# Sensitivity of generalised least squares to correlation structure misspecifications

Daniel Mahar
Supervised by Dr. Luke Prendergast
La Trobe University

## 1. Background

In a wide range of emerging fields, particularly within medicine, education and psychology, an important statistical design is *repeated measures*. Repeated measures is a design that uses the same individuals throughout the course of the experiment. There are two primary methods for conducting a repeated measures study. In the first situation, the individuals perform a series of tests or measurements under different conditions. Here, the results are considered to be independent of time and if multiple tests are conducted, the individual results on one test do not influence the results on any other tests. The second major use for repeated measures is for a *longitudinal study*. A longitudinal study is where the same measurements are taken from the same individuals over a period of time, which can be hours, weeks, months or even years, depending on the study in question. The purpose of a longitudinal study is to measure change over time so as to, for example, measure the short and long term effectiveness of a new treatment for people with high blood pressure.

A key aspect of a repeated measures design is that since the same individuals are being used throughout, there will be multiple observations for each individual and as such, any given observation for each individual won't be independent of the other measurements for that same individual. Consequently, there will be a *correlation* structure between the observations for each individual. Correlation is a statistical measure of how strongly variables are associated with each other, either positively or negatively. For longitudinal studies, each time point acts as a variable and as such, the response for each individual at each time point is correlated with the response for each individual at the other time points.

Correlations themselves are derived from the variances and covariances of the error terms for each individual where for variables X, Y;

$$\text{Correlation}(X, Y) = \rho(X, Y) = \frac{\text{Covariance}(X,Y)}{\sqrt{\text{Variance}(X).\text{Variance}(Y)}} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2.\sigma_Y^2}}$$

Throughout this paper, covariance matrices will be used to represent the covariance structures and hence the correlation structures using the following assumptions for simplicity. Assume that there are three responses per individual. Let $\sum$ denote the covariance of the error term for each individual. Then the general covariance structure is a symmetric matrix such that

$$\sum = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$$

## 2. Methods of Analysis

There are a wide range of possible statistical methods that can be used to analyse a longitudinal study. This paper will focus on four such methods. Firstly, we will consider Multiple Linear Regression, a popular model for analysis but one which is completely unsuited to repeated measures design. Secondly we will consider a commonly used model for repeated measures design, the Repeated Measures ANOVA (RMANOVA) and note its weakness when considering a longitudinal study. Finally, we will consider two more recent methods, Generalised Least Squares (GLS) and Linear Mixed Effects (LME), which allow different correlation structures to be chosen.

Assume that for the $i^{th}$ individual within the longitudinal study that

$$\mathbf{Y_i} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \mathbf{X\beta} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

where $\mathbf{Y_i}$ is the response vector, $\mathbf{X\beta}$ is the matrix of variables and $\varepsilon$ denotes the variance of each response.

Multiple Linear Regression has two major assumptions. Firstly it assumes that observations are independent within and across individuals and secondly it assumes constant variance within and across individuals. These two assumptions cause the following covariance structure to arise.

$$\sum = \sigma^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ where } \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ is the correlation structure.}$$



*Postal Address:* Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au

The reason Multiple Linear Regression is not suited to repeated measures design is that as shown above, the assumptions made result in zero correlation between variables for each individual.

Repeated Measures ANOVA (RMANOVA) also has two major assumptions, namely equal variance and equal covariances. These assumptions lead to the following covariance structure.

$$\sum = \sigma^2 \begin{pmatrix} 1 & p & p \\ p & 1 & p \\ p & p & 1 \end{pmatrix} \text{ where } |p| \leq 1.$$

Unlike the Multiple Linear Regression model, the RMANOVA model does include a correlation coefficient. RMANOVA is widely used for repeated measures design, in particular for time independent studies where it is reasonable to assume that correlation in constant. However, in longitudinal studies, the constant correlation may not be as reasonable an assumption as people change over time, irrespective of any treatment they might be undergoing. Consequently, a correlation structure which has smaller magnitudes of correlation for time points which are further apart may be more appropriate.

Generalised Least Squares (GLS) has the same model structure as shown earlier for the Multiple Linear Regression, however, it has some key differences in assumptions. Firstly, it allows for errors to be correlated and for variances to be unequal and lastly it allows different correlation structures to be chosen. These changes in assumptions fix the issues associated with using Multiple Linear Regression for repeated measures designs.

Linear Mixed Effects Model (LME) is a model containing both fixed and random effects where, for the $i^{th}$ individual;

$$\mathbf{Y_i} = \mathbf{X\beta} + \mathbf{Z_i\mu_i} + \boldsymbol{\varepsilon_i}$$

where $\mathbf{Z_i\mu_i}$ is the matrix of random effects for each individual. As with the GLS, different correlation structures can be chosen for the LME.

## 3. Correlation Structures

There are a wide range of different correlation structures that can be chosen for the GLS and MLE, however, in this paper we will be focusing on three of the most widely used.

(i) Compound Symmetry: $\sigma^2 \begin{pmatrix} 1 & p & p \\ p & 1 & p \\ p & p & 1 \end{pmatrix}$



*Postal Address:* Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au

The GLS, equipped with Compound Symmetry is an equivalent model to the Repeated Measures ANOVA shown earlier. The LME with a random intercept and Compound Symmetry is equivalent to the GLS with Compound Symmetry up until degrees of freedom.

(ii) AR(1): $\sigma^2 \begin{pmatrix} 1 & p & p^2 \\ p & 1 & p \\ p^2 & p & 1 \end{pmatrix}$

Since $|p| \leq 1$, the AR(1) correlation structure allows for a lower correlation between more distant events when time is a variable, such as for longitudinal studies. This is a more intuitive structure than then compound symmetry structure given what we know about how people change over time.

(iii) Unstructured: $\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$

An unstructured correlation structure allows for every variance and each unique covariance to be different, noting that this is still a symmetric matrix. As such, it has no limitations and is consequently very flexible. One drawback of the unstructured correlation structure is that it can lead to computational issues and in situations where there are a small number of observations, it is sometimes unable to obtain results at all due to the large number of parameters needing to be evaluated.

## 4. Simulation

In order to assess how sensitive each of the various methods are to different correlation structures, a simulation was coded in R to assess the *coverage probability* of the 95% confidence interval. Coverage probability denotes the proportion of occasions that the true value, in this case 0, lies within the confidence interval. This simulation consisted of four observations each for twenty individuals, where by the data produced arose from a multivariate normal distribution with the AR(1) structure used for the covariance structure of the error terms. In each instance, the simulation was run 1000 times in order to obtain the asymptotic approximation for the true coverage probability. Each set of 1000 runs was then repeated ten times and averaged to obtain the final coverage probability.



*Postal Address:* Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au

## 5. Results

| | 0.2 | 0.5 | 0.8 |
|---|---|---|---|
| Multiple Linear Regression | 0.881 (min 0.857) | 0.802 (0.782) | 0.716 (0.674) |
| GLS-Compound Symmetry | 0.937 (0.932) | 0.930 (0.919) | 0.934 (0.926) |
| GLS-AR(1) | 0.932 (0.919) | 0.931 (0.916) | 0.934 (0.923) |
| GLS- Unstructured | 0.930 (0.918) | 0.928 (0.917) | 0.932 (0.921) |
| LME-Compound Symmetry* | 0.955 (0.948) | 0.945 (0.937) | 0.950 (0.939) |
| LME-AR(1)* | 0.958 (0.949) | 0.957 (0.941) | 0.954 (0.940) |
| LME- Unstructured * | 0.938 (0.925) | 0.936 (0.922) | 0.935 (0.924) |

Table 1. Coverage probability for treatment effect using data with AR(1) covariance structure with p = {0.2, 0.5, 0.8}. LME models used random intercept, grouped by individual.

* Note, the LME models had significant issues obtaining confidence intervals (CI). The Compound Symmetry model failed to obtain a CI on 0.9% of occasions, the AR(1) model on 24.8% of occasions and the Unstructured on 50.1% of occasions. The most likely explanation for this is a convergence issue within the LME function in R, with the function unable to always converge within the allowed number of iterations.

From Table 1 above, the Multiple Linear Regression model performed poorly as expected, especially as the correlation coefficient moved further away from zero. All three of the GLS models performed quite well with a coverage probability around 93%. Interestingly, there was little difference in choosing the wrong correlation structure which suggests that the GLS model is not sensitive to correlation structure misspecifications. The LME models with Compound Symmetry and AR(1) performed excellently, with a coverage probability hovering around the desired 95% range. The unstructured LME model, however, performed slightly worse than the other two models which was something of a surprising result.

As noted earlier on, the GLS with compound symmetry and the LME with the random intercept and compound symmetry are equivalent models up to degrees of freedom (df). Both models produced the same β coefficients and the same standard error values but due to a difference in how the two models determine the degrees of freedom, there is a difference in the coverage probability as shown in Table 1. More



*Postal Address:* Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au

specifically, the degrees of freedom were calculated as follows, noting that there are twenty individuals who each have four observations for a total number of eighty observations.

GLS (df): 80 – Number of variables (time points) = 80-4 = 76.

LME(df): 80 – Number of random intercepts (1 per individual, taken at time=0) – number of remaining variables (time points not including time = 0) = 80-20-3 = 57.

As 57 and 76 are both not a low magnitude and the difference of 19(df) is not that significant, the resulting confidence intervals are of a similar width and hence, the coverage probabilities are quite similar as shown in Table 1. For a more complicated model ,however, for example an LME that had both a random intercept and a random slope coefficient, the degrees of freedom would be 18 (20 individuals -2 variables), which would have a greater effect on the difference in coverage probability as, all thing being equal, a smaller number of degrees of freedom creates a wider confidence interval and thus a higher coverage probability.

## 6. Conclusion

The aim of this project was to examine various correlation structures and develop R code for the various methods of analyzing longitudinal data in order to determine how sensitive each method was to misspecifications in the correlation structure. The major focus of the project was writing code in R to run the simulation required to complete the examination of the various methods.

The simulation showed that Multiple Linear Regression model, which assumes no correlation within and between groups is a poor method of analysing longitudinal data. The Generalised Least Squares (GLS) showed little change to correlation structure misspecifications with the compound symmetry structure (equivalent to running the RMANOVA) performing as well as the AR(1) model which was the correct model. Likewise for the Linear Mixed Effects (LME) model, using a random intercept with compound symmetry yielded similar results to the correct AR(1) model.

Whilst the GLS and MLE aren't particularly sensitive to correlation misspecifications on this relatively simple model, there is still the primary issue of justification. As statisticians working in fields where the majority of collaborators are not statisticians, it is imperative that we can justify why we're choosing a particular model or a particular correlation structure, even if not doing so would lead to the same end conclusion by non statisticians.

Complications with the LME model proved to be a significant hurdle in the project and with more time, a greater understanding of why confidence intervals were not always able to be calculated may be forthcoming, and possible changes to the LME model in R in order to solve the issue might be developed.



*Postal Address:* Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au

## 7. Acknowledgements

Thank you to AMSI for providing this opportunity to research an emerging area of statistics, CSIRO and the University of New South Wales for organizing and hosting the Big Day In and to Dr. Luke Prendergast for all his time, support and help throughout the course of this project.

## References

Sheather, SJ 2009, *A modern approach to regression with R*. Springer, New York.

West, BT, Welch, KB & Galecki, AT, 2007, *Linear Mixed Models: A Practical Guide Using Statistical Software*, Chapman & Hall/CRC, New York



*Postal Address:* Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au