

## Index of Lecture 11b

Page	Title
1	Overview: dealing with clustering
2	Fixed effects modelling
3	Stratification (binary data)
4	Robust standard errors
5	Generalized estimating equations (GEE)
6	Working correlation matrices
7	GEE in practice
8	Summary for GEE
9	Summary of analyses for pig data
10	Means/margins for mixed models
11	Means/margins for GLMMs and GEE

## OVERVIEW: DEALING WITH CLUSTERING

### Detection of clustering:

- primarily through understanding of data structure,
- no recommended statistical tests for clustering!<sup>1</sup>

### 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 (only partially valid reason),
- the assumptions required,
- interest in clustering per se.

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

<sup>1</sup> 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 fixed 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 <sup>2</sup>).

---

<sup>2</sup> 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<sup>3</sup>:

- 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*<sup>4</sup> 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.

---

<sup>3</sup> 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.

<sup>4</sup> “Marginal” refers to the mean across the population of clusters (herds).

## WORKING CORRELATION MATRICES

For a series  $Y = (Y_1, \dots, Y_n)$  of observations on a subject, the correlation matrix  $\text{Corr}(Y)$  is the  $n \times n$ -matrix of all pairs of correlations,<sup>5</sup>

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

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

- independence:  
 $\sim$  independent or uncorrelated obs., or no assumptions about  $\text{Corr}(Y)$

$$\text{Corr}(Y) = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

- exchangeable:  
 $\sim$  hierarchical data struct. with  $\rho = \text{ICC}$  (also compound symmetry)

$$\text{Corr}(Y) = \begin{pmatrix} 1 & & & \\ \rho & 1 & & \\ \rho & \rho & 1 & \\ \rho & \rho & \rho & 1 \end{pmatrix},$$

- autoregressive ar(1):  
 $\sim$  first order autoreg., for repeated measures, has decaying corr. with time

$$\text{Corr}(Y) = \begin{pmatrix} 1 & & & \\ \rho & 1 & & \\ \rho^2 & \rho & 1 & \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$

- Toeplitz:  
 $\sim$  “stationary” ( $\rho$  depends on distance only)

$$\text{Corr}(Y) = \begin{pmatrix} 1 & & & \\ \rho_1 & 1 & & \\ \rho_2 & \rho_1 & 1 & \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}.$$

<sup>5</sup> 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<sup>6</sup> 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  $X$ .

---

<sup>6</sup> 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<sup>7</sup> (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”,
  - \* implemented in SAS and R, but not Stata.

---

<sup>7</sup> 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):

Model	effect of <code>ar_g1</code>		intercept	
	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 ( <code>exch</code> 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 appears to exist in the data,
- ordinary LR and robust SE considerably off, probably due to strong herd confounding (for `ar_g1`).

## 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,<sup>8</sup>
  - \* 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,<sup>9</sup>
  - \* 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, e.g. if outcome was transformed prior to mixed model analysis:

- setting  $u = 0$  no longer gives means in random effects distribution but predictions can be interpreted 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.

---

<sup>8</sup> With the usual assumption that random effects have mean 0, e.g.  $u \sim N(0, \sigma_h^2)$ .

<sup>9</sup> 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<sup>9</sup> 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.

Common software limitations (Stata, SAS):

no predictions on original scale in GLMMs,<sup>10</sup>

- 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<sup>11</sup>.

---

<sup>10</sup> After `meqrlogit` in Stata, the `margins` command gives the error message: “default prediction is a function of possibly stochastic quantities other than e(b)”.

<sup>11</sup> Using the standard approximation formula in logistic regression models:

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