Clusters with random size: maximum likelihood versus weighted estimation

Lisa Hermans[1], Geert Molenberghs[12], Michael G. Kenward[3], Wim Van der Elst[1], Vahid Nassiri[2], Marc Aerts[1], Geert Verbeke[12]

[1] I-BioStat, Universiteit Hasselt, Belgium
[2] I-BioStat, KU Leuven, Belgium
[3] Dept. of Medical Statistics, London School of Hygiene & Tropical Medicine, UK

E-mail for correspondence: lisa.hermans@uhasselt.be

Abstract: There are many contemporary designs that do not use a random sample of a fixed, *a priori* determined size. In case of informative cluster sizes, the cluster size is influenced by the the cluster's data, but here we cope with some issues that even occur when the cluster size and the data are unrelated. First, fitting models to clusters of varying sizes is often more complicated than when all cluster have the same size. Secondly, in such cases, there usually is no so-called complete sufficient statistic (Molenberghs et al., 2014).

Keywords: Likelihood inference; Pseudo-likelihood; Random cluster size.

1 Introduction

In applied statistics, situations exist where there is no fixed sample size. Molenberghs et al. (2014) provide an overview of various situations. Examples include: sequential trials, incomplete data, censored survival data, etc. Here we focus on hierarchical or clustered data. Random cluster sizes can occur for any outcome type, including continuous data, binary data, counts, and failure times. We will focus on cases where the cluster size is variable but independent of observed and unobserved outcomes. As a simple cluster paradigm, we consider the normal compound-symmetry (CS) model.
Molenberghs et al. (2011) introduced the split-sample methodology, i.e., a form of pseudo-likelihood where a sample is subdivided into subsamples. These subsamples are analyzed as if they were unrelated and afterwards the

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

results are appropriately averaged. There are many options for splitting the data, but here we use splitting along the cluster sizes. For the subsamples, closed-form estimators then exist, whereas they do not in the sample as a whole. A weighted combination of the subgroup-specific estimators is needed. However, Molenberghs et al. (2014) and Hermans et al. (2014) show that there may not be an optimal set of weights, resulting from calculations on incomplete sufficient statistics in the context of weighted averages.

2 Split-sample methods for clusters of variable size

2.1 Compound-symmetry model

Let $\boldsymbol{Y}$ be a vector of length n, following the compound-symmetry normal law $\boldsymbol{Y} \sim N(\mu \mathbf{1}_n, \sigma^2 \boldsymbol{I}_n + d\boldsymbol{J}_n)$. In general, both $\boldsymbol{Y}$ and n are random variables. Let there be a sample of N independent clusters, with K different cluster sizes n_k $(k = 1, \ldots, K)$ with multiplicity c_k. Denote the outcome vector for the ith $(i = 1, \ldots, c_k)$ replicate among the cluster of size n_k by $\boldsymbol{Y}_i^{(k)}$. For a sample with constant cluster size $(K = 1)$, compound-symmetry models allow closed-form solutions for the estimators. These sufficient statistics are complete, the estimator is unique minimum variance unbiased, and the mean parameter estimator and the variance parameter estimator are independent. Hermans et al. (2015) show, based on likelihood calculations, that in case $K \geq 2$ all these results disappear and there is no closed form solution. Likelihood calculations for K cluster sizes with common mean and variance parameter across all clusters, do not lead to explicit solutions, unless the variance components are known or the cluster size is constant. This suggest further study of weighted averages, e.g., of the form

$$\widetilde{\mu} = \sum_{k=1}^{K} a_k \widehat{\mu_k}, \qquad \widetilde{\sigma^2} = \sum_{k=1}^{K} b_k \widehat{\sigma_k^2}, \qquad \widetilde{d} = \sum_{k=1}^{K} g_k \widehat{d_k}, \tag{1}$$

where μ_k, σ_k^2, and d_k are the cluster-specific parameters. This idea is very similar to that in Molenberghs et al. (2011), who splits a sample in subsamples, that are analyzed seperately and than combined in an overall estimator.

2.2 Pseudo-likelihood for split samples

A pseudo-likelihood function is one that replaces a given likelihood function due to computational convenience. The likelihood contribution of a cluster is now a product of contributions for the various sub-vectors. Molenberghs et al. (2011) partitioned a sample in dependent or independent subsamples and used pseudo-likelihood for the fit. Referring to the compound symmetry

model described above, a pseudo-likelihood, for estimating a single vector (μ, σ^2, d) from a dataset divided into K subgroups, each containing c_k replicates, can be written as:

$$p\ell(\boldsymbol{\theta}) = \sum_{k=1}^{K} \ell(\boldsymbol{\theta_k} | \boldsymbol{y}_1^{(k)}, \ldots, \boldsymbol{y}_{c_k}^{(k)}), \tag{2}$$

with $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2, d_k)$. All $\boldsymbol{\theta}_k$ are assumed to be formally different, $\boldsymbol{\theta}$ stacks all vectors $\boldsymbol{\theta}_k$ and the parameter of interest $\boldsymbol{\theta}^*$, is found from an appropriate combination of the $\boldsymbol{\theta}_k$.

2.3 Weighting schemes

Referring to the setting in Section 2.1, note that subjects in different subsamples are allowed to have the same distribution, but that subjects in the same sub-sample must have the same distributions. Consider pseudo-likelihood in the general case (2). Assume that $\boldsymbol{\theta}^*$ is a vector of length p, and that each $\boldsymbol{\theta}_k$ is a separate copy of $\boldsymbol{\theta}^*$. Then $\boldsymbol{\theta}$ is a vector of length $K \cdot p$ and A is a $(p \times K \cdot p)$ matrix. The generic combination rules become:

$$\widetilde{\boldsymbol{\theta}}^* = \sum_{k=1}^{K} A_k \widehat{\boldsymbol{\theta}}_k, \qquad \text{var}(\widetilde{\boldsymbol{\theta}}^*) = \sum_{k=1}^{K} A_k V_k A_k', \tag{3}$$

with $V_k = I_0(\widehat{\boldsymbol{\theta}}_k)^{-1}$. We use the symbol $\widetilde{\boldsymbol{\theta}}^*$ to emphasize that this is not necessarily the maximum likelihood estimator even though, in our formalism, $\widehat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator when restricting attention to subsample k.

Not every choice of the A_k leads to an unbiased estimator, but to ensure an unbiased expectation of $\widetilde{\boldsymbol{\theta}}^*$, we impose $\sum_{k=1}^{K} A_k = I_p$. Two obvious choices are constant, $A_k = (1/K)I_p$, and proportional weights, $A_k = (c_k/N)I_p$. Constant weights are an intuitive choice when partitioning in sub-samples of equal size, however the latter one is more obvious for sub-samples of varying size. This leads us, using Lagrange multipliers, to the optimal weights, $A_k^{opt} = \left(\sum_{m=1}^{K} V^{-1_m}\right)^{-1} V_k^{-1}$. These then lead to the maximum likelihood estimator. However, not in every case will there be a closed-form solution for V_k and if there are these may depend on unknown parameters. To solve this dilemma, consider first scalar weights by demanding A_k to be diagonal. Each component of $\boldsymbol{\theta}^*$, θ_r^* say, is determined as a linear combination, $\widetilde{\theta}_r^* = \sum_{k=1}^{K} a_{k,r} \widehat{\theta}_{k,r}$, with $A_k = \text{diag}(a_{k,1}, \cdots, a_{k,p})$. The resulting optimum will not necessarily be equal to the MLE, but the weights can be chosen for computational convenience. A second option is iterated optimal weights. The data need to be analyzed only once, to find $\hat{\boldsymbol{\theta}}_k$. From these, an initial estimator for $\boldsymbol{\theta}^*$ is computed using a simple weighting method, e.g., constant or proportional weights. Using $\boldsymbol{\theta}^{*(t)}$ and calculating $V_k^{(t+1)}$, $\boldsymbol{\theta}^{*(t+1)}$ can be

determined as $\boldsymbol{\theta}^{*(t+1)} = \left(\sum_{k=1}^{K}\left[V_k^{(t+1)}\right]^{-1}\right)^{-1}\sum_{k=1}^{K}\left[V_k^{(t+1)}\right]^{-1}\widehat{\boldsymbol{\theta}_k}$. This is repeated until convergence. From this we deduce the approximate optimal weights, a non-iterative approximation.

3 Partitioned-sample analysis for the compound symmetry model

The weights discussed in the previous section can be constructed for this specific case. Due to the independence of the mean and the variance components, the optimal and scalar weights do not make a difference for the mean parameter, but are different for the variance parameters. The weights depend on the parameters, but by plugging in the cluster-size specific mean and variance components, the expressions can be used for approximate weighting. But also the principles of iterated and approximate weights can be applied, as in Section 2.3.
The scalar weights are found to be:

$$a_k = \frac{\frac{c_k n_k}{\sigma^2+n_k d}}{\sum_{m=1}^{K}\frac{c_m n_m}{\sigma^2+n_m d}}, \tag{4}$$

$$b_k = \frac{c_k(n_k-1)}{\sum_{m=1}^{K} c_m(n_m-1)}, \tag{5}$$

$$g_k = \frac{\frac{c_k n_k}{\frac{\sigma^4}{n_k-1}+2\sigma^2 d+n_k d^2}}{\sum_{m=1}^{K}\frac{c_m n_m}{\frac{\sigma^4}{n_m-1}+2\sigma^2 d+n_m d^2}}, \tag{6}$$

with $\sum_{i=1}^{K} a_k = \sum_{i=1}^{K} b_k = \sum_{i=1}^{K} g_k = 1$. These weights again depend on the parameters and they can again be made part of an iterative scheme. Calculations show that the variance of the weighted estimator of the mean equals that of the maximum likelihood, so the weighted split-sample parameter is the maximum likelihood estimator. This is to be expected due to the the independence of the mean estimator from the variance components estimators for a given cluster size. Thus, the optimally weighted estimator and the scalar estimator coincide for the mean. This is not true for the variance components, however.
By approximating these weights for the case where cluster sizes are large, we derive that these weights are almost identical to the proportional weights, which makes them a sensible option for practice.
All this can also be applied when there is only one cluster per sub-sample, a so called cluster-by-cluster analysis. Then, $c_k \equiv 1$, $K \equiv N$, and n_k will no longer be unique. Since we make use of the fact that the cluster size is constant within a stratum, and not that the cluster sizes must be different

TABLE 1. ML and weighted split-sample estimates (standard errors): (a) ML: maximum likelihood; (b) REML: restricted maximum likelihood; (c) Prop.: proportional weights; (d) Equal: equal weights; (e) Appr.sc.: like proportional weights, except that for b_k is used; (f) Scalar: scalar weights, with the sub-sample specific weights plugged in for the parameters figuring in the weights; (g) Opt.: optimal weights, with the sub-sample specific weights plugged in for the parameters figuring in the weights.

Par.	ML	REML	Prop.	Equal	Appr.sc.	Scalar	Opt.
μ	0.89450	0.89450	0.87305	0.85896	0.87305	0.91831	0.91831
σ^2	0.01950	0.01950	0.02200	0.02657	0.01951	0.01930	0.00603
d	0.01449	0.01467	0.00895	0.00858	0.00895	0.00085	0.00220
s.e.(μ)	0.01228	0.01234	0.01035	0.01478	0.01035	0.00761	0.00761
s.e.(σ^2)	0.00087	0.00087	0.00160	0.00431	0.00101	0.00097	0.00040
s.e.(d)	0.00245	0.00248	0.00208	0.00355	0.00208	0.00048	0.00044

between different strata, this is no problem. Result can be combined using again the weighted estimators or a two-stage approach. For the latter one, unbiasedness is not necessarily obtained.

4 Case study: a developmental toxicity study

The chemical compound di(2-ethylexyl)phthalate (DEHP) is used as plasticizer for numerous devices. Due to a possible presence in human and animal tissue, caused by leaks in plastic containers, toxic effects need to be investigated. The study was conducted in timed-pregnant mice during the period of major organogenesis (Tyl et al., 1988). A total of 1082 live fetuses were dissected. Our focus is on the continuous weight outcome. Fetuses are clustered within mothers. The CS model is fitted to the fetal weight outcome to examine the performance of the weighted estimators in Table 1. All split-sample estimators perform well in comparison with the (restricted) maximum likelihood estimators, only the optimal weights give slightly deviating results, which is because uncertainty due to the dependence of the weights on parameters is currently ignored. Using the delta method, this can be rectified. Importantly, the weighted estimators are a magnitude faster than the likelihood-based ones.

5 Concluding remarks

The use of weighted estimators reduced computation time and en enhances computation stability. They are simple to use, especially the proportional weights, and have a high efficiency.

Acknowledgments: Geert Molenberghs, Mike Kenward, Marc Aerts, Geert Verbeke and Wim van der Elst gratefully acknowledge support from IAP research Network P7/06 of the Belgian Government (Belgian Science Policy) and Geert Molenberghs and Geert Verbeke from ExaScience Project.

References

Hermans, L., Molenberghs, G., Aerts, M., Kenward, M.G., and Verbeke, G. (2014). A note on incomplete sufficient statistics. *Submitted for publication.*

Hermans, L., Molenberghs, G., Kenward, M.G, Van der Elst, W., Nassiri, V., Aerts, M., and Verbeke G. (2015). Clusters with random size: maximum likelihood versus weighted estimation. *Working paper.*

Molenberghs, G., Kenward, M.G., Aerts, M., Verbeke, G., Tsiatis, A.A., Davidian, M., Rizopoulos, D. (2014). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research*, **23**, 11 – 41.

Molenberghs, G., Verbeke, G., and Iddi S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics and Probability Letters*, **81**, 892 – 901.

Tyl, R.W., Price, C.J., Marr, M.C., and Kimmel, C.A. (1988). Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice. *Fundamental and Applied Toxicology*, **10**, 395 – 412.