

**Menelaos Pavlou<sup>1</sup>**

Andrew Copas<sup>2</sup>

Shaun Seaman<sup>3</sup>

**Efficient weighted generalised estimating equations  
when the cluster size or covariate structure are  
informative**

1. University College London, UK; 2. MRC Clinical Trials Unit, London, UK
3. MRC Biostatistics Unit, Cambridge, UK

# Overview

- Informative cluster size – definition.
- Example - teeth data.
- Possible populations for inference.
- Informative covariate structure - an example.
- Estimation: Cluster Weighted GEE (WIGEE) with independence correlation matrix.
- Bias from use of realistic correlation matrix.
- Adapting WIGEE for efficient and unbiased inference with realistic correlation matrix (WRGEE).
- Conclusions.

## Informative Cluster size

- Informative cluster size definition (repeated measures)
  - $y_{ij}$  is the outcome variable for cluster  $i$ , unit  $j$ .
  - $\mathbf{x}_{ij}$  is the vector of explanatory variables for cluster  $i$ , unit  $j$ .
  - $n_i$  is the cluster size, number of units/measurements.
  
- Informative cluster size arises :
  - (i) When outcome not independent of the cluster size, i.e.  $E(Y|X,n) \neq E(Y|X)$ .
  - (ii) Cluster size variability is an inherent feature of the data.
  - (iii) Not scientifically meaningful to include  $n$  as part of  $X$ .

## Informative Cluster Size - Example

- *Dental Studies* – Explore factors associated with periodontal disease :
  - mouth = cluster.
  - tooth = unit.
- (i) Dental health (disease status of teeth) may be associated with number of teeth, because factors causing disease also cause tooth loss.
- (ii) No “missing data”, but... variable cluster size.

## 3 Populations for inference

1. *Population of measurements (PM)*: all measurements.
2. *Population of clusters 1 (PC1)*: one measurement (i.e. one ‘unit’) selected at random from each cluster.
3. *Population of clusters 2 (PC2)*: one measurement selected with each distinct value of  $X$  from each cluster.

Example:

- $Y$  health outcome,
- $X$  therapy , patients switch therapies.
- Fewer measurements  $\rightarrow$  Earlier switch & Higher Outcome  $Y$   
 $\rightarrow$  Informative Cluster size & non “size balanced” covariate

## 3 Populations for inference

- PC1 and PC2 are identical if  $X$  is either:
  - cluster constant, or
  - categorical cluster varying but '*size balanced*' – distribution of  $X$  same across all cluster sizes.
- PC2 and PM are identical for cluster varying continuous  $X$ .
- PC1 and PM differ under *informative cluster size*.

## Informative Covariate Structure - Example

- *Informative covariate structure*: number of measurements before switch linked to health status before switch.
- Can arise even when cluster sizes are equal for all clusters.
- Define  $n_x$  : number of measurements which share common values  $X=x$  within a cluster,  
e.g.  $X_i=(0,0,1,1,1,1)$ ,  $n_0=2$ ,  $n_1=4$ ,  $n=6$ .
- Informative covariate structure arises when  $E(Y|X, n_x, n) \neq E(Y|X, n)$ .

# Generalized Estimating Equations

- Repeated Measures – Fit Marginal models of the form  $g[E(Y)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$ ,  $g$  is the link function.
- Model the mean and the covariance structure.
- Similar to score equations if data were multivariate normal.

$$U(\beta, \rho) = \sum_{i=1}^N \left( \frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} (Y_i - \mu_i) = 0$$

- Variance  $V$  must be “guessed” and modelled.
- Use of working correlation matrix  $R(\rho)$  to increase efficiency.
- $\rho$  = correlation parameter (nuisance),  $\phi$  = scale parameter.

# Marginal models and Populations for inference

- Weighted Independence GEE (WIGEE) with weights selected according to the population for inference:

$$U(\beta, \rho) = \sum_{i=1}^N \left( \frac{\partial \mu_i}{\partial \beta} \right)^T \boxed{W_i} V_i^{-1} (Y_i - \mu_i) = 0$$

$$V_i = \phi A_i^{1/2} R_i(\rho) A_i^{1/2}$$

$$A_i = \text{diag}(V(\mu_i))$$

- Population of measurements :**  $W_i = I_{n_i}$
- Population of clusters 1:**  $W_i = \frac{I_{n_i}}{n_i}$
- Population of clusters 2:**  $W_i$  diagonal,  $w_i(j, j)(x) = \frac{I(X = x)}{n_x}$

**!** The correlation structure assumed MUST be independence.

- Justification: PC1 and PC2 arise by selection from PM, these weights are *inverse selection probabilities*...

# Why not use a realistic correlation matrix in general?

- Inverse correlation matrix in GEE can be viewed as a weight matrix
- Use of independence correlation matrix, all measurements equally weighted.
- Issue: Informative cluster size & realistic correlation matrix, weights to each measurement altered, total cluster weight changes, inference about none of populations!

i.e. Informative cluster size/structure & Realistic correlation  $\rightarrow$  Biased estimation!

- *Proposal*: In special cases however, use realistic working correlation matrix with further weighting to be unbiased.

# Special Cases: Efficient Weighted GEE (WRGEE) for ICS

- **Cluster constant** covariates  $\longrightarrow$  unbiased GEE based on realistic correlation matrix and additional weights

$$U(\beta, \rho) = \sum_{i=1}^N \left( \frac{\partial \mu_i}{\partial \beta} \right)^T W_i^* V_i^{-1} (Y_i - \mu_i) = 0$$

- $W_i^*$  diagonal matrix with cluster constant elements. It can be proved that these are:

$$PM: w_i^*(j,j) = \frac{n_i}{\text{sum}(R_i^{-1})} \qquad PC: w_i^*(j,j) = \frac{1}{\text{sum}(R_i^{-1})}$$

## Special case: WRGEE for Informative Covariate Structure

- When  $X$  is categorical cluster varying and the covariate structure is informative we suggest using a block diagonal correlation matrix according to the sub-clusters defined by common values of the cluster varying covariate.
- I.e. split the cluster into these sub-clusters and apply the previous method for cluster constant covariates

## Concluding remarks 1

- Previous authors have defined informative cluster size, and proposed WIGEE for PM and PC1.
- Concentrated on covariates  $X$  either cluster constant or cluster size balanced.
- *Define* PC2, more intuitive [as PC1 in special cases] for cluster varying  $X$ .
- *Define* “Informative Covariate Structure”, can arise when  $X$  is cluster varying categorical.
- As with informative cluster size, standard GEE can be biased under informative covariate structure and WIGEE can be recommended.

## Concluding remarks 2

- In special cases, GEE with a realistic correlation matrix and additional weights.
- Variable efficiency gains, Depending on the nature of X: highest for X cluster varying continuous and size balanced.
- Caution when using WIGEE (or WRGEE): lower efficiency than standard GEE if cluster size/structure are non-informative.
- When cluster size varies, use these special methods only used where it is strongly suspected that they are required.

## Supplement if needed - Example on weights

- Suppose that  $X_i=(0,0,1,1,1,1)$
- Population of measurements :  $W_i=\text{diag}(1,1,1,1,1,1)$
- Population of clusters 1:  $W_i=\text{diag}(1/6,1/6,1/6,1/6,1/6,1/6)$ .
- Population of clusters 2:  $W_i=\text{diag}(1/2,1/2,1/4,1/4,1/4,1/4)$ .

## Supplement if needed - Simulation Studies

- Informative Cluster Size – Cluster constant covariates: PM, PC1  
→ Efficiency Gains up to 8%.
- Informative Cluster Size – Non Size balanced Cluster varying categorical covariates: PM, PC1, PC2.  
→ Efficiency Gains up to 25%.
- Informative Covariate Structure:  
→ Efficiency Gains up to 25%.