Index of Lecture 4b

Practical Information

Today's lecture:

- follow-up from previous lectures:
  - ∗ prediction in linear/logistic regression (lecture 3a), backtransformation in linear regression,
  - ∗ presentation of categorical predictor effects (new slide),
- 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,
- any follow-up from yesterday's lab session?

News/Schedule:

- new VHM 812 quiz coming up (Wednesday),
- work for Friday: logistic regression exercise 2 (VER 16.2; continuation of VER 16.1 with same dataset).

Introduction to Logistic Model-Building

Synthesis:

- similar to linear regression model-building,
- major principles and tools still apply, such as
  - causal diagrams (confounding and intermediate effects),
  - statistical testing to assess (significance of) effects, however different types of tests (Wald/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,
- also brief discussion of pairwise comparisons (categorical predictors) in linear and logistic models; usually covered in more detail in VHM 802.

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,[1]
- check association between `dcpct` and `dneo`/`dclox`,
- 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 significance in both models),
- `dclox`: substantial confounding effect by `dcpct` (40% change[2], change in significance).

[1] Strictly speaking, the association should exist in non-exposed subjects.

[2] 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.

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

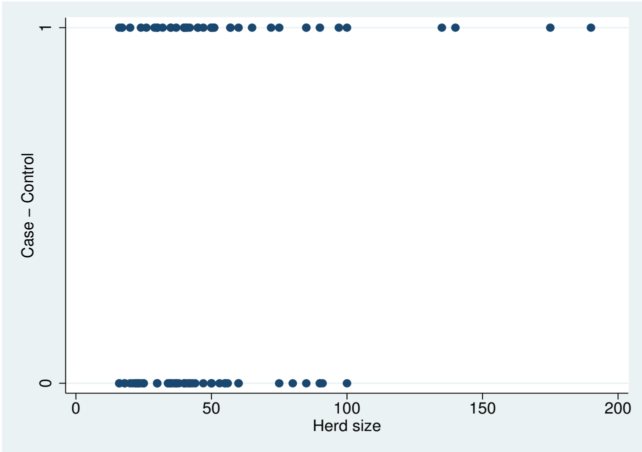
	dclox	
dneo	0	1
0	$a + 0 + 0 + 0$ $= a$	$a + 0 + 0.446 + 0$ $= a + 0.446$
1	$a + 3.184 + 0 + 0$ $= a + 3.184$	$a + 3.184 + 0.446 - 2.552$ $= a + 1.078$

- 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 $= -2.552$.

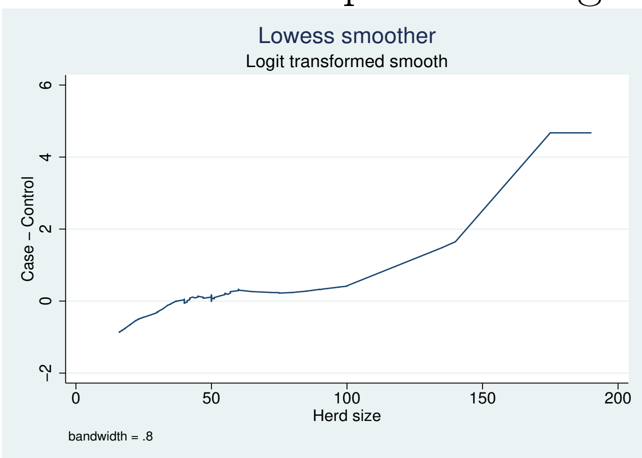
Evaluating Linearity by Plots

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[3]: 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.

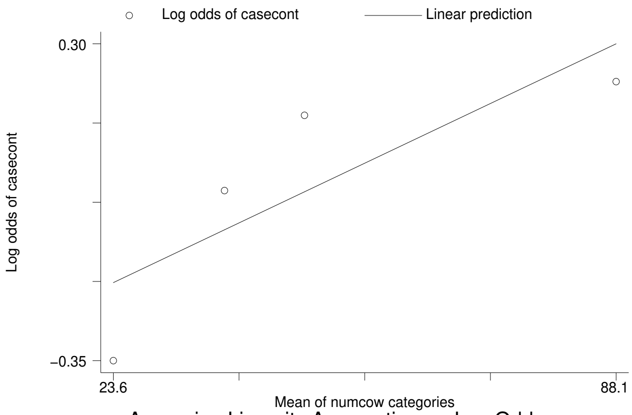
[3] For simple associations only; no multivariable counterpart exists.

Evaluating Linearity by Categorization

Same as in linear regression:
(but more important here because of less alternatives)[4]

- divide predictor's range into a few intervals[5] 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):



Assessing Linearity Assumption −− Log Odds

	Interval			Data			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

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

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

Evaluating Linearity by Polynomials

– entirely same methods[6] as for linear models, both for simple polynomials and advanced polynomials (see do-file),

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

	Linear polynomial			Quadratic polynomial		
Coef.	Estim.	SE	P	Estim.	SE	P
interc.	-0.71	.41	–	-0.34	.76	–
`numcow`	0.015	.008	0.056	0.001	.026	0.97
`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).

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

Model/Variable Selection in Logistic Models

- Main tool: statistical tests of nested models (only),
- Supplementary tools: measures of model fit adjusted for no. of parameters,
  - AIC statistic[7]: general for likelihood-based inference, and preferable to BIC[8],[9]
  - R^2-type statistics exist for logistic regression but not recommended (difficult to interpret),
  - used for overview and "informal" model choice; also maybe 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 based on Wald tests,
- Initial screening of (a large number of) predictors at liberal P,
- Recommendation: manual selection among models built upon consideration of causal diagram and effects in data.

Illustration: dropping `dclox` from model (`dneo##dclox`,`dcpct`,`dbarn2`):

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 significant interaction).

[7] Akaike's Information Criterion (AIC): computed as $-2\ln L + 2\cdot\#$param.

[8] Bayesian Information Criterion (BIC): computed as $-2\ln L + \ln(n)\cdot\#$param., where n = no. of obs.; being developed within the Bayesian framework, its guidelines are less (little?) useful in classical statistics.

[9] 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 no. of observations,
  - ∗ "best AIC" is more liberal (⇒ larger models) than LR-tests for small df (up till df=7), more conservative (⇒ smaller models) for large df,
  - ∗ "best BIC" is very conservative (⇒ 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:[10]
(this is not an endorsement of their use for mechanical 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

[10] Burnham and Anderson 2002, *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*, 2nd ed., p. 170.

Presentation of Categorical Predictor Effects

Things to consider in linear/logistic models when presenting results for a categorical (> 2 categories) predictor:

- unless pre-determined hypotheses exist, the overall P-value (for H_0 : equal effects across all categories) should be presented and used for discussion,
- for significant or "interesting" predictors, estimates with SE or CI should be shown in one of two possible layouts:
  - ∗ differences to a "reference" (or "baseline") category: the typical format of software packages when a particular parametrization has been selected (e.g., `regress` and `logit`),
  - ∗ estimates for all categories with suitably defined (and reported) averaging across or values set for all other predictors,
- estimates should be backtransformed as needed (e.g., values may be given both as coefficients on logit scale and as ORs),
- unless pre-determined hypotheses exist, pairwise comparisons should be conducted between *all* category pairs[11],
  - ∗ adjusted for multiple comparisons if many categories are present and strict overall significance is required, otherwise unadjusted,
  - ∗ many methods for adjustment for multiple comparisons exist[12]; the simplest and most flexible is the Bonferroni method: divide the significance level (0.05) by the number of comparisons (m) or multiply each P-value by m.
- results of pairwise comparisons may be reported (P-values) or indicated by letter coding[13].

[11] Not only for comparisons with a reference category.

[12] Discussed in more detail in VHM 802.

[13] Most common system: two categories with the same letter indicated are not significantly different.

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 distrib.),
- assumption of independence between outcomes.

Examples:

- binary / binomial distribution – logistic regression (logit link),[14]
- multinomial distribution (ordinal data) – ordinal logistic regr.,
- Poisson distribution (count data) – Poisson regression (log link).

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

- similarities to linear models make analysis simpler/more intuitive,
- flexibility in modelling by choice of link function and distribution,
- common algorithms (software): maximum likelihood and quasi-likelihood[15] estimation, inference (Wald type and likelihood-based) and model checking procedures.
- (link-)specific tools for interpretation, e.g. odds-ratio.

[14] Other links sometimes used for proportions: probit and complementary log-log.
[15] Quasi-likelihoods were developed for specific situations where ordinary likelihood functions do not exist, e.g. overdispersed models (later lectures).