Index of Lecture 5b: Logistic regression wrap-up

Practical Information

News/Schedule:

- today is the last (shared) regression session,
- schedule from next week onward: Thursday lectures and Monday labs,
- note: next Thursday (February 13), we have been moved to 280N.

Today's session:

- second logistic regression exercise (VER 16.2): review/discussion,
- what to use, or modify, from (logistic) model output in a manuscript? (4bL−10)
- logistic regression diagnostics (remaining parts of Lecture 5aL),
- no review planned of today's few extra slides (VER Sections 16.13−15), and no accompanying do-file.

What are we going to do with the last logistic regression exercise (VER 16.3)?

Sample Size Considerations for Logistic Modelling

Sample size calculation in logistic models is not straightforward,

- not widely available in standard software packages (e.g., Stata's comprehensive sample size module does not include logistic regression),
- a few calculators exist[1], but all with limited scope,
- with a single categorical predictor (in particular a binary predictor), methods for comparing proportions can be used,
- a general simulation approach to sample size calculation can be used (VER Section 2.11.8), but will require considerable programming effort.

Adequate sample size for multivariate models? — a general recommendation exists[2]:

> "one needs to have at least 10 events and 10 non-events per predictor parameter (including the intercept)",

- example: for the Nocardia dataset, the 54 cases (and 54 controls) would suffice for $\approx 5-6$ predictor parameters, including the intercept.

The simplicity of this rule has been much criticized, and it has been argued to be inappropriate for many situations.[3]

[1] Algorithms by E. Demidenko implemented at: `https://www.eugened.org/power-sample-size-calculator`.
[2] Based on Peduzzi et al. (1996), *Journal of Clinical Epidemiology* 99, 1373−1379.
[3] See e.g. a fairly recent article: Van Smeden et al. (2019), *Statistical Methods in Medical Research* 28, 2455−2474.

Exact Logistic Regression

Fact: Inference in logistic regression relies on "asymptotic[4] statistical results" but difficult to know ...

- when specific approximations are valid or not,[5]
- when a dataset is large enough for all approximations to be non-problematic.

Alternative approach: exact logistic regression — an entirely different way of conducting statistical inference (estimation, CI, tests) analogous to Fisher's exact test,

- computationally demanding ⇒ cannot do all analyses as exact logistic regression,
- limited availability in standard software packages (but Stata: `exlogistic` command),
- some parts of the analysis and results will appear differently than in ordinary logistic regression, despite the model being exactly the same.

Henrik's (pragmatic) view (absolutely not everyone will agree):

- maybe trying to compensate for a too small dataset with advanced computational methods is not ideal; e.g., a small dataset may also have limited external validity,
- trying to fit a large number of predictors into a small or moderate-sized dataset probably signals that the objective, rather than the analysis method, is wrong.

[4] Approximations valid for large sample sizes, where sample size can refer to the total number of observations or the number of replicates per covariate pattern.

[5] One general option is to conduct a simulation study based on a final model to confirm its inference — extra work...!

Conditional Logistic Regression

Matched case-control design:

- for each case: one (1:1) or several (1:m) controls are selected randomly from a subpopulation matched to that case,
- exposure variables are recorded for cases and controls,
- a matching attempts to make cases and controls equal on known confounders, thereby emphasizing their differences on exposure variables of interest,
- examples of matching variables: time, location, age, sex,
- all comparisons between cases and controls should be within (instead of across) the matched sets.

Conditional logistic regression = special type of logistic regression analysis:

- not the same as usual (or ordinary) logistic regression,
- correct analysis for matched case-control data.

Coverage of conditional logistic regression in VHM 802/812:

- introduction and demonstration by a worked example (VER 16.16; see 2019 course webpages for slides and do-file),
- useful to know about, but not trained specifically (or part of course curriculum).