Index of Lecture 4b: Model building for logistic regression

Practical Information

News/Schedule:

- lab session for VHM 802/812 on Monday:
  - logistic regression, incl. VER 16.1 (basics) and 16.2 (model building),
  - also possible to review linear regression model building,
- homework for Wednesday: preparation for discussion of second logistic regression exercise (VER 16.2); note: the causal diagram is non-standard, so don't spend too much time on it.

Today's session:

- model-building exercise (VER 15): interpretations of causal diagram (Q3-Q5, Q9),
- last part of previous lecture: likehood-based inference and prediction in multiple logistic regression,
- new material on logistic regression:
  - model-building for logistic regression: principles and procedures (focusing mostly on differences to linear regression),
  - model selection criteria,
  - brief intro to generalised linear models,
- dicussion of presentation of predictor effects (from logistic model example).

Introduction to Logistic Model-Building

Synthesis:

- similar to linear regression model-building,[1]
- major principles and tools still apply, such as
  - causal diagrams (confounding and intermediate/intervening effects),
  - statistical testing to assess (significance of) effects — however different types of tests (Wald and likelihood-ratio tests, both with reference χ^2-distribution),
  - model fit statistics to compare (non-nested) models — however different type of statistics (AIC/BIC instead of R^2-type statistics),
  - exploration of form of effects (linearity, interaction),
  - validation of model assumptions by residuals and diagnostics (next lecture).

List of topics covered in detail in this lecture:

- confounding and interaction,
- evaluation of linearity of continuous predictors (adaptation of tools from linear to logistic regression),
- model selection procedures.

[1] Due to having the same expression for the linear predictor, i.e. $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$.

Confounding in Logistic Models

Same approach as in linear models:

- draw causal diagram to determine potential confounders,
- analyze with/without, and note "substantial" changes in estimates.

Illustration: confounding by `dcpct` on effects of `dneo` or `dclox` in Nocardia data?

- `dcpct` not "causally later" than `dneo` and `dclox`,
- `dcpct` significant ($z = 3.15, P = 0.002$) in "full" model,[2]
- check association between `dcpct` and `dneo`/`dclox`, in the controls (case-control study),
- comparison of estimates for `dneo` and `dclox` with/without `dcpct`:

Effect	Without `dcpct` ("crude")	With `dcpct` ("full")
`dneo`	2.38 (SE=0.55, $P < 0.0005$)	2.21 (SE=0.58, $P < 0.0005$)
`dclox`	-1.01 (SE=0.53, $P = 0.058$)	-1.41 (SE=0.56, $P = 0.011$)

Conclusions:

- `dneo`: no serious confounding effect by `dcpct` (7% change, strong signif. in both models),
- `dclox`: substantial confounding effect by `dcpct` (40% change[3], change in significance).

[2] Strictly speaking, the association should exist in non-exposed subjects, see do-file.
[3] Calculated as 0.40/1.01 = 40%, using the guidelines from Section 13.5.2 of VER2, as listed for Example 16.4 in the VER2 Errata (note: typo in text: 30% computed wrongly).

Interaction in Logistic Models

- same construction as in linear models using cross-product terms and/or dummy variables,
- similar statistical assessment using Wald/LR-tests,
- same interpretation on logit scale, with interaction plot, but should convert to odds-ratios (maybe probability scale).

Illustration: interaction between `dneo` and `dclox` in model for Nocardia data also including `dcpct`:

$$\text{logit}(\hat{p}) = -3.777 + 0.023\,\texttt{dcpct} + 3.184\,\texttt{dneo} + 0.446\,\texttt{dclox} - 2.552\,\texttt{neoclox}$$

Estimated $\text{logit}(\hat{p})$ for combinations of (`dneo`,`dclox`) and a fixed value of `dcpct` (using $a = -3.777 + 0.023\,\texttt{dcpct}$):

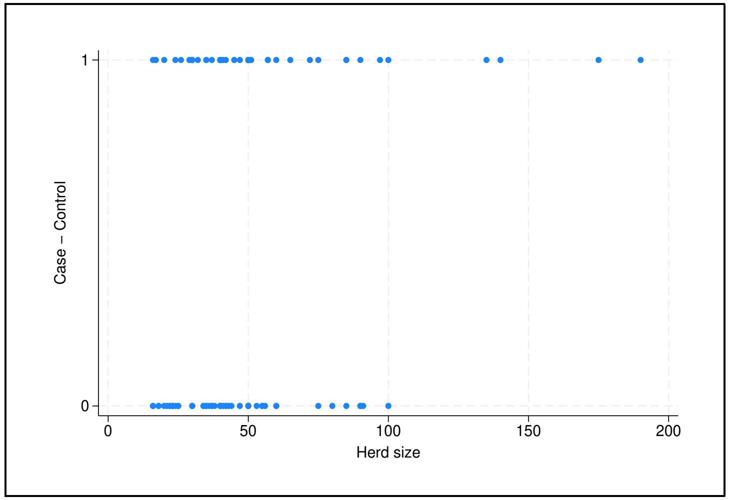
	dclox	
dneo	0	1
0	$a + 0 + 0 + 0$ $= a$	$a + 0 + 0.446 + 0$ $= a + 0.446$
1	$a + 3.184 + 0 + 0$ $= a + 3.184$	$a + 3.184 + 0.446 - 2.552$ $= a + 1.078$

- effect of `dclox` when `dneo` absent: 0.446 (OR = 1.56),
- effect of `dneo` when `dclox` absent: 3.184 (OR = 24.1),
- effect of `dclox` when `dneo` present: −2.106 (OR = 0.122),
- effect of `dneo` when `dclox` present: 0.632 (OR = 1.88),
- "added effect" of both products simultaneously: −2.552 (does not correspond to any OR!).

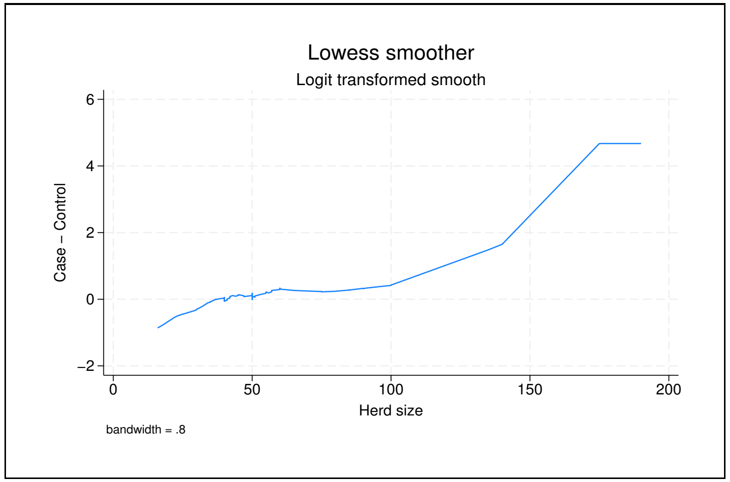
Evaluating Linearity by Plots

More difficult than in linear regression(!):

- less information in binary data ⇒ simple scatterplots often useless with continuous predictor; illustration for `numcow`:



- residuals of little use without replication/grouping (next lecture); `numcow` has almost no replication,
- new tool[4]: lowess smoothed scatterplots on logit scale:



  * bandwidth=0.8 ∼ 80% of points included in weighted average,
  * beware: smoother may be noisy towards ends of data range,
  * interpretation: linear up to ≈ 60, then flat up to ≈ 100.

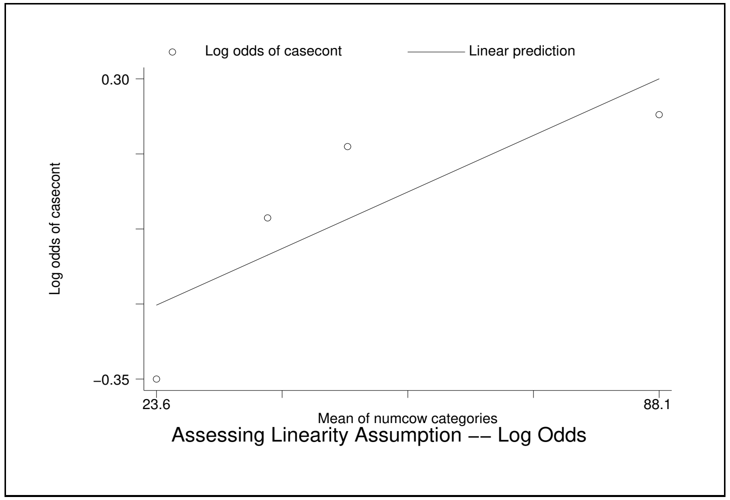
[4] For simple associations only; no multivariable counterpart exists.

Evaluating Linearity by Categorization

Same as in linear regression: (but more important here because of less alternatives)[5]

- divide predictor's range into a few intervals[6] based on cut-points with biological meaning or corresponding roughly to relevant percentiles,
- fit model with categorical (ordinal) predictor, and assess whether estimates follow a linear pattern.

Illustration with 4 intervals for `numcow` (univariate model):



Findings:

- curve appears non-linear (but non-significant estimates),
- created with `lintrend` and `egen cut` commands.

	Interval (x)			Data (y)			Estimates	
	min	max	mean	nobs	prop	logit	$\hat{\beta}$	SE
1	16	30	23.6	29	0.41	-0.35	0	—
2	32	41	37.9	26	0.50	0.00	0.348	0.54
3	42	53	48.2	26	0.54	0.15	0.502	0.54
4	55	190	88.1	27	0.56	0.22	0.571	0.54

[5] The idea applies to both simple and multivariable logistic regression models.
[6] It is often useful to try some different numbers of categories.

Evaluating Linearity by Polynomials

Entirely same methods[7] as for linear models, both for simple polynomials and advanced polynomials (see do-file),

- construct predictors corresponding to desired model form, and assess significance,
- illustrated by quadratic term for `numcow` (univariate):

	Linear polynomial			Quadratic polynomial		
Coef.	Estim.	SE	*P*	Estim.	SE	*P*
intercept	-0.71	.41	—	-0.34	.76	—
`numcow`	0.015	.008	0.056	0.0010	.026	0.97
$\texttt{numcow}^2$	—	—	—	0.000098	.00018	0.59

  * no significance for quadratic term,
  * note: variable values for intercept and linear term,
  * centering of `numcow` reduces high collinearity (but not a serious problem here).

Summary of analyses for linearity of `numcow`:

- indications of non-linearity (categorized version, smoothed scatterplot),
- no significance for non-linear model (though close with advanced methods, see do-file),

⇒ simple linear modelling seems preferable (any effect is weak, and `numcow` drops out quickly in multivariable models anyway).

[7] Applicable to both simple and multivariable logistic regression models.

MODEL/VARIABLE SELECTION IN LOGISTIC MODELS (VER)

- Main tool: statistical tests of nested models (only),
- Supplementary tools: measures of model fit adjusted for the number of parameters,
  - R^2-type statistics exist, but not recommended (difficult to interpret),
  - AIC statistic[8]: general for likelihood-based inference, and preferable to BIC[9],[10]
  - may be used to compare non-nested models (if no nesting possible),
  - should (in my view) not replace tests between nested models,
- Automated procedures (forward/backward/stepwise): as in linear models (same advantages/drawbacks), here usually based on Wald tests,
- Initial screening of (a large number of) predictors at liberal P,
- Recommended approach: manual selection among models built upon consideration of causal diagram and effects in data.

Illustration:
dropping `dclox` from model:

Model	$2\ln L$	AIC	BIC	G^2	df	P
`dneo##dclox,dcpct,dbarn2`	-97.31	109.31	125.40	—	—	—
`dneo,dcpct,dbarn2`	-107.25	115.25	125.98	9.95	2	0.007

- quite strong significance of `dclox` (despite non-significant main effect and weakly sign. interaction).

[8] Akaike's Information Criterion (AIC): computed as $-2\ln L + 2\cdot\#\text{param}$.
[9] Bayesian Information Criterion (BIC): computed as $-2\ln L + \ln(n)\cdot\#\text{param.}$, where n = number of observations.
[10] For both AIC and BIC, the sign is unimportant, but smaller/lower values correspond to "better" models, e.g. $1.2 < 3.5$, or $-5.3 < -2.5$.

More about the AIC and BIC

The "best" model selection statistic…

- people have very different opinions about this, sometimes tending to "semi-religious" convictions,
- preferences may depend on general statistical approach (Bayesian vs. classical/ frequentist) and proposed use (e.g., as a guidance or as a definitive model selection tool),
- AIC and BIC are both based on the likelihood function with (different) penalties for the number of observations,
  - "best AIC" is more liberal ($\Rightarrow$ larger models) than LR-tests for small df (up till df$=$7), more conservative ($\Rightarrow$ smaller models) for large df,
  - "best BIC" is very conservative ($\Rightarrow$ small models),
  - use only for models with comparable likelihoods (e.g., same data/scale!) — a very common source of errors.

Guidelines for interpreting AIC differences exist:[11] (this is not an endorsement of their use for automated model selection!)

	$AIC_2 - AIC_1$	level of support for Model 2
Model 1 has	$0-2$	substantial
lower (better)	$4-7$	considerably less
AIC	>10	essentially none

[11] Burnham and Anderson 2002, *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*, 2nd ed., p. 170; guidelines (from a Bayesian framework) for the BIC are less (little?) useful in classical statistics.

Presentation of Predictor Effects (Recap)

General model building recommendations (lecture 3bL) also apply to logistic models, although the need to backtransform is more evident (and also more attractive, as ORs).

Stata listing of potential final model — what is useful and what is not?

```
. logit casecont i.dneo##i.dclox dcpct i.dbarn /* potential final model */

  Logistic regression                                   Number of obs =    108
                                                        LR chi2(6)    =  52.59
                                                        Prob > chi2   = 0.0000
  Log likelihood = -48.563447                           Pseudo R2     = 0.3513
  ------------------------------------------------------------------------------
      casecont | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
  -------------+----------------------------------------------------------------
          dneo |
           yes |   3.840981   .9613888     4.00   0.000     1.956693    5.725268
         dclox |
           yes |     .85191   1.086531     0.78   0.433    -1.277651    2.981471
    dneo#dclox |
       yes#yes |  -2.843304   1.269394    -2.24   0.025     -5.33127   -.3553377
         dcpct |   .0215351   .0082369     2.61   0.009     .0053911    .0376792
         dbarn |
      tiestall |  -1.488824    .662023    -2.25   0.025    -2.786365   -.1912831
         other |  -.4818957   1.137152    -0.42   0.672    -2.710672     1.74688
         _cons |  -3.272424   1.063408    -3.08   0.002    -5.356666   -1.188183
  ------------------------------------------------------------------------------
```

GENERALISED LINEAR MODELS (GLMs)

Main idea: extend scope of linear modelling by transforming, using the link function g, the parameter μ of principal interest (usually, the mean) to a more appropriate scale for linear modelling:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

Components of a generalised (not general!) linear model:

- link function g,
- set of explanatory variables (represented by x's),
- distribution of outcome Y (usually one of few standard distributions),
- assumption of independence between outcomes.

Examples:

- binary / binomial distribution – logistic regression (logit link),[12]
- multinomial distribution (ordinal data) – ordinal logistic regression,
- Poisson distribution[13] (count data) – Poisson regression (with log link).

GLMs – a unified framework for many different models/analyses:

- similarities to linear models, due to predictors entering in the same way: quantitative/qualitative predictors, main effects, interactions, dummy variables . . . ,
- common algorithms (implemented in software): maximum likelihood estimation, inference (Wald type and likelihood-based) and model checking procedures,

⇒ make analysis simpler/more intuitive.

[12] Other link functions sometimes used for proportions: probit and complementary log-log.
[13] The so-called negative binomial distribution is also frequently used for count data.