

## Index of Lecture 12b2: Alternative methods, incl. generalized estimating equations

Page	Title
1	Practical information
2	Overview: dealing with clustering
3	Fixed effects modelling
4	Stratification (binary data)
5	Robust standard errors
6	Generalized estimating equations (GEE)
7	Working correlation matrices
8	GEE inference and variants
9	Choice of correlation “model”
10	Summary for GEE
11	Summary of analyses for simulated data
12	Summary of analyses for pig data
13	Means/margins for (linear) mixed models
14	Means/margins for GLMMs and GEE

## PRACTICAL INFORMATION

### Today's lecture:

- “**alternative** methods” — to the linear and logistic **mixed models** already covered:
  - \* perhaps more relevant for discrete than continuous data<sup>1</sup>,
  - \* brief review of several methods, contrasting their properties,
  - \* main focus on **generalized estimating equations** (but really just an introduction to a large topic),
  - \* illustrations and comparisons of methods by familiar **2-level datasets**: simulated datasets (Chapter 20) and pig\_adg data.
- **textbook coverage** for lecture:
  - \* alternative methods: VER2/MER, Section 20.5,
  - \* generalized estimating equations: VER2/MER, Section 23.5.

### Schedule — next session Friday (April 9), 9-11am, on VHM 812 stream:

- a **review** session for both recorded video lectures (with discussion, and potential software demonstrations),
- **practice problem** (VER20) posted for earlier sessions, and we may also touch briefly on VER22 exercise.

---

<sup>1</sup> Linear mixed models remain the most popular choice when its assumptions can be met.

## OVERVIEW: DEALING WITH CLUSTERING

### Detection of clustering:

- primarily through **understanding of data structure**,
- statistical testing for clustering is **not recommended!**<sup>2</sup>

### Approaches to model clustering (in the course):

- mixed (random effects) models — covered already,

### Approaches to account for clustering (in the course):

- fixed effects,
- stratification, for binary data (Mantel-Haenszel procedure),
- robust standard errors,
- **generalized estimating equations** (GEE).

### Choice of approach should take into account:

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

---

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

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

---

<sup>3</sup> 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 (OR) 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):

- o Mantel-Haenszel  $OR = 2.009 \Rightarrow \hat{\beta} = \ln(2.009) = 0.698$  – close to mixed model,
- o estimated  $SE = .046$ , also close to mixed model (computed by backtransforming OR and its confidence interval bounds to logit scale<sup>4</sup>).

---

<sup>4</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:

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

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

### Robust variance estimation (Huber-White, “sandwich”):

- o base SEs on properties of the estimation method valid for a wider class of models than assumed,
- o method can be targeted to deal with clustering, where the assumption of within-group independence is critical (only this version is of interest here).

### Advantages and disadvantages of robust SE method:

- + simple to use and general method (available in Stata for wide range of models),
- +/- robust SEs have a different interpretation than usual SEs,
- +/- **does not affect the estimate**, only its standard error,
  - gives no insight into the type or magnitude of clustering.

**Results:** fairly close to mixed model SEs (except: linear model, herd-level  $X$ ).

## GENERALIZED ESTIMATING EQUATIONS (GEE)

Initial remarks about GEE for GLMs<sup>5</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>6</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 observations within clusters (herds, subjects):
  - \* a setting in the algorithm, not a model assumption,
  - \* **hierarchical data**: use exchangeable type ( $\sim$  equal correlations  $\rightarrow$  next slide),
- invented in the 1980s and very much used since.

---

<sup>5</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>6</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  $\sim Y$  as a multivariate outcome.

**Examples of working correlation matrices**<sup>7</sup> commonly used with GEE (shown for 4 observations per subject):

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

$$\text{Corr}(Y_i, Y_j) = 0 \quad \text{for } i \neq j,$$

- **exchangeable:**  
 $\sim$  hierarchical data structure,  
 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 autoregressive,  
 for repeated measures, has  
 decaying correlation 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>7</sup> As the matrices are symmetric, for clarity the values above the diagonal are left blank.

## GEE INFERENCE AND VARIANTS

### GEE settings:

- strongly recommended to use **robust standard errors** instead of model-based ones (**caution**: not the Stata default),
- **note**: GEE with independent correlation matrix  $\sim$  robust variance estimation!,
- **hypothesis testing**: Wald tests mostly used (other tests exist),
- parameter estimates and SEs are **asymptotically unbiased** (when # clusters is large) under weak conditions,<sup>8</sup>
- **performance in finite samples**: standard guideline on sample size is that at least 30 clusters is needed to avoid biases,

### 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,
  - \* implemented in SAS and R, but not Stata.

---

<sup>8</sup> Importantly, **misspecification** of the working correlation matrix **does not invalidate** the estimates, but may lead to some loss of efficiency.

## CHOICE OF CORRELATION “MODEL”

How to choose the “best” (working) correlation structure for GEE?

- in principle<sup>9</sup>, the QIC statistic (an analog of the AIC) can be compared between different correlation structures, and the one with smaller QIC value is preferable,
- **guidelines** (from Hardin and Hilbe (2003), *Generalized Estimating Equations*, 1st ed., p. 141–42):
  - \* for small cluster size and complete data, use unstructured,
  - \* for repeated measures over time, use a structure with time dependence, e.g. ar(1),
  - \* for 2-level hierarchical structure, use exchangeable,
  - \* for small number of clusters, independence structure may be best,
  - \* if more than one structure meets the guidelines, use the QIC statistic to choose between them,
  - \* unstructured correlations may be used for exploratory analysis.
- another **recommendation** (Hardin & Hilbe): **alternating logistic regression** is preferable to ordinary GEE for logistic regression if “the focus of the analysis does include the association parameters” (i.e., association within clusters).

---

<sup>9</sup> My (Henrik’s) experience with the QIC statistic has been mixed, and I don’t feel comfortable with using QIC to choose the correlation structure.

## SUMMARY FOR GEE

**Advantages and disadvantages** of (classical) GEE method:

- + no assumptions about herd effects,
- + good/robust theoretical properties (with robust SE),
- + computationally feasible for large data sets,
- + possible/necessary to use knowledge of clustering,
- +/- **population-averaged** instead of cluster-specific estimates,<sup>10</sup>
  - less flexible with respect to multiple levels,
  - no direct modelling of correlation structure.
  - statistical choice of “working correlation structure” less clear,
  - choice between different GEE variants remains “subjective”.

**Results** (exchangeable correlation 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>10</sup> Recall that with identity link, e.g. linear models, there is no distinction between PA and CS estimates  
⇒ GEE is perfectly valid for normally distributed outcomes, but often linear mixed models are preferred, due to their more clear-cut inference and larger flexibility.

## SUMMARY OF ANALYSES FOR SIMULATED DATASETS

Collected **results** for the 4 datasets (two outcomes, two predictor levels):

Estimate (SE)	Continuous outcome (true= 5)		Binary outcome (true= 0.693) <sup>a</sup>	
Model/approach	herd-level $X$	cow-level $X$	herd-level $X$	cow-level $X$
unadjusted	3.557 (.200)	4.982 (.199)	0.529 (.042)	0.586 (.042)
mixed	3.796 (1.50)	4.968 (.149)	0.620 (.204)	0.697 (.046)
fixed effects	n/a	4.968 (.149)	n/a	0.704 (.046)
stratification	n/a	not discussed	n/a	0.698 (.046)
robust SE	3.557 (1.71)	4.982 (.142)	0.529 (.211)	0.586 (.044)
GEE (exch. corr)	3.797 (1.49)	4.968 (.141)	0.559 (.177)	0.569 (.042)

<sup>a</sup> cluster-specific true value, population-averaged equivalent  $\approx 0.597$

(computed as  $0.693/\sqrt{1 + 0.346 \cdot 1} = 0.597$ .)

### Extra notes:

- **cow-level predictor**: due to large sample size for  $X$ , all estimates are close to their (respective) true values,
- **herd-level predictor**: with an effective sample size of only 50 replicates per group, no close agreement with the true values can be expected.

## SUMMARY OF ANALYSES FOR PIG DATA

Estimates and SE (on logit scale):

Model	effect of ar_g1		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 (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 (LINEAR) 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>11</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>12</sup>
  - \* gives largest uncertainty (SE) for predictions because nothing is known about random effects from the data,
  - \* usual software default, e.g. in Stata’s margins command.

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

- similar to 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.

---

<sup>11</sup> With the usual assumption that random effects have mean 0, e.g.  $u \sim N(0, \sigma_h^2)$ .

<sup>12</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>12</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 (technical: VER2/MER Ex. 22.10).

**Common software limitations** (Stata 13, SAS, R): no predictions on original scale in GLMMs,<sup>13</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>14</sup>.

---

<sup>13</sup> In Stata 14+, only the `meqrlogit` and `meqrpoisson` commands do not allow estimation on original scale.

<sup>14</sup> Using the (by now) well-known approximation formula in logistic regression models:

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

SUMMARY OF ANALYSES FOR SIMULATED DATASETS

Collected **results** for the 4 datasets (two outcomes, two predictor levels):

Estimate (SE)	Continuous outcome (true= 5)		Binary outcome (true= 0.693) <sup>a</sup>	
Model/approach	herd-level $X$	cow-level $X$	herd-level $X$	cow-level $X$
unadjusted	3.557 (.200)	4.982 (.199)	0.529 (.042)	0.586 (.042)
mixed	3.796 (1.50)	4.968 (.149)	0.620 (.204)	0.697 (.046)
fixed effects	n/a	4.968 (.149)	n/a	0.704 (.046)

<sup>a</sup> cluster-specific true value, population-averaged equivalent  $\approx 0.597$

(computed as  $0.693/\sqrt{1 + 0.346 \cdot 1} = 0.597$ .)

SUMMARY OF ANALYSES FOR SIMULATED DATASETS

Collected **results** for the 4 datasets (two outcomes, two predictor levels):

<b>Estimate (SE)</b>	<b>Continuous</b> outcome (true= 5)		<b>Binary</b> outcome (true= 0.693) <sup>a</sup>	
<b>Model/approach</b>	herd-level $X$	cow-level $X$	herd-level $X$	cow-level $X$
unadjusted	3.557 (.200)	4.982 (.199)	0.529 (.042)	0.586 (.042)
mixed	3.796 (1.50)	4.968 (.149)	0.620 (.204)	0.697 (.046)
fixed effects	n/a	4.968 (.149)	n/a	0.704 (.046)
stratification	n/a	not discussed	n/a	0.698 (.046)

<sup>a</sup> cluster-specific true value, population-averaged equivalent  $\approx 0.597$

(computed as  $0.693/\sqrt{1 + 0.346 \cdot 1} = 0.597$ .)

SUMMARY OF ANALYSES FOR SIMULATED DATASETS

Collected **results** for the 4 datasets (two outcomes, two predictor levels):

Estimate (SE)	Continuous outcome (true= 5)		Binary outcome (true= 0.693) <sup>a</sup>	
Model/approach	herd-level $X$	cow-level $X$	herd-level $X$	cow-level $X$
unadjusted	3.557 (.200)	4.982 (.199)	0.529 (.042)	0.586 (.042)
mixed	3.796 (1.50)	4.968 (.149)	0.620 (.204)	0.697 (.046)
fixed effects	n/a	4.968 (.149)	n/a	0.704 (.046)
stratification	n/a	not discussed	n/a	0.698 (.046)
robust SE	3.557 (1.71)	4.982 (.142)	0.529 (.211)	0.586 (.044)

<sup>a</sup> cluster-specific true value, population-averaged equivalent  $\approx 0.597$

(computed as  $0.693/\sqrt{1 + 0.346 \cdot 1} = 0.597$ .)