# Index of Lecture 11a

## Overview: Dealing with Clustering

Detection of clustering:

- primarily through understanding of data structure,
- no recommended statistical tests for clustering![1]

Approaches to model clustering (in course):

- mixed (random effects) models – covered already,

Approaches to account for clustering (in course):

- fixed effects,
- stratification, for binary data (Mantel-Haenszel procedure),
- robust standard errors,
- generalized estimating equations (GEE) – the main topic of this lecture.

Choice of approach:

- ease of use (easy does not mean valid…),
- the assumptions required,
- (degree of) interest in clustering per se.

Datasets in lecture: 2-level simulated datasets + pig data.

[1] The reason is that even small (possibly statistically non-significant) clustering can have substantial impact (as discussed in terms of variance inflation), so we prefer to account for clustering whenever its presence makes biological sense.

## Fixed Effects Modelling

Fixed effects modelling: enter the herds (clusters) into the model as a categorical variable, to estimate a separate parameter for each herd (but one).

Advantages and disadvantages of fixed effects modelling:

- + generally very easy to carry out,
- + avoids distributional assumption about herd effects,
- + avoids taking the herds as representative for a population (which might be inappropriate if the number of herds is small),
- +/− estimates are cluster-specific and specific to actual herds in the study (cannot be generalized to a population),
- − does not allow for herd-level predictors,
- − does not give an estimate of the variance between herds,
- − may lead to biased estimates for other fixed effects when the number of herds is large, in particular for non-normal models.

Results for simulated datasets (cow-level predictor only):

- linear model: $\hat{\beta} = 4.968\,(.149)$ − close to mixed model,
- logistic model: $\hat{\beta} = 0.704\,(.046)$ − close to mixed model.

Estimates for the intercept $\sim$ reference herd $\Rightarrow$ not comparable to mixed models or models without herd effects.

## Stratification (Binary Data)

Stratification = Mantel-Haenszel procedures using herds (clusters) as strata:

– combined odds-ratio across herds (binary within-herd predictor; binary outcome),

– test for association in two-way table (categorical within-herd predictor; categorical outcome), adjusted for herds.

Advantages and disadvantages of stratification:

+ easy to carry out,

+ avoids any assumptions about the herds,

+/– cluster-specific estimates (OR),

– restricted scope and limited analysis,

– no insight into the type or magnitude of clustering.

Results for simulated dataset (cow-level predictor only)

- Mantel-Haenszel odds-ratio = 2.009 $\Rightarrow \hat{\beta} = \ln(2.009) = 0.698$ – close to mixed model,
- estimated SE = .046, also close to mixed model (computed by backtransforming OR and its CI bounds to logit scale [2]).

[2] The CI for the M-H OR was constructed on logit scale, where estimates follow a normal distribution to a good approximation, and transformed to odds-ratio scale.

# Robust Standard Errors

Background: usual (model-based) statistical methods:

$$\text{data} \mapsto \text{model} \mapsto \begin{cases} \text{estimates, test statistics} \\ \text{standard errors, } P\text{-values} \end{cases}$$

- statistical models are not always (never) true!
- robust methods are designed to be less sensitive to model deviations.

Robust variance estimation (Huber-White, "sandwich"):

- base SEs on properties of the estimation method valid for a wider class of models than assumed,
- variance estimate can be split across groups if the assumption of within-group independence is critical.

Advantages and disadvantages of robust SE method:

+ simple to use (readily available in Stata),

+ general method, extends to other models,

+ possible to use knowledge about clusters,

+/− robust SEs have different interpretation than usual SEs,

+/− does not affect the estimate, only its standard error,

− no insight into the type or magnitude of clustering.

Results: fairly close to mixed model SEs, except for linear model with predictor at herd level (1.712 vs. 1.496).

## Generalized Estimating Equations (GEE)

Initial remarks about GEE for GLMs[3]:

- an estimation procedure (set of equations from which estimates are constructed iteratively) rather than a model,
- partially specified model involving only assumptions about the *marginal*[4] means and variances,
- gives population-averaged (or marginal) estimates,
- no herd effects, hence no assumed distribution for them,
- no likelihood function or likelihood-based inference.

Original form of GEE for GLMs:

- framework of longitudinal data (repeated measures),
- algorithm estimates a working correlation matrix for the correlation of obs. within clusters (herds, subjects):
  - ∗ a setting in the algorithm, not a model assumption,
  - ∗ hierarchical data: use exchangeable type where all observations in a cluster are equally correlated,
- invented in the 1980s and much used since.

[3] Recall that generalized linear models (GLMs) constitute a general class of models including logistic and Poisson regression, specified by a distribution (family) and a link function.

[4] "Marginal" refers to the mean across the population of clusters (herds).

## Working Correlation Matrices

For a series $Y = (Y_1, \ldots, Y_n)$ of observations on a subject, the correlation matrix $\mathrm{Corr}(Y)$ is the $n \times n$-matrix of all pairs of correlations,[5]

$$\mathrm{Corr}(Y) = \begin{pmatrix} 1 & & & & \\ \mathrm{Corr}(Y_1, Y_2) & 1 & & & \\ \mathrm{Corr}(Y_1, Y_3) & \mathrm{Corr}(Y_2, Y_3) & 1 & & \\ \vdots & \vdots & \vdots & \vdots & \\ \mathrm{Corr}(Y_1, Y_n) & \mathrm{Corr}(Y_2, Y_n) & \mathrm{Corr}(Y_3, Y_n) & \cdots & 1 \end{pmatrix}.$$

Examples of working correlation matrices commonly used with GEE (shown for 4 observations):

- independence: $\sim$ independent or uncorrelated obs., or no assumptions about $\mathrm{Corr}(Y)$
  $$\mathrm{Corr}(Y) = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{pmatrix},$$
- exchangeable: $\sim$ hierarchical data struct. with $\rho = \mathrm{ICC}$ (also compound symmetry)
  $$\mathrm{Corr}(Y) = \begin{pmatrix} 1 & & & \\ \rho & 1 & & \\ \rho & \rho & 1 & \\ \rho & \rho & \rho & 1 \end{pmatrix},$$
- autoregressive `ar(1)`: $\sim$ first order autoreg., for repeated measures, has decaying corr. with time
  $$\mathrm{Corr}(Y) = \begin{pmatrix} 1 & & & \\ \rho & 1 & & \\ \rho^2 & \rho & 1 & \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$
- Toeplitz: $\sim$ "stationary" ($\rho$ depends on distance only)
  $$\mathrm{Corr}(Y) = \begin{pmatrix} 1 & & & \\ \rho_1 & 1 & & \\ \rho_2 & \rho_1 & 1 & \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}.$$

[5] As the matrix is symmetric, for clarity the values above the diagonal are left blank.

## GEE in Practice

GEE settings:

- strongly recommended to use robust standard errors instead of model-based ones (caution: not the Stata default):
  note: GEE with indep. corr. matrix $\sim$ robust variance estimation,
- hypothesis testing: Wald tests mostly used,
- performance in finite samples: standard guideline on sample size is that at least 30 clusters is needed to avoid biases,
- normally distributed outcomes: GEE applies[6] but often linear mixed models are preferred, due to the more clear-cut inference.

Guidelines for choice of (working) correlation structure (Hardin and Hilbe (2003), *Generalized Estimating Equations*, p. 141-42):

- for small cluster size and complete data, use unstructured,
- for repeated measures over time, use a structure with time dependence,
- for 2-level hierarchical structure, use exchangeable,
- for small number of clusters, indep. structure may be best,
- if more than one structure meets the guidelines, use the QIC statistic (AIC analogue) to choose between them,
- unstructured correlations may be used for exploratory analysis.

Results (exchangeable corr. structure): very close to linear mixed model (continuous data), but population-averaged estimates for binary data: 0.559 (.177) and 0.569 (.042) for herd- and cow-level $X$.

[6] With identity link, there is no distinction between cluster-specific and population-averaged estimates.

## Summary for GEE

Advantages and disadvantages of (classical) GEE method:

+ no assumptions about herd effects,
+ good/robust theoretical properties[7] (with robust SE),
+ computationally feasible for large data sets,
+ possible/necessary to use knowledge of clustering,

+/− population-averaged instead of cluster-specific estimates,

− less flexible with respect to multiple levels,

− no direct modelling of correlation structure.

− statistical choice of “working corr. structure” less clear.

Different versions of GEE for GLMs:

- differ by algorithms and settings for the part of the algorithm dealing with correlation structure – one important example: alternating logistic regression (ALR):
  - ∗ GEE-type procedure based on odds-ratios rather than working correlations (arguably more appropriate for binary data),
  - ∗ allows for either 2 or 3 hierarchical levels,
  - ∗ recommended (Hardin and Hilbe, 2013) over ordinary GEE for logistic regression if “the focus of the analysis does include the association parameters” (i.e., assoc. within clusters),
  - ∗ implemented in SAS and R, but not Stata.

[7] Parameter estimates and SEs are asymptotically unbiased (when # clusters is large) under weak conditions; also, misspecification of the working correlation matrix does not invalidate the estimates, but may lead to some loss of efficiency.

## Summary of Analyses for Pig Data

Estimates and SE (on logit scale):

| | effect of `ar_g1` | | intercept | |
|---|---|---|---|---|
| Model | Coef. | SE | Coef. | SE |
| ordinary LR | 0.647 | 0.220 | -0.145 | 0.156 |
| fixed effects LR | 0.365 | 0.268 | N/A | |
| Mantel-Haenszel | 0.346 | 0.261 | N/A | |
| robust variance | 0.647 | 0.267 | -0.145 | 0.279 |
| random effects LR | 0.437 | 0.258 | 0.020 | 0.301 |
| GEE (`exch` corr.) | 0.354 | 0.215 | 0.018 | 0.271 |

Conclusion:

- estimate by random effects model (GLMM) somewhat larger than for GEE; more than explained by being a cluster-specific estimate, as seen from

$$0.437/\sqrt{1 + 0.346 \cdot 0.877} = 0.383,$$

  but not critically off considering the SE,
- Mantel-Haenszel and fixed effects estimates should be closer to GLMM than GEE estimates; some herd-level confounding (for `ar_g1`) appears to exist in the data,
- ordinary LR and robust SE considerably off, probably due to herd confounding.

## Means/Margins for Mixed Models

Basic fact: decisions are *required* (or implicit) on how to deal with the random effects.

Simplest situation: predictions on same scale as the random effects (say $u$):

- setting $u = 0$ corresponds to predictions that are means in the random effects distribution,[8]
    - similar to setting $\varepsilon = 0$ for prediction in linear models,
    - assumes prediction is for new "cluster" from population, as opposed to cluster(s) in the dataset,[9]
    - gives largest uncertainty (SE) for predictions because nothing is known about random effects from the data,
    - usual software default, e.g. Stata's `margins` command.

More complex situation: (non-linear) transformation of means, e.g. if outcome was transformed for mixed model analysis:

- similar situation as in linear models, where (back)transforming means gives medians, not means,
- for $u=0$, interpret transformed values as medians,
- exact means can be obtained analytically (using formulae) for some transformations and generally by simulation; in practice, the median interpretation is often sufficient and satisfactory.

---

[8] With the usual assumption that random effects have mean 0, e.g. $u \sim \mathrm{N}(0, \sigma_h^2)$.
[9] Sometimes termed referred to as "broad" and "narrow" inference spaces, respectively; Littell et al. (2006), *SAS for Mixed Models*, 2nd ed., SAS Publishing.

## MEANS/MARGINS FOR GLMMs AND GEE

Main issue: distinction between CS (cluster-specific) and PA (population-averaged) interpretations,

- PA predictions usually desired for broad[9] inference space,
  - ∗ directly with GEE (e.g. Stata's `margins` command),
  - ∗ caution needed (see below) for GLMMs,
- CS predictions may be of interest for specific clusters (included in the data), e.g. farms or schools — only available for GLMMs (technical: VER2/MER Ex. 22.10).

Common software limitations (Stata 13, SAS, R): no predictions on original scale in GLMMs,[10]

- can work around these by manual backtransformation, but it will not give proper PA predictions: *medians* instead of *means* (as discussed on previous page), and thus not truly population-averaged,
- analytical adjustment to achieve PA estimates may be possible, e.g. in logistic regression[11].

[10] In Stata 14+, only the `meqrlogit` and `meqrpoisson` commands do not allow estimation on original scale.

[11] Using the standard approximation formula in logistic regression models:

$$\beta^{\mathrm{PA}} \approx \beta^{\mathrm{SS}} / \sqrt{1 + 0.346\,\sigma_h^2}\,.$$