

## Index of Lecture 4b: Model building for logistic regression

Page	Title
1	Practical information
2	Introduction to logistic model-building
3	Confounding in logistic models
4	Interaction in logistic models
5	Evaluating linearity by plots
6	Evaluating linearity by categorization
7	Evaluating linearity by polynomials
8	Model/variable selection in logistic models
9	More about the AIC and BIC
10	Presentation of categorical predictor effects
11	Generalised linear models

## PRACTICAL INFORMATION

### News/Schedule:

- today is my third (and last) lecture on logistic regression,
- homework for Friday: first home assignment (both courses), but we postpone the next logistic regression exercises (VER 16.2-3) until Tuesday,
- don't forget this week's quiz on logistic regression (VHM 812).

### Today's session:

- last bits from previous lecture: **prediction** in 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**,
- \* presentation of categorical predictor effects (both linear and logistic models),
- any **follow-up** from yesterday's lab session?

## INTRODUCTION TO LOGISTIC MODEL-BUILDING

### Synthesis:

- similar to linear regression model-building,<sup>1</sup>
- 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  $\chi^2$ -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 (for categorical predictors) in linear and logistic models; usually covered in more detail in VHM 802 (+ VHM 801).

---

<sup>1</sup> Due to having the same expression for the linear predictor, i.e.  $\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$ .

## 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,<sup>2</sup>
- 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 <code>dcpct</code> (“crude”)	With <code>dcpct</code> (“full”)
<code>dneo</code>	2.38 (SE = 0.55, $P < 0.0005$ )	2.21 (SE = 0.58, $P < 0.0005$ )
<code>dclox</code>	-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<sup>3</sup>, change in significance).

<sup>2</sup> Strictly speaking, the association should exist in non-exposed subjects, see do-file.

<sup>3</sup> 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:

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

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

- 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 (no corresponding OR!).

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

## EVALUATING LINEARITY BY PLOTS

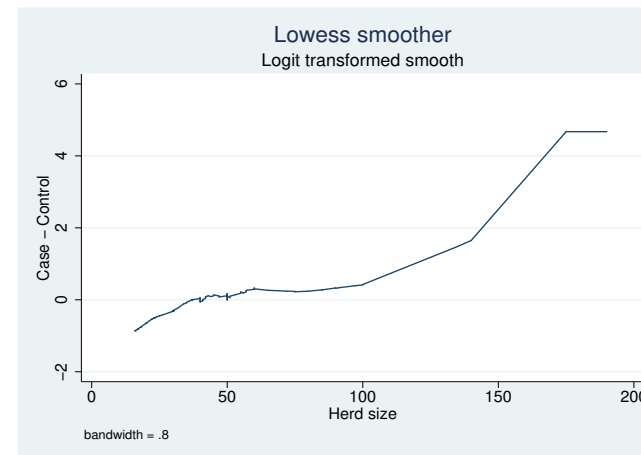
More difficult than in linear regression (!):

- less information in binary data  $\Rightarrow$  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<sup>4</sup>: lowess smoothed scatterplots on logit scale:

- \* bandwidth=0.8  $\sim$  80% of points included in weighted average,
- \* beware: smoother may be noisy towards ends of data range,



- \* interpretation: linear up to  $\approx$  60, then flat up to  $\approx$  100.

<sup>4</sup> For simple associations only; no multivariable counterpart exists.

## EVALUATING LINEARITY BY CATEGORIZATION

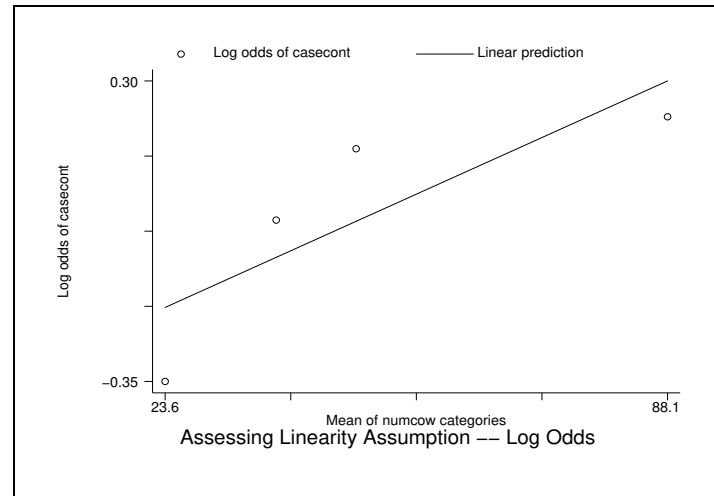
Same as in linear regression: (but more important here because of less alternatives)<sup>5</sup>

- divide predictor's range into a few intervals<sup>6</sup> 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

<sup>5</sup> The idea applies to both simple and multivariable logistic regression models.

<sup>6</sup> It is often useful to try some different numbers of categories.

## EVALUATING LINEARITY BY POLYNOMIALS

Entirely same methods<sup>7</sup> 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):

Coef.	Linear polynomial			Quadratic polynomial		
	Estim.	SE	<i>P</i>	Estim.	SE	<i>P</i>
intercept	-0.71	.41	—	-0.34	.76	—
numcow	0.015	.008	0.056	0.0010	.026	0.97
numcow <sup>2</sup>	—	—	—	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).

<sup>7</sup> 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 number of parameters,
  - \* **AIC** statistic<sup>8</sup>: general for likelihood-based inference, and preferable to BIC<sup>9,10</sup>,
  - \* **R<sup>2</sup>-type** statistics exist, but not recommended (difficult to interpret),
  - \* 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 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 dcl<sub>ox</sub> from model:

Model	$2 \ln L$	AIC	BIC	$G^2$	df	$P$
dneo##dcl <sub>ox</sub> ,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 dcl<sub>ox</sub> (despite non-significant main effect and weakly sign. interaction).

<sup>8</sup> Akaike's Information Criterion (AIC): computed as  $-2 \ln L + 2 \cdot \# \text{param.}$

<sup>9</sup> Bayesian Information Criterion (BIC): computed as  $-2 \ln L + \ln(n) \cdot \# \text{param.}$ , where  $n$  = number of observations.

<sup>10</sup> 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:**<sup>11</sup>  
 (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 lower (better) AIC	0 – 2	substantial
	4 – 7	considerably less
	> 10	essentially none

<sup>11</sup> 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 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<sup>12</sup>,
  - \* 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<sup>13</sup>; 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<sup>14</sup>.

<sup>12</sup> Not only for comparisons with a reference category.

<sup>13</sup> Discussed in more detail in VHM 802, see GO, Chapter 5; also, Chapter 26 of PSLS introduces the Tukey method.

<sup>14</sup> 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**  $\mu$  of principal interest (usually, the mean) to a more appropriate scale for linear modelling:

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

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),<sup>15</sup>
- multinomial distribution (ordinal data) – ordinal logistic regression,
- Poisson distribution<sup>16</sup> (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.

<sup>15</sup> Other link functions sometimes used for proportions: **probit** and **complementary log-log**.

<sup>16</sup> The so-called **negative binomial distribution** is also frequently used for count data.