## Simulating Clustered Data

Another difficulty encountered in several of the simulation studies conducted as part of this thesis was how to simulate clustered data. Such data can be generated using either marginal or conditional simulation models, which were introduced in Section . In general, the choice between these two approaches will depend on whether the simulated data will be analysed using marginal or conditional models, since parameters often differ between these models. This is not an issue for the log binomial model, however, since the marginal and conditional model parameters coincide, apart from the intercept [1]. Since the intercept was not of interest in this thesis, the type of simulation model (marginal or conditional) could be chosen independently of the intended analysis method.

Initially, I investigated using marginal simulation models to generate clustered binary outcomes. As discussed in Section , a range of methods are available for generating such outcomes under a marginal model. Each method was assessed for suitability based on the following requirements, determined by the simulation scenarios of interest:

The probability of success can vary within clusters;

An exchangeable correlation structure can be accommodated;

The method is computationally feasible for both small and large cluster sizes; and

Pairwise correlations can take on a wide range of positive values.

Based on these criteria, the method of Qaqish [2] was chosen for further investigation. This method involves generating the outcome for the first unit in the cluster () from a Bernoulli distribution with probability of success . Outcomes for subsequent units from the same cluster (;) are then generated from a Bernoulli distribution with conditional probability of success , given by the marginal probability of success () plus a weighted sum of the differences between the previously observed outcomes and the corresponding marginal success probabilities, i.e.

 ,

where  and  are the observed outcome and marginal success probability respectively for subject , and the weights  depend on the pairwise correlations . This method meets the requirements described above; it allows the probability of success to vary within clusters, it can accommodate any correlation pattern, including exchangeable, it can generate datasets quickly and easily using available SAS code [3], even when the cluster size is large, and it compares favourably with other methods in terms of the allowable range of pairwise correlations [4-6].

In practice, I found that this method worked well in simple scenarios. For instance, this method was successfully used to simulate clustered binary outcomes where the probability of success for each subject varied within clusters and was either 0.4 or 0.6, depending on a single binary predictor of interest (see Chapter 7). Use of this method in more complex scenarios was problematic, however. The conditional probability of success, defined by , often exceeded one which caused the method to fail. Since the observations from a cluster are generated sequentially, with the conditional probability of success dependent on previous outcomes, this suggests that the problem may be overcome by reordering the observations in the cluster. However, given that there are  possible permutations for a cluster of size , use of this marginal approach quickly becomes impractical when simulating multiple datasets [2]. As a result, I used conditional simulation models, rather than marginal simulation models, to generate clustered binary outcomes for the majority of simulation settings considered in this thesis.

## Avoiding Invalid Success Probabilities

A difficulty with simulating either independent or clustered data under the log binomial regression model is that the probability of success is only constrained to be greater than zero, whereas probabilities must lie between zero and one. As a result, simulating data under the log binomial model can lead to invalid success probabilities. This is generally only a problem when the probability of success depends on unbounded continuous covariates. For categorical covariates and/or bounded continuous covariates, simulation parameters can be chosen such that invalid probabilities cannot occur. For example, if the probability of success depends on treatment and a binary covariate, with a baseline success probability of 0.1 and a treatment relative risk of 2, the relative risk for the baseline covariate must be no more than 5 to avoid success probabilities exceeding 1 for intervention subjects with the baseline covariate.

To avoid success probabilities exceeding 1 when simulating independent data, a number of choices are available. First, simulation models can be limited to contain categorical covariates only. This approach restricts the usefulness of the simulation study, since continuous covariates are often controlled for in RCTs in practice. Second, continuous covariates can be categorised for analysis. This may lead to a loss of information and power [7]. Further, categorisation limits the applicability of the results of the simulation study, since continuous covariates may not be categorised in practice. Third, the continuous covariate can be generated from a truncated distribution, thus allowing the simulation parameters to be chosen to avoid success probabilities exceeding 1. For example, if the probability of success depends on treatment and a truncated continuous covariate, with a baseline success probability of 0.1, a treatment relative risk of 2, and a maximum covariate value of 1, the relative risk for the continuous covariate must be no more than 5 to avoid success probabilities exceeding 1 for intervention subjects with the maximum covariate value. This may limit the generalisability of the simulation results, depending on where the distribution is truncated. A normally distributed covariate truncated at the mean is unlikely to be representative of many baseline covariates considered in practice. In contrast, truncation at 2 standard deviations above the mean would have little impact on the range of covariate values generated and may therefore be a reasonable choice. A fourth option is for the value of the continuous covariate to be resampled until the success probability is valid. This effectively truncates the covariate distribution but the point of truncation will differ according to subgroups defined by any categorical covariates that affect the probability of success. This approach limits the amount of resampling required but will have a differential impact between treatment groups whenever treatment has an effect. Finally, success probabilities can be truncated, such that if the probability exceeds 1 based on the covariate values, it can be set to 1. The impact of this approach would likely depend on the proportion of probabilities that require truncation. A high proportion of truncated probabilities could result in substantial bias, while a low proportion would be expected to have little influence on the results.

None of the approaches described above provide a perfect solution to the problem of avoiding invalid success probabilities. The fourth approach involving resampling of continuous covariates has been used in previous simulation studies comparing methods for estimating relative risks [8-9] and was thus chosen for this thesis to allow comparison of results with those reported previously. For the simulation study conducted in Chapter 4, the probability of generating a covariate value above the point of truncation was very small for all scenarios considered, with a maximum value of only 0.03 for simulation scenarios involving a single continuous covariate, and 0.05 for simulation scenarios involving both a binary and a continuous covariate. This suggests that the need to resample was rare and the chosen approach for avoiding invalid success probabilities should therefore have little impact on the results.

A similar issue arises when simulating clustered outcomes under the log binomial regression model with a random cluster effect (assumed to be normally distributed). In this case, resampling can again be used to avoid the possibility of generating invalid success probabilities, which effectively truncates the random effects distribution. This approach was used for the simulation studies described in Chapters 5 to 7. Truncation points were more than two standard deviations above the mean for the vast majority of simulation scenarios investigated. This indicates that the need to resample the random effect was generally very small and the impact of resampling on the results should therefore be limited.

References

1. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research* 2004; 13(4): 309-323.

2. Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* 2003; 90(2): 455-463.

3. Qaqish B. *clfsim.sas*. 2003 [cited 17/03/09]; Available from: <http://www.bios.unc.edu/distrib/gee/clf/clfsim.sas>.

4. Chaganty NR, Joe H. Range of correlation matrices for dependent Bernoulli random variables. *Biometrika* 2006; 93(1): 197-206.

5. Farrell PJ, Sutradhar BC. A non-linear conditional probability model for generating correlated binary data. *Statistics & Probability Letters* 2006; 76(4): 353-361.

6. Farrell PJ, Rogers-Stewart K. Methods for generating longitudinally correlated binary data. *International Statistical Review* 2008; 76(1): 28-38.

7. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 2006; 25(1): 127-141. DOI: 10.1002/Sim.2331

8. Carter RE, Lipsitz SR, Tilley BC. Quasi-likelihood estimation for relative risk regression models. *Biostatistics* 2005; 6(1): 39-44.

9. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology* 2007; 60(9): 874-882.