

Index of Lecture 5b

Page	Title
0	Practical information
1	Introduction to conditional logistic regression (CLR)
2	Dataset <code>sal_outbrk</code> (VER)
3	Unconditional associations for matched data
4	CLR model
5	CLR analysis of <code>sal_outbrk</code> data
6	Statistical analysis of CLR model

PRACTICAL INFORMATION

Today's lecture: ,

- continuation of logistic regression model evaluation (Javier),
- last about prediction and `margins` command (Henrik),
- Exercises 2 + 3 for logistic regression (VER 16.2 + 16.3),
- questions and discussion on regression — the last combined session for VHM 802 and 812 for a long time,
- conditional logistic regression: for matched case-control studies.

INTRODUCTION TO

CONDITIONAL LOGISTIC REGRESSION (CLR)

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,
- matching attempts to make cases and controls equal on known confounders, thereby *emphasizing their difference on exposure variables of interest*,
- examples of matching variables: time, location, age, sex,
- all comparisons between cases and controls should be within (not 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 CLR in VHM 812/802:

- introduction and demonstration by an example,
- Section 16.15 in VER (about 3 pages),
- important to know about but not trained specifically.

DATASET SAL_OUTBRK (VER)

- subset of real dataset from a Salmonella typhimurium (phage type 12) outbreak investigation (Denmark 1996),
- 39 case persons (who got diseased) and 73 control persons (1-2 per case) matched for age, sex and residence (municipality),
- exposure variables determined by interviews,
- study aim: determine cause/agent of Salmonella outbreak.

Variable	Description	Values
match_grp	matched set id	(nominal)
casecontrol	case-control status	0/1 (control/case)
age	age (years)	2.53–64.44
gender	gender	0/1 (male/female)
eatbeef	ate beef in prev. 72 hours	0/1 (no/yes)
eatpork	ate pork in prev. 72 hours	0/1 (no/yes)
eatpoul	ate poultry in prev. 72 hours	0/1 (no/yes)
eateggs	ate eggs in prev. 72 hours	0/1 (no/yes)
slt_a	ate pork from sl.house A	0/1 (no/yes)
dlr_a	ate pork from wholesaler A	0/1 (no/yes)
...

Sample data:
(matched set
no. 23)

Variable	eatpork		eatbeef		slt_a		dlr_a	
Status	+	-	+	-	+	-	+	-
case	1	0	0	1	1	0	0	1
control	2	0	1	1	1	1	0	2

UNCONDITIONAL ASSOCIATIONS

FOR MATCHED DATA

Unconditional associations:

- initial step in model building/variable selection,
- can confirm/validate findings of complex model,
- for matched data: *usual methods inappropriate*.

Dichotomous (and categorical) exposure variables:

- use Mantel-Haenszel statistic stratified by matched sets (for 1:1 matching \sim McNemar's test; VER Chapter 13),
- sparse¹ tables in each matched set, but validity of statistic depends only on total counts².

Continuous exposure variables:

- 1:1 matching: use paired t -test or non-parametric test,
- 1: m matching: average exposure among controls, and use paired t -test or non-parametric test³.

Some examples (from `sal_outbrk` data):

- `eatpork`: M-H OR = 2.14 (0.69, 6.65), $P = 0.17$,
- `slt_a`: M-H OR = 3.87 (1.45, 10.3), $P = 0.002$.
- `d1r_a`: M-H OR = 7.75 (1.20, 49.9), $P = 0.004$.

¹ A sparse table has small counts, possibly one or several zeros.

² See Kleinbaum *et al.* (1982), *Epidemiologic Research*, Section 17.3.1.

³ A practical example is: Mortensen *et al.* (2002), *Prev. Vet. Med.* **53**, 83–101.

CLR MODEL⁴

$\text{logit}(p_i) = \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \alpha_{\text{stratum}(i)}$, where

- * p_i = prob. of subject i to be a case ($Y_i = 1$) given the predictors (x_{1i}, \dots, x_{ki}) ,
- * α -values: $\alpha_1, \dots, \alpha_s$ in the s strata (matched sets).

Comments:

- new: the stratum effects $\alpha_1, \dots, \alpha_s$,
- no intercept β_0 (but not of interest in case-control studies anyway),
- the β 's in (1) have same interpretation as usual,
- model formulated for prospective study, but the Y_i 's are not observed outcomes,
- additional assumption in (1) (relative to ordinary LR): additive stratum effects (on logit scale), i.e. same OR in all strata for each of the predictors.

Problems with model:

- large number of α -parameters difficult to estimate,
- less information in data than intuitively expected, e.g.
 - * a stratum where x_1 (say) is constant contributes no information about the OR for x_1 ,
- no stratum-level predictors allowed (but see L5b–6).

⁴ A detailed discussion of model and analysis can be found in: Hosmer & Lemeshow, *Applied Logistic Regression*, 2nd/3rd ed.

CLR ANALYSIS OF SAL_OUTBRK DATA

Unconditional associations:

- `slt_a` and `dlr_a`: $OR > 1$, $P \ll 0.05$,
- `eatpork` and `eateggs`: $OR > 1$, $0.10 < P < 0.20$,
- several missing values among predictors.

Best multivariable model: sole(!) predictor `slt_a`,

- $\hat{\beta}_1 = 1.485$, $SE = 0.518$, 95% CI: 0.470 – 2.50,
- $OR = e^{1.485} = 4.42$, $SE = 2.29$, 95% CI: 1.60 – 12.2
(note that $e^{0.47} = 1.60$ and $e^{2.5} = 12.2$),
- likelihood-ratio test: $G^2 = 10.0$, $df = 1$, $P = 0.002$.

Other models examined:

- added `dlr_a`: no significant effect of `dlr_a` ~ explains the same as `slt_a` (and is an intervening variable),
- added `eatpork` or `eateggs`: no significance of adding these variables,
- added interaction `sex*slt_a`: no significance ~ same effect of `slt_a` in men and women,
- added interaction `age*slt_a`: no significance ~ risk of getting sick not age-dependent.

Follow-up on analysis: *Salmonella typhimurium* (phage type 12) was isolated at slaughterhouse A!

STATISTICAL ANALYSIS OF CLR MODEL

Maximum likelihood (ML) estimation:

- usual ML does not work (too many parameters),
- conditional⁵ ML estimation:
eliminates the stratum parameters (sometimes called “nuisance parameters”⁶), and no estimates for stratum effects are obtained,
- ignoring the matching \sim ordinary LR (OLR):
potential loss of efficiency and bias — not recommended⁷
(possible alternative is OLR with matching variables as confounders, but its validity is not clear (HS)).

Analysis and model building:

- confidence intervals and tests based on likelihood (preferable) or Wald procedures (differences may be appreciable),
- model evaluation by residuals/diagnostics (CLR-specific):
in Stata 13 with `predict` or (add-on) `clfit` commands,
- model selection often by forward-type selection because of low information content in data,
- stratum-level predictors *can* be examined, but only in interactions with within-stratum predictors; evaluation of such interactions requires good data.

⁵ In each stratum, the likelihood is conditional on the number of cases and total number of subjects.

⁶ The term “nuisance” indicates the parameter to be of no real interest.

⁷ OLR for `s1t_a` model: $\hat{\beta}_1 = 1.168$ (SE = 0.416), OR = 3.21, 95% CI: 1.42 – 7.27.