hosmer-Lemeshow chi-square test: based on G groups constructed from predicted probabilities and $\sim \chi^2(G-2)$; example → next slide.

[4] Test for mice data (slides 3aL−6/9) with $k=2$ and $J=12$.
[5] Version for individual Pearson residuals: $X^2_{\text{ind}} = \sum_i r_i^2 \sim \chi^2(n-k)$, but rarely has enough replication to be useful.
[6] The deviance (test) is not meaningful when calculated from binary data, e.g. with Stata's `glm` command.

Hosmer-Lemeshow Test: Details

Construction of groups and test statistic:

- recommended $G = 10$ groups (unless sample size is very large[7]),
- algorithm attemps to create G groups of approximately equal size
  — caution: $G < 10$ may result, and the test has very low power if $G < 6$,
- test uses similar formula as for Pearson chi-square test,

$$\text{Hosmer-Lemeshow statistic} = \sum_g (\text{pos}_g - e_g)^2 / [e_g(m_g - e_g)/m_g],$$

  where the expected count e_g for group g equals $\sum_i \hat{p}_i$ for observations i in group g.

Example: Nocardia mastitis model (VER Example 16.7),

Group (g)	Prob. range	Positive (pos_g)	Total (m_g)	Expected (e_g)	Contribution
1	0 – 0.039	1	11	0.27	1.99
2	0.039 – 0.181	2	12	2.18	0.02
3	0.181 – 0.256	3	12	3.00	0.00
4	0.256 – 0.383	6	7	2.51	1.42
5	0.383 – 0.412	4	10	3.92	0.00
6	0.412 – 0.751	8	14	8.48	0.07
7	0.751 – 0.842	35	40	33.64	0.35
					H-L = 3.85 $\sim \chi^2(5)$

∗ H-L test clearly non-significant; largest contributions from groups 1 and 4.

[7] Modifications have been proposed, see overview in e.g. Nattino et al. (2020), *Biometrics* 76, 549-560.

Predictive Ability of Logistic Regression Models

Synthesis: R^2-type statistics not helpful and need to look for methods adapted to binary outcomes:

- classification tables from predictions ($\hat{p}$): Se, Sp, % correctly classified — all based on chosen cutoff[8] for "positive",
- ROC(Receiver Operating Characteristics)-curves from predictions ($\hat{p}$) (next slide).

— splits of data into estimation/assessment parts also relevant here → reliability.

Illustration: classification table for Nocardia model (cutoff = 0.5):

Predicted status	Observed status: case	control	total
case	41	9	50
control	13	45	58
total	54	54	108

- sensitivity = **0.759**,
- specificity = **0.833**,
- positive predictive value = **0.820**,
- negative predictive value = **0.776**,
- % correctly classified = **79.6**.

Besides the choice of cutoff, also the population prevalence[9] is important for assessing classification performance.

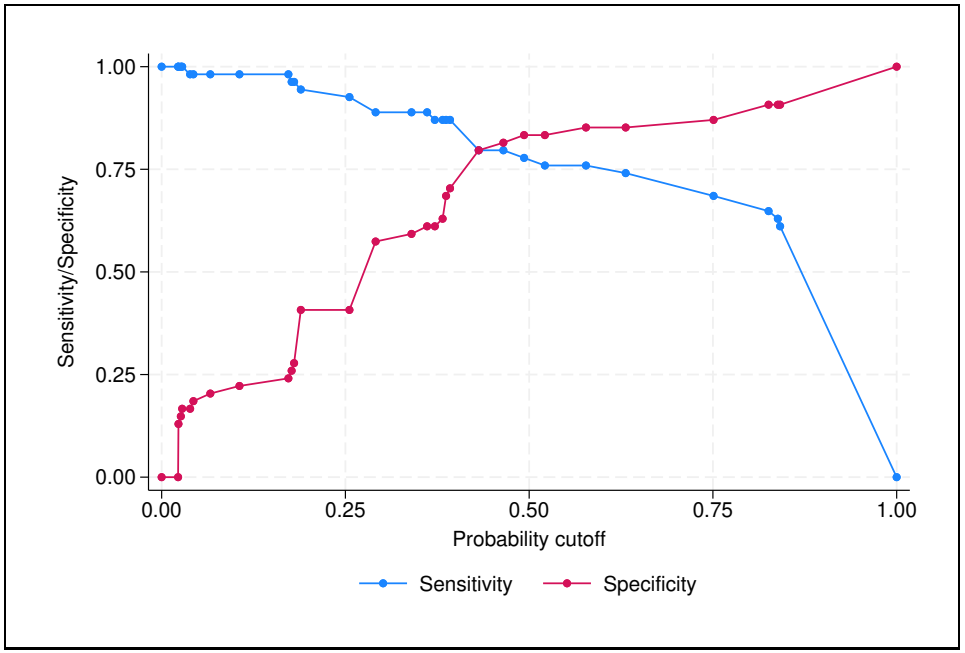
[8] The typical software default is: "positive" for $\hat{p} > 0.5$ and "negative" for $\hat{p} \leq 0.5$, but alternative cutoffs may be considered, in particularly when the proportion of events is far from 0.5.

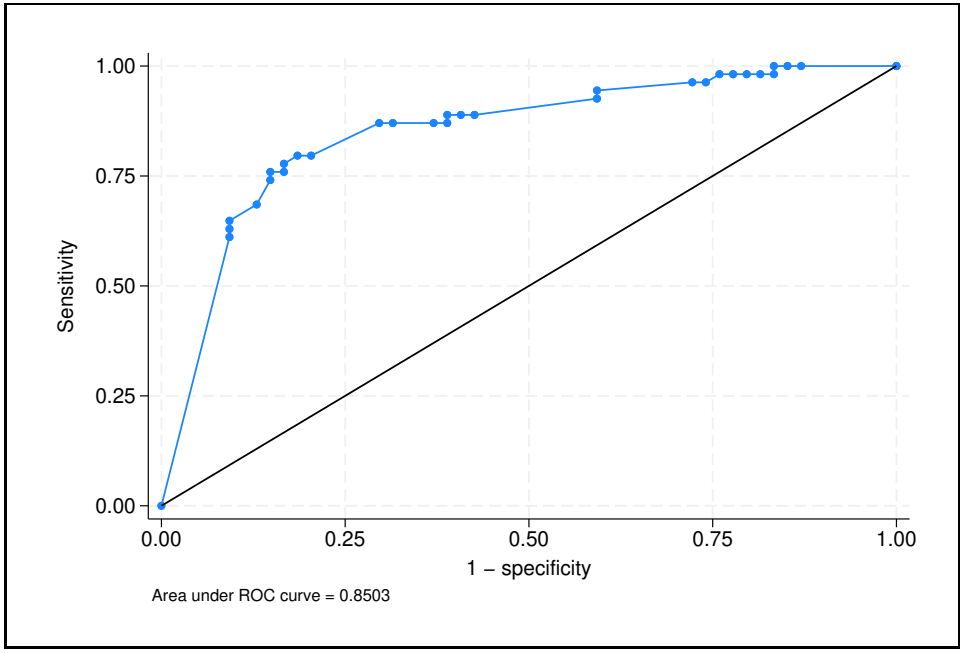
[9] In a case-control study, the prevalence in the data is artificial, and the classification rates have no meaning beyond the sample itself.

ROC Analysis for Logistic Regression Models

Added value of ROC analysis: considers discrimination at all cutoffs instead of one,

- graphs: one- and two-curve displays (see below, VER Examples 16.9 and 16.10),
- AUC (area under curve, in 1-curve graph): single value ∼ discriminatory ability.[10] [11]





[10] AUC can be interpreted as the (approximate) probability that in a randomly chosen pair of observations from the study sample with different outcomes and different predicted probabilities, say $\hat{p}_1$ for the event and $\hat{p}_0$ for the non-event, it holds that $\hat{p}_1 > \hat{p}_0$.

[11] Guideline for interpretation of AUC values (from Hosmer & Lemeshow (2000), *Applied Logistic Regression*): i) AUC = 0.5: no discrimination; ii) $0.5 <$ AUC < 0.7: poor discrimination; $0.7 \leq$ AUC < 0.8: acceptable discrimination; $0.8 \leq$ AUC < 0.9: excellent discrimination; AUC ≥ 0.9: outstanding discrimination.

Reliability (Cross-validation) in Logistic Regression Models

Synthesis: principles entirely the same as for linear models, but options more limited:

- R^2-based statistics no longer useful → other measures of predictive ability, e.g. AUC or percent correctly classified,
- less common access through software built-in features ⇒ manual coding attractive (as an alternative to relying on any ad-hoc implementations in add-on packages):
  - requires access to statistical programming language where loops can be implemented easily, e.g. Stata.

Example: leave-one-out cross-validation for Nocardia data model:

- all classification statistics at cutoff 0.5 unchanged![12], including the percent correctly classified (79.6%),
- AUC drops from 0.85 to 0.79,
  - some shrinkage on cross-validation,
  - AUC crosses guideline cutoff for "excellent" discrimination.[13]

[12] This can happen if none of the leave-one-out cross-classified predicted probabilities switch across the cutoff.
[13] An illustration of the issue with setting cutpoints for interpretations...

Residuals and Diagnostics

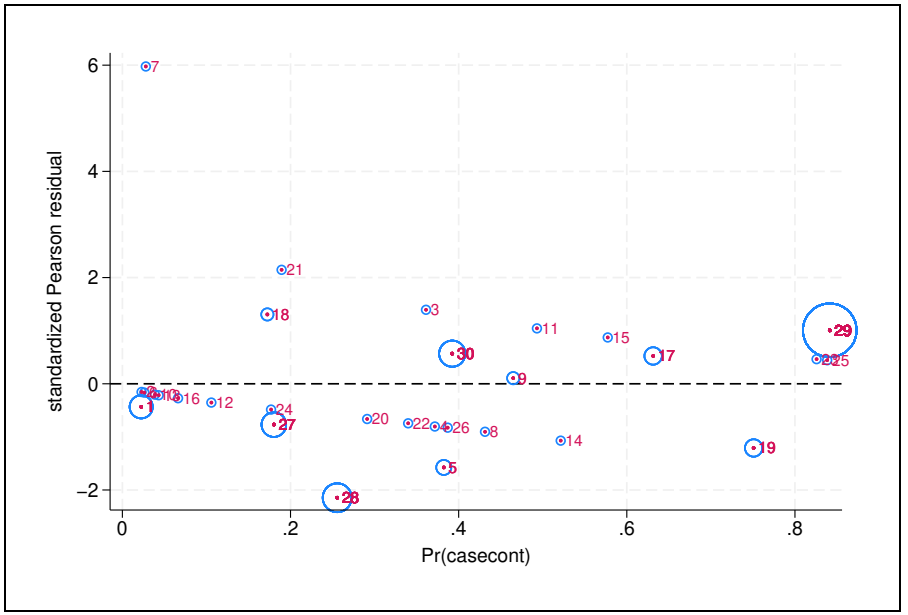
Synthesis: some major differences to linear models exist:

- diagnostics best done per covariate pattern, but still potential problems with sparse covariate patterns which do not provide much information,[14]
- no approximate $N(0,1)$ distribution expected for residuals and less specific guidelines/cutoffs $\Rightarrow$ look for extreme values rather than compliance with rules.

Residual types (per covariate pattern j):

- Pearson residual r_j, defined on slide 5aL−3,
- standardized Pearson residual (further standardized to a variance closer to 1): $r_{Sj} = r_j/\sqrt{1-h_j}$, where h_j is the leverage for covariate pattern j (next slide),
- deviance residual d_j, constructed so that the deviance statistic equals the sum of squared contributions, $D = \sum_j d_j^2$.[15]

Residual plot for Nocardia data



[14] One option may be to reduce the number of covariate patterns for diagnostic purposes (slides 5aL−13/14).
[15] The mathematical form of d_j is less informative; deviance is a concept from generalised linear models.

Leverage for Logistic Regression

As in linear models, leverage measures "outlyingness" among the predictors (their covariate patterns), but with an additional dependence on the outcome probability $\hat{p}$, whereby the leverage h_j for the j^{th} covariate pattern tends to be:[16]

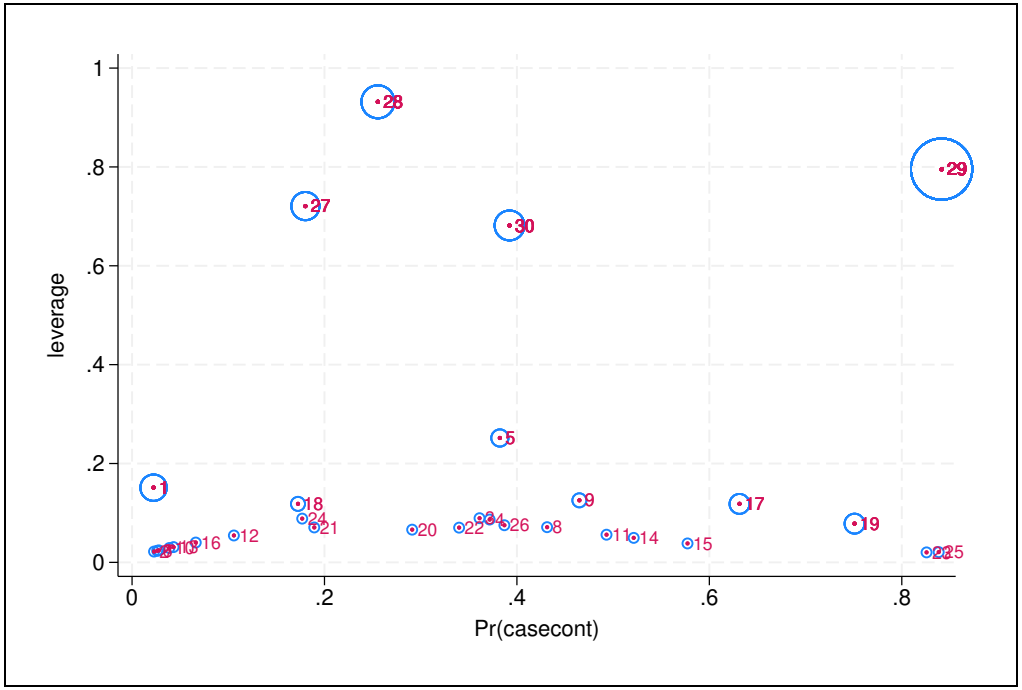
- h_j small for $\hat{p}_j < 0.1$ or $\hat{p}_j > 0.9$,
- h_j small to moderate for $0.3 < \hat{p}_j < 0.7$,
- h_j large for $0.1 < \hat{p}_j < 0.3$ or $0.7 < \hat{p}_j < 0.9$,
- always: $0 < h_j < 1$.

Interpretations of highest leverage values:

- mostly in range indicated,
- also for largest groups and with `dcpct=100`.

note: no cutoff to define "high" h_j.

Leverage against $\hat{p}$ for Nocardia data



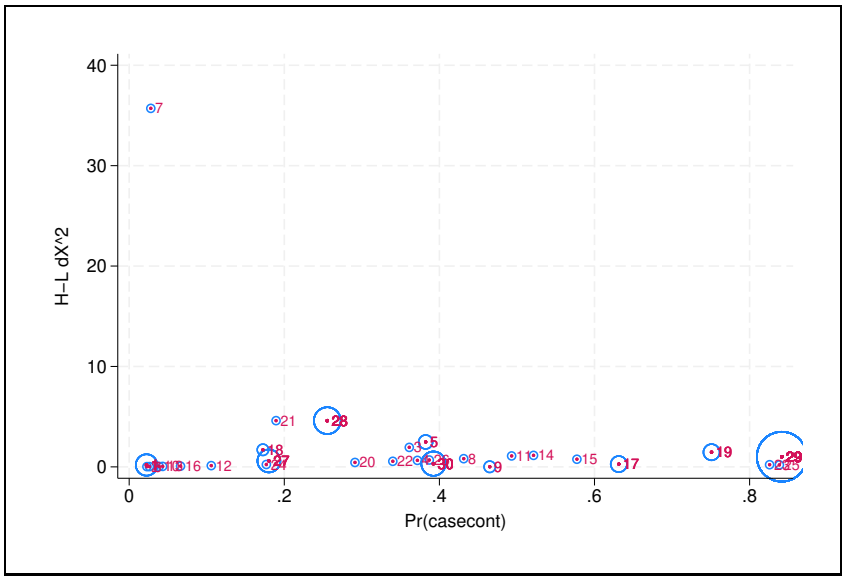
[16] Based on Hosmer & Lemeshow (2000), Section 5.3.

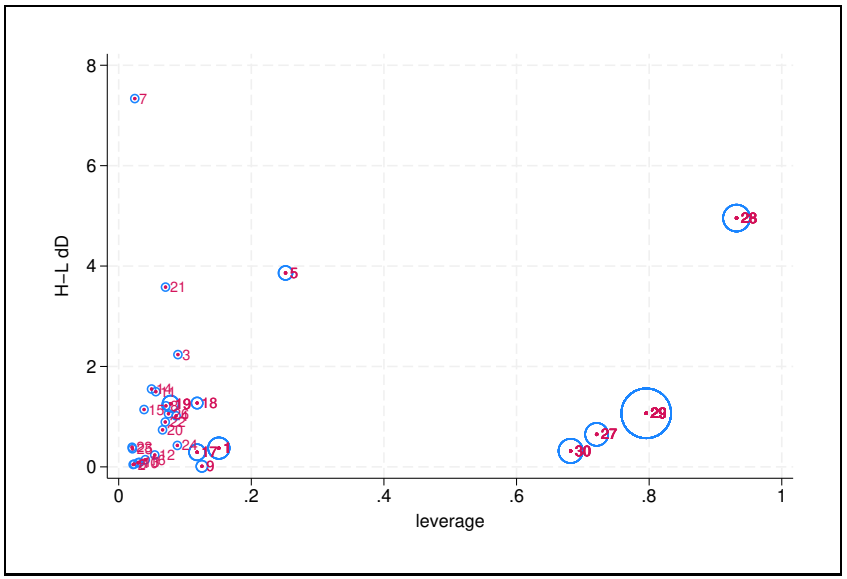
Influence on Model Fit

Idea: quantify how covariate patterns impact the Pearson X^2 and deviance D statistics,

- Pearson and deviance residuals defined from the contributions to respective statistics from each covariate pattern,
- ΔX^2 and ΔD statistics approximate the decrease in X^2 and D, respectively, from omitting a covariate pattern.[17]

Graphical exploration: plot ΔX^2 (or ΔD) against ($\hat{p}_j$) or against leverage.





- covariate pattern # 7 has huge and strong impacts on X^2 and D, respectively.

[17] In formulas, $\Delta X^2_j = r^2_j/(1-h_j) = r^2_{Sj}$, and $\Delta D_j = d^2_j/(1-h_j)$.

Influence on Parameter Estimates

Delta-beta statistic ($\Delta\beta_j$): approximates the change in full parameter vector if covariate pattern j is omitted[18] [19] — can also be plotted against ($\hat{p}_j$) or leverages.

Summary of most extreme residuals/diagnostics for covariate patterns of the Nocardia data model (VER Example 16.11):

\# 7: a single case herd with low predicted value
⇒ very large residuals and model fit diagnostics — may be an unusual herd, but may also be an artifact from a covariate pattern without replication,

\# 28: a total of 11 herds (`dneo=no`, `dclox=yes`, `dcpct=100`) with very high $\Delta\beta$ value
— there is only one additional herd with this `dneo*dclox` combination
⇒ estimation of the interaction without these 11 herds becomes problematic (boundary problem on logit scale), so the interaction depends crucially on # 28,

\# 29: a total of 38 herds with quite large $\Delta\beta$ — not surprising because $\approx \frac{1}{3}$ of the data.

Conclusion: not clear any changes to the model are necessary (but the information could be reflected in the reporting of results).

Possible next step: analyse without some of these covariate patterns and note the differences in results.

[18] Stata only offers the overall $\Delta\beta$ statistic, other software packages also offer statistics for each parameter, similar to the DFBETA statistics for linear models.

[19] The $\Delta\beta$ statistic is calculated from the other statistics as: $\Delta\beta_j = r_{Sj}^2\, h_j/(1-h_j)$.

Grouped Nocardia Model

Objective: refit Nocardia model with sufficient replication to make covariate-pattern based diagnostics more interesting:

- problematic predictor = `dcpct`: semi-continuous with range 0−100 and 19 distinct values; however 66/108 herds have value 100,
- categorization approach: define categories as 0−49, 50−99, and 100 (42, 18 and 66 herds, respectively) → `dcpct3`,
- model with `dcpct3` has slightly better fit (i.e., lower $\ln L$) than with `dcpct` as linear (and 1 parameter more, so higher AIC), and 11 distinct covariate patterns:

```
. logit casecont i.dcpct3 i.dneo##i.dclox
                                                Log likelihood = -51.632242
------------------------------------------------------------------------------
    casecont | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      dcpct3 |
          50 |   1.361002    .819178     1.66   0.097    -.2445579    2.966561
         100 |   2.026562   .6855237     2.96   0.003     .6829604    3.370164
        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
       _cons |  -3.531226   .9364287    -3.77   0.000    -5.366593    -1.69586
------------------------------------------------------------------------------
```

Cov	Pos	Neg	dcpct3	dneo	dclox
1	1	11	0	no	no
2	4	7	0	yes	no
3	0	1	0	yes	yes
4	0	2	50	no	no
5	1	0	50	no	yes
6	7	3	50	yes	no
7	1	4	50	yes	yes
8	1	7	100	no	no
9	2	9	100	no	yes
10	33	5	100	yes	no
11	4	5	100	yes	yes

Diagnostics for Grouped Nocardia Model

Covar. pattern j	Prop. cases	Fit $\hat{p}_j$	Residuals r_j	r_{Sj}	d_j	Leverage h_j	Influence ΔX^2	ΔD	$\Delta\beta$
1	.083	.028	1.14	1.36	0.93	0.29	1.85	1.23	0.76
2	.364	.416	-0.35	-1.07	-0.36	0.89	1.15	1.17	9.50
3	.000	.082	-0.30	-0.31	-0.41	0.06	0.09	0.18	0.01
4	.000	.102	-0.48	-0.52	-0.66	0.17	0.27	0.52	0.05
5	1.00	.152	2.36	2.50	1.92	0.11	6.23	4.21	0.74
6	.700	.735	-0.25	-0.47	-0.25	0.70	0.22	0.21	0.51
7	.200	.258	-0.30	-0.42	-0.30	0.50	0.17	0.18	0.17
8	.125	.182	-0.42	-0.80	-0.44	0.73	0.64	0.70	1.74
9	.182	.259	-0.58	-2.50	-0.61	0.95	6.23	6.72	108
10	.868	.844	0.42	1.08	0.43	0.85	1.17	1.22	6.69
11	.444	.403	0.25	0.52	0.25	0.76	0.27	0.27	0.87

- no indication of lack of fit: both the Pearson ($X^2 = 8.22$) and the deviance ($D = 6.32$) statistics are non-significant with df$= 5$, and the by far largest contribution is from a covariate pattern (# 5) with no replication,
- leverage not too informative (because all predictors categorical),
- a notable difference between the Pearson (deviance) residuals and the corresponding influence statistics (ΔX^2 and ΔD) for # 9 (due to high h_j),
- extreme $\Delta\beta$ for # 9 (same as for ungrouped model), and high for # 2 (not seen before) and #10 (also in ungrouped model).

Summary for Logistic Regression Diagnostics

What can residuals/goodness-of-fit tests/diagnostics show or detect in logistic models?

- observation-level residuals:
  exploratory listing of observations with the "wrong outcome" (i.e., $y_i = 1$ when $\hat{p}_i$ small, or $y_i = 0$ when $\hat{p}_i$ large), indicating possible errors in individual observations,
- covariate pattern residuals and goodness-of-fit tests:
  inadequacies in the modelling of the predictors in model, e.g. non-linearity or missing interactions; but cannot detect missing predictors or other groupings of observations not explicitly included,
- diagnostics (per covariate pattern):
  consequences of current model, in particular high influence of specific covariance patterns on the model fit and parameter estimates — an aid to avoid misinterpretations of the results.

Some words of advice and caution (from Hosmer & Lemeshow, 2000):[20]

> In logistic regression we have to rely primarily on visual assessment, as the distribution of the diagnostics under the hypothesis that the model fits is known only in certain limited settings. [...] All of the diagnostics are evaluated by covariate pattern; hence any approximations to their distributions [...] depend on the number of subjects with that pattern. When a fitted model contains some continuous covariates then [... (the necessary)] asymptotic results cannot be relied upon. Thus, in practice, an assessment of "large" is, of necessity, a judgment call based on experience and the particular set of data being analyzed.

[20] Tentative recommendations for "interesting" values of the diagnostics: $\Delta X^2, \Delta D \geq 3.84 (= \chi^2_{.95}(1))$; $|\Delta\beta| \geq 1$.