ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY
DEPARTMENT OF STATISTICS
POSTGRADUATE PROGRAM

COVARIANCE ESTIMATORS FOR GENERALIZED ESTIMATING EQUATIONS IN LONGITUDINAL ANALYSIS WITH SMALL SAMPLES

By
MARI K. BARAZIAN

supervised by
Pr. V. Vasdekis

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
June, 2017

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ

ΕΚΤΙΜΗΤΕΣ ΣΥΝΔΙΑΚΥΜΑΝΣΗΣ ΓΙΑ ΤΙΣ ΓΕΝΙΚΕΥΜΕΝΕΣ ΕΞΙΣΩΣΕΙΣ ΕΚΤΙΜΗΣΗΣ ΣΕ ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΕΣ ΜΕΤΡΗΣΕΙΣ ΜΕ ΜΙΚΡΑ ΔΕΙΓΜΑΤΑ

ΜΑΡΙ Κ. ΜΠΑΡΑΖΙΑΝ

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Ιούνιος, 2017

To my parents, my brother and Paul

Acknowledgements

First and foremost, I would like to express my profound gratitude to my supervisor Pr. Vasilios Vasdekis who never hesitated to share with me his knowledge. The guidance and the assistance he offered me were really helpful throughout my study. All the experience I gained from this Master program has prepared me well for my future professional career.

Moreover, I would like to acknowledge my parents Kyriako and Sofia and my brother Haik for their endless love, support and understanding. Last but not least, I would like to thank Paul for being always there for me.

Vita

I was born in Athens, Greece. I studied Mathematician at the National and Kapodistrian University of Athens. I proceeded with my studies at the department of Statistics of the Athens University of Economics and Business as a postgraduate.

Up until now, my studies gave me the opportunity to enrich my knowledge and fulfill my dream; understand the numbers and get involved with them, considering the fact that they are a significant part of our daily life.

Abstract

Mari Barazian

COVARIANCE ESTIMATORS FOR GENERALIZED ESTIMATING EQUATIONS IN LONGITUDINAL ANALYSIS WITH SMALL SAMPLES

June, 2017

The Generalized Estimating Equations (GEE) statistical method is a simple and efficient approach to estimate the regression coefficient of a marginal model for correlated responses when the associational structure is regarded as a "nuisance". Its most common use is to fit marginal models for longitudinal data in several fields such as biomedical studies and social sciences. The most attractive feature of the GEE methodology is that consistent estimates for marginal regression coefficients are obtained even if the correlation structure is misspecified. However, the technique requires that the sample size is large. The variance-covariance matrix of the regression parameter coefficients is often estimated by the so-called "sandwich" variance estimator, which is robust and performs well when the size of the sample is large. However, when the sample size is small, the "sandwich" estimator does not have a good performance. Specifically, in that case, bias and inefficiency appear. The main goal is to find ways in order to decrease the bias and improve the efficiency. For this reason, some recently developped modified variance estimators have been proposed. The current GEE methodology focuses on the modeling of the working correlation matrix assuming a known variance function. However, Wang, Y.-G., Lin, X. and Zhu, M. (2005) showed that the correct choice of the correlation structure may not necessarily improve the estimation efficiency for the regression parameters if the variance function is misspecified.

The purpose of this thesis is to provide a review on recent developments of modified variance estimators. One of the most attractive parts is the comparison of their small-sample performance and the presentation of the most important results which were obtained through simulations and one real data example. Simulation shows that the modified estimators do well in reducing bias and increasing efficiency compared to the GEE estimates. Hypothesis testing that is used is based on Wald tests and t-tests on different variance estimators. Finally, the "geesmv" R package which incorporates all of those variance estimators is used for programming purposes.

Περίληψη

Μαρί Μπαραζιάν

ΕΚΤΙΜΗΤΕΣ ΣΥΝΔΙΑΚΥΜΑΝΣΗΣ ΓΙΑ ΤΙΣ ΓΕΝΙΚΕΥΜΕΝΕΣ ΕΞΙΣΩΣΕΙΣ ΕΚΤΙΜΗΣΗΣ ΣΕ ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΕΣ ΜΕΤΡΗΣΕΙΣ ΜΕ ΜΙΚΡΑ ΔΕΙΓΜΑΤΑ

Ιούνιος, 2017

Η Στατιστική μέθοδος των Γενικευμένων Εξισώσεων Εκτίμησης είναι μία απλή και αποτελεσματική προσέγγιση ώστε να εκτιμήσουμε τον συντελεστή παλινδρόμησης ενός περιθώριου μοντέλου όταν υπάρχουν συσχετίσεις στις μεταβλητές απόκρισης και η δομή της συσχέτισης θεωρείται ως παράμετρος ενόχλησης. Η πιο συνηθισμένη χρήση τους έγκειται στην προσαρμογή περιθώριων μοντέλων σε βιοϊατρικές μελέτες και κοινωνικές επιστήμες όταν τα δεδομένα μας αφορούν επαναλαμβανόμενες μετρήσεις στο πέρασμα του χρόνου. Το πιο ελκυστικό στοιχείο των Γενικευμένων Εξισώσεων Εκτίμησης είναι το γεγονός ότι ακόμα και στην περίπτωση όπου η δομή της συσχέτισης δεν είναι επακριβώς προσδιορισμένη, μπορούμε να λάβουμε συνεπείς εκτιμητές των συντελεστών παλινδρόμησης στα περιθώρια μοντέλα. Ωστόσο, κάτι τέτοιο προϋποθέτει την ύπαρξη μεγάλου μεγέθους δείγματος. Ο πίνακας διακύμανσης-συνδιακύμανσης των συντελεστών παλινδρόμησης εκτιμάται συνήθως από τον επονομαζόμενο 'σάντουϊτς' εκτιμητή διακύμανσης, ο οποίος όμως δεν αποφέρει καλή συμπεριφορά στην περίπτωση του μικρού μεγέθους δείγματος. Στην περίπτωση αυτή, γίνεται αισθητή η εμφάνιση μεροληψίας και αναποτελεσματικότητας. Για τον λόγο αυτό έχουν προταθεί αρκετοί διορθωτικοί εκτιμητές διακύμανσης προκειμένου να επιτευχθεί μείωση της μεροληψίας και βελτίωση της αποτελεσματικότητας. Η παρούσα μεθοδολογία των Γενικευμένων Εξισώσεων Εκτίμησης εστιάζει στη μοντελοποίηση του πίνακα συσχέτισης υπό την υπόθεση ύπαρξης γνωστής συνάρτησης διακύμασης. Ωστόσο, οι Wang, Lin, Zhu (2005) έδειξαν ότι η σωστή επιλογή της δομής συσχέτισης πιθανόν να μη βελτιώσει την αποτελεσματικότητα της εκτίμησης για τις παραμέτρους παλινδρόμησης εάν η συνάρτηση διακύμανσης δεν είναι σωστά προσδιορισμένη.

Ο σκοπός αυτής της Διπλωματικής εργασίας είναι να παρουσιάσει την πιο πρόσφατη έρευνα και τις τελευταίες πρακτικές ανάπτυξης σχετικά με τους διορθωτικούς εκτιμητές διακύμανσης και να συγκρίνει τη συμπεριφορά τους πάνω σε μικρά δείγματα τόσο σε θεωρητικό υπόβαθρο όσο και σε πρακτικό μέσω της προσομοίωσης αλλά και ενός παραδείγματος με πραγματικά δεδομένα. Η προσομοίωση υποδεικνύει ότι οι διορθωτικοί εκτιμητές διακύμανσης συμπεριφέρονται αρκετά καλά στη μείωση της μεροληψίας και στην αύξηση της αποτελεσματικότητας συγκρινόμενοι με τους εκτιμητές των Γενικευμένων Εξισώσεων. Οι έλεγχοι υποθέσεων που διεξάγονται στηρίζονται στους Wald tests και t-tests για κάθε εκτιμητή διακύμανσης και στο κατάλληλο μέγεθος του δείγματος ώστε να περιοριστεί το σφάλμα τύπου I. Τέλος, το πακέτο που χρησιμοποιείται για την ανάπτυξη του κατάλληλου κώδικα είναι το "geesmv" της γλώσσας R, το οποίο περιέχει ενσωματωμένους όλους τους διορθωτικούς εκτιμητές διακύμανσης με τους οποίους ασχολούμαστε.

Contents

List of Figures

List of Tables

https://doi.org/10.26219/heal.aueb.6669

1 Introduction

When measurements are taken from the same subject more than one time, the responses are no longer independent. Longitudinal studies are characterized by repeated measurements of the same individuals allowing the study of change over time. In these studies it is reasonable to assume that the subjects are independent, but the repeated measurements taken on each subject may not be uncorrelated. The purpose of longitudinal data analysis is to model the relationship of the repeated measurements of each subject to the associated covariates. Moreover, the primary goal of a longitudinal study is to specify the change in response through time and the factors that influence this change. With repeated measures on individuals one can capture within-individual change, as G. Fitzmaurice et al. (2004) mentioned. An exceptional feature of longitudinal data is that they are clustered. The clusters are composed of the repeated measurements obtained from a single individual at different occasions. Observations within a cluster will typically exhibit positive correlation and this correlation must be accounted for in the analysis.

There are three types of models for longitudinal data analysis: (1) transition or fully conditional models (Korn and Whittemore, 1979, Rosner, 1984 and Zeger and Qaqish, 1988 etc.), (2) random effects models (Rao, 1965, Laird and Ware, 1982 and Stiratelli, Laird and Ware, 1984 etc.) and (3) marginal models (Liang and Zeger, 1986, Zeger and Liang, 1986 and Prentice and Zhao, 1991 etc.). Transition models are used to specify the conditional distribution of each response given the past responses. Random effects models describe the natural heterogeneity among subjects. Marginal models are used to characterize the marginal expected value of a subject's response as a function of the subject's covariates. Diggle, Liang and Zeger (1994) discussed these models in detail.

There are several approaches in order to analyze repeated measurements, with the mixed-effects models and the Generalized Estimating Equations (GEE) to be the most popularly applied. The distinct condition of mixed-effects models is that some subset of the regression parameters typically vary from one individual to another accounting for natural forms of heterogeneity in the population. This means that the individuals in the population have their own subject-specific mean response feature over

time. Moreover, the mean response is modeled as a combination of population characteristics, that are assumed to be shared by all individuals and are called *fixed effects* and subject-specific effects that are unique to a particular individual and are called *random effects*. The term *mixed* is used to demonstrate that the model contains both fixed and random effects.

The GEE method is developed from the theory of Generalized Linear Models (GLM) by Nelder and Wedderburn (1972). GEE is used to estimate the regression parameters in marginal models of longitudinal data in which the link function and the variance function take the forms of those in GLM.

Generalized Linear Models (GLM) are widely used to estimate regression coefficients of a linear model. This class of models can be applied to both continuous and discrete data and introduce a likelihood-based method in which the distribution of the response variable is a member of the exponential family. The main feature that characterizes this methodology is the existance of one random component and one systematic component called linear predictor and the link between those two. The monotone link function which relates the expected responses and the linear predictor may not be linear.

Another method that is closely related to the GLM is the quasi-likelihood method (QL) proposed by Wedderburn (1974) and explored by McCullagh and Nelder (1989). The QL method requires only the first two moments, mean and variance, for estimating the regression parameters when the distribution may not be from an exponential family. The main thing that differentiates the QL method of the GLM method is that the former does not have a likelihood function as it does not assume a full distributional specification, so the inference about the parameters relies solely on limit theory results. However, both of these methods assume independence of the observations.

Despite the fact that these methods are really useful to a statistician, there are some important limitations to their use. The lack of independence among the repeated measures of the same individual or the existence of clustered data make the application of the GLM or the QL method inappropriate. Clustered data are data whose observations come into clusters showing that there are subjects with common

characteristics. Moreover, clustered data arise when each individual is measured repeatedly through time so we expect the responses within a cluster to be correlated.

However, one could use maximum likelihood methods that typically take into account the dependencies within a cluster but these methods have two important disadvantages; they are computationally difficult and they are sensitive to the misspecification of the correlation structure.

A straightforward application of the generalized linear models to longitudinal data is not correct, as mentioned before, due to the lack of independence among the repeated measurements of the same individual. The approach for extending generalized linear models to longitudinal data leads to a class of regression models known as *marginal models*. The name of these models indicates that the model for the mean response depends only on the covariates of interest and not on any random effects or previous responses. An asset of marginal models is that they require only a regression model for the mean response and not any distributional assumption for the observations. The only assumptions that marginal models rely on are those about the mean response. The avoidance of distributional assumptions leads to a method of estimation known as Generalized Estimating Equations (GEE).

The GEE technique is asymptotic. Thus, in the case of small sample sizes, GEE may result in biased estimates. Notice that the GEE function is an extension of the quasi-likelihood which is the true likelihood when the distribution is from an exponential family. This motivates us to use the bias-correction technique in maximum likelihood estimation to reduce the bias. Under general conditions, maximum likelihood (ML) estimators are consistent, but they are not unbiased generally.

The GEE method that was proposed by Liang and Zeger (1986) and Zeger and Liang (1986) is a synthesis of the GLM and the QL method. On one hand, it requires correct specification of the model for the response mean and on the other hand it allows us to adopt a working assumption for the correlation structure. As in the GLM procedure, the mean is also described through a link function that is the connection between the response variable and the linear predictor. The GEE methodology has three main advantages that attract a statistician's attention; (i) When the inference is intended to be population-based, GEE treats the variance-covariance matrix of the

responses as a "nuisance" parameter. (ii) The regression parameter estimates are consistent and asymptotically normal even if the correlation structure of the responses is misspecified and (iii) GEE is computationally simple enough, as it relaxes the distribution assumption; it is necessary to specify correctly the marginal mean and variance as well as the link function between the mean and the covariates of interest.

As it is well known, the variance estimators of parameters of interest are really useful in hypothesis testing. In order to obtain valid inference, it is actually important to have accurate estimates. As it was mentioned before, the GEE methodology with the classic "sandwich" variance estimator does not have a good performance when the sample size is small. As a result, considerable bias appears (Gunsolley JC, Getchell C, Chinchilli VM., 1995) which in turn leads to inflated type I errors and smaller coverage rates of the resulting confidence intervals (Wang M, Long Q., 2011). This specific feature was the main factor for developing several modifications of variance-covariance estimators in order to improve the small-sample performance.

This Master thesis is organised as follows: After the section of Introduction we continue to the Section 2 in which we provide the GEE theory and methodology as well as we introduce the notations of all nine variance estimators of GEE with their theoretical and practical comparisons. In Section 3, the simulation procedure follows in order to compare the performance of different variance estimators and analyze their performance in controlling the bias and the inflated type I error. The R package "geesmv" is proved really helpful for our simulation study. Moreover, our simulation results are getting more understandable through a real data example in Section 4. Last but not least, in Section 5, we give the conclusion of this thesis with a brief discussion and some future research.

2 Generalized Estimating Equations Theory and Modified Variance Estimators With Small Samples

2.1 Background Study

2.1.1 Basic Aspects of Longitudinal Analysis

The fact that measurements of the same individuals are taken more than once through time defines longitudinal studies. The result is the direct study of change, while the goal is to characterize the change in response over time and the factors which influence that change. With repeated measurements on different individuals one can capture within-individual change, providing not only comparisons among different individuals but also information about how individuals change during the corresponding period. Another feature that differentiates longitudinal data from other type data is that they are *clustered*. The clusters are composed of the repeated measurements taken from the same individuals at different occasions. Additionally, observations of the same cluster will possibly exhibit positive correlation.

It gets obvious from the above that longitudinal data have to deal with two types of dependence: (1) homogeneity of the responses of the same individual and (2) heterogeneity across different individuals. In a repeated measurements design the response variable can be in the form of count data, such as the number of children laid; binary, such as the gender of the people (male or female); categorical, such as the type of damage to a machine, which can be aggregated into counts; lastly, it can be in the form of continuous data, such as the growth of a child's height. These responses may have come from a study where the subjects have undergone some treatment. Randomisation is required to allocate subjects to treatment groups so that bias is avoided. Lindsey (1993, p.9) notes that randomisation allows for statements of causality, since which treatment a subject receives is not influenced by the response that the subject gives. It also minimises the effects of inter-response variability by distributing it randomly over treatments, thereby ensuring homogeneity of variability. In order to attribute causality, the relationship between the cause and the effect needs to be strong, and the relationship should be consistent in different populations and under different circumstances. In addition, the cause needs to lead to a single effect

(specificity) and the cause must precede the effect in time (temporality).

2.1.2 Models for Repeated Measurements

There is a great variety of different types of models which can be used in order to analyze repeated measurements. Linear mixed effects models are the most commonly used.

2.1.2.1 Random Effects Models

In random effects modeling one or more variables are declared as random factors. If a model also contains fixed factors, then the model is referred to as a mixed model. Random factors have a distribution assumed for the different levels while the values for the levels of a fixed factor are fixed, known values which are chosen at the beginning of the experiment and the effects of each level on the response are estimated as model coefficients. When a factor is declared to be a random factor, then inferences can be made on the population from which the levels of the random factor have been chosen. Correlation can also be incorporated into the model, since observations that share the same level of the random effect are modeled as correlated. A great variety of bibliography related to random effects models is available (Crowder & Hand, 1990; Davis, 2002; Fitzmaurice et al., 2004)

In a repeated measures ANOVA, a random effect for the individuals of the study can be included in the model. As a result, positive correlation is induced between repeated measurements through the covariance matrix of the random effects while concerning the mean structure, random effects can be thought of as randomly varying intercepts which account for all unmeasured factors (Fitzmaurice et al., 2004).

The repeated measures ANOVA model can be written as:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + b_i + e_{ij}$$

where b_i is a random individual-specific effect and e_{ij} is a within-individual measurement of error (Crowder & Hand, 1990; Fitzmaurice et al., 2004).

There are two standard assumptions when using ANOVA for repeated measures; (1) The observations on different subjects at each of the repeated measurement times are independent and (2) these observations are distributed as multivariate normal. Therefore, the b_i's are assumed to be normally distributed with mean zero and $var(b_i) = \sigma_b^2$ as well as the e_{ij} are assumed to be normally distributed with mean zero and $var(e_{ij}) = \sigma_e^2$. Thus, repeated measures ANOVA has two different sources of variability; on one hand, the subject variability (σ_b^2) and on the other hand the within subject variability (σ_e^2). In addition, the b_i's of the different individuals are uncorrelated and the errors e_{ij}'s are uncorrelated for different time points and for different individuals. Lastly, it is assumed that all the correlations in the outcome variable between repeated measurements are equal and variances of the outcome variable are the same at each of the repeated measurements (which is known as sphericity). An example of a covariance matrix that satisfies the sphericity condition is the compound symmetric (CS) covariance matrix (Hand & Crowder, 1996, p. 41) :

$$\begin{pmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \dots & \sigma_b^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 + \sigma_e^2 \end{pmatrix}$$

The mean response can be written as follows since the means of b_i's and e_{ij}'s are equal to zero:

$$E(y_{ij}) = \mu_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}$$

Fitzmaurice et al. (2004, p. 14) note that regression models have a wide range of uses. Regression models include linear regression, linear logistic regression and Poisson or log-linear regression models. Linearity means that all of these models for the mean or a transformation of the mean are linear in the regression parameters. The regression parameters in the model express how the covariates are related to the mean of the response variable. The covariates can be quantitative or categorical (such as gender or treatment group). Models which only include categorical covariates are actually

ANOVA models.

2.1.2.2 Linear Mixed Effects Models

Linear Mixed Effects Models represent one of the most widely used methods of including the covariance matrix in the statistical analysis. Mixed effects models are those where the mean is modeled through both random and fixed effects.

Fixed effects are those factors in a model for which the designer of the experiment had deliberately chosen certain levels and which are the only levels of interest, rather than randomly sampling levels from an infinite population of possible levels (Vittinghoff et al., 2005). When a researcher chooses individuals for a study in such a way that both males and females are included, then gender can be considered as a fixed effect.

When the researcher does not explicitly choose the levels of a factor, but rather the levels are a sample of the possible levels available, then this is known as a random effect (Fitzmaurice et al., 2004). In the Potthoff and Roy dataset the children included in the study are an example of a random effect, as they were randomly selected from a larger population of children. Including individual specific random effects into a model can be used to account for correlation among repeated measurements (Fitzmaurice et al., 2004; Vittinghoff et al., 2005).

Linear mixed effects models are a special case of mixed effects models in which both the fixed and random effects occur linearly in the model function. The most common formulation of the model is that of Laird and Ware (1982):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \text{ for } i = 1, 2, ..., K$$

and

$$\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$$

where $\mathbf{y}_i$ $(n_i \times 1)$ are independent and normally distributed, $\boldsymbol{\beta}$ is the p-dimensional vector of fixed effects, $\mathbf{b}_i$ is the q-dimensional vector of random effects, $\mathbf{X}_i$ $(n_i \times p)$ and $\mathbf{Z}_i$ $(n_i \times q)$ are known fixed effects and random effects regressor matrices respectively,

and $\boldsymbol{\epsilon}_i$ is the n_i-dimensional within-individual error vector. It is assumed that $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ are independent for different individuals and that they are independent of each other for the same individual. A structure needs to be chosen for $\boldsymbol{\Sigma}$, the covariance matrix of $\mathbf{b}_i$. The consequences of these structural choices will be the main consideration of the following chapters.

However, during this specific research, two main thoughts appeared. The first one is the consequences which may appear when using an over-simplified model, namely the ordinary linear regression model which assumes independence of repeated measurements, to analyze repeated measures data. The second thought stands for wondering if an appropriate model is chosen, what the consequences are of using an incorrect parameterisation of the covariance structure for the estimates of the fixed effects and inferences about these estimates.

2.2 Hierarchical versus Marginal Modeling Approaches

The marginal modeling approach and the hierarchical modeling approach both assume correlation into the model. The former assumes a model which holds averaged over all the clusters (also referred to as population averaged). Thus, the coefficients can be interpreted as the average change in the response for a unit change in the predictor over the entire population. The second is the hierarchical or conditional modeling approach which assumes a model specific to each cluster (also referred to as subject specific). Coefficients can then be interpreted as the change in the response in each cluster in the population for a unit change in the predictor, and the marginal information can be obtained by averaging over all the clusters.

In the analysis that follows we highlight the main parts of hierarchical modeling as well as marginal modeling. According to Verbeke and Molenberghs (2000), hierarchical modeling implies a two-stage process; In the first stage, which can be named as the calculation stage, it is assumed that the following linear regression relationship holds:

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$$

where $\mathbf{Z}_i$ $(n_i \times q)$ is a matrix of known covariates, $\boldsymbol{\beta}_i$ $(q \times 1)$ is a vector of unknown subject-specific regression coefficients and $\boldsymbol{\epsilon}_i$ is the vector of residuals of length n_i. This regression equation models how the i^{th} subject's response evolves over time. All $\boldsymbol{\beta}_i$ estimates for the observed $\mathbf{y}_i$ for each subject are obtained separately.

In the second stage, which can be interpreted as the analysis stage, a multivariate regression model for the subject-specific regression coefficients, $\boldsymbol{\beta}_i$, is assumed to be:

$$\boldsymbol{\beta}_i = \mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i$$

where $\mathbf{K}_i$ is a matrix of known covariates, $\boldsymbol{\beta}$ $(p \times 1)$ is a vector of unknown regression coefficients and $\mathbf{b}_i$ is a vector of independent elements of length q. Consequently, we obtain:

$$\begin{aligned}
\mathbf{y}_i &= \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \\
&= \mathbf{Z}_i(\mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i) + \boldsymbol{\epsilon}_i \\
&= \mathbf{Z}_i\mathbf{K}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \\
&= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i
\end{aligned}$$

where $\mathbf{X}_i$ is the fixed effects regressor matrix. The estimates $\hat{\boldsymbol{\beta}}$ are used to provide inferences for $\boldsymbol{\beta}$.

However, this two-stage process is not that innocent; Firstly, information is lost in summarising the $\mathbf{y}_i$ by the estimated vector of subject-specific regression coefficients, $\hat{\boldsymbol{\beta}}$. Secondly, there is the problem that the covariance matrix of $\hat{\boldsymbol{\beta}}$ is highly dependent on the number of measurements available for each subject and also on when the measurements were taken.

Marginal models are those that are most commonly used in order to make inference about population means. Marginal models for longitudinal data model the mean response and the within-subject association among repeated responses obtained separately (Davis, 2002; Fitzmaurice et al., 2004). Marginal modeling approach assumes that the marginal expectation $(E(y_{ij}) = \mu_{ij})$ can be related to the covariates through a known link function (g):

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}$$

Moreover, the conditional variance of each y_{ij} given the covariates, depends on the mean in the following way:

$$var(y_{ij}) = \phi v(\mu_{ij})$$

where $v(\mu_{ij})$ is a known variance function of the mean and ϕ is a scale parameter (Davis, 2002; Fitzmaurice et al., 2004).

There is a great controversy in bibliography about which modeling approach is more preferable. Lee and Nelder (2004) argue that the conditional modeling approach is preferable to the marginal modeling approach since both marginal inferences and conditional inferences can be obtained, i.e. one can have both $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ and $E(\mathbf{y}_i|\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$. Since the expected value for the mean of the random effects is constrained to equal zero, this means that the fixed effects estimates of a conditional model have the same meaning as those of the marginal model. The authors show that if the individuals in a study have significant random treatment effects (e.g. random time effects), these will be confounded with the fixed treatment effects in a marginal model, whereas for a conditional model these two different treatment effects will have separate estimates. The marginal estimates for the fixed effects are then only useful if there is no interaction effect between the subject and the treatment and this can only be checked by means of a conditional model. In addition, the authors conclude that conditional models allow for the estimation of two different types of error: random error and subject-specific error, which is not possible through the marginal modeling approach.

2.3 Generalized Estimating Equations

2.3.1 Introduction to Generalized Estimating Equations

In this section we outline the main idea of the GEE method in the context of repeated measurements. The interest in analyzing longitudinal data is to describe the dependence of the outcome on predictor variables (marginal expectation of the outcome Y as a function of covariates X, i.e. $E(Y|X)$). Repeated measurements tend to be correlated since they are made on the same subject. For example, two measurements from the same subject are likely to be more correlated than two measurements from different subjects. Since most statistical tests assume independence of observations, it is crucial to take the within-subject correlation into account to obtain correct statistical analysis. If we do not take the correlation into account, it can lead to a wrong test statistic and inference, for example standard errors will likely be too small. (Burton et al., 1998 , Zeger & Liang, 1986).

There are many techniques for analysis when the outcome variable is approximately normally distributed (e.g. fitting growth curves for each subject using repeated measurements). Difficulty in analysis comes from the lack of multivariate joint distribution of the outcome variable, hence likelihood methods are not available or are difficult to compute (Liang & Zeger 1986).

Linear models for normally distributed data have been expanded to non-normal data using generalized estimation methods and quasi-likelihood when there is a single observation for each subject (no repeated measurements). Quasi-likelihood approach does not assume distribution, it only specifies a linear function between marginal expectation of the outcome variable and covariates and assumes that variance (of the outcome variable) is a known function of its expectation (Zeger & Liang 1986).

As it was mentioned above Generalized Estimating Equations is a general statistical method to fit marginal models for correlated or clustered responses and it uses a robust sandwich estimator to estimate the variance-covariance matrix of the regression coefficient estimates. We begin by introducing some useful notation. We assume that K subjects are measured repeatedly over time. Let Y_{ij} denote the response variable for the i^{th} subject on the j^{th} measurement, given longitudinal data consisting of

K subjects, $i = 1, 2, ..., K$ and $j = 1, 2, ..., n_i$. The response variable could be continuous, binary or a count. The nature of the response variable does have really important and useful implications for model specification; however the notation does not distinguish between the different types of responses. Also, let $\mathbf{X}_{ij}$ be a $p \times 1$ covariates vector. $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$ denotes the response vector with the mean vector noted by $\boldsymbol{\mu_i} = (\mu_{i1}, \mu_{i2}, ..., \mu_{in_i})'$ where μ_{ij} is the corresponding j^{th} mean for subject i. Although there exists within-subject correlation, the observations across subjects are assumed to be independent. The marginal model specifying an association between μ_{ij} and the covariates of interest $\mathbf{X}_{ij}$ is given by

$$g(\mu_{ij}) = \mathbf{X}'_{ij}\boldsymbol{\beta} \tag{1}$$

with g as a known link function and $\boldsymbol{\beta}$ an unknown $p \times 1$ vector of regression coefficients. The conditional variance of Y_{ij} given $\mathbf{X}_{ij}$ is $Var(Y_{ij}|\mathbf{X}_{ij}) = v(\mu_{ij})\varphi$ with v as a known variance function of μ_{ij} and φ a scale parameter which may need to be estimated.

For the case of univariate QL, the estimates $\hat{\boldsymbol{\beta}}$ for a GLM are solutions of likelihood equations:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{K} (\frac{\partial \mu_i}{\partial \boldsymbol{\beta}})' \frac{(y_i - \mu_i)}{v(\mu_i)} = \mathbf{0}$$

for variance function $v(\mu)$.

The estimators $\hat{\boldsymbol{\beta}}$ are asymptotically normal with model-based covariance matrix approximated by

$$\mathbf{V} = \Big[\sum_{i=1}^{K} (\frac{\partial \mu_i}{\partial \boldsymbol{\beta}})' \big[v(\mu_i)\big]^{-1} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \Big]^{-1}$$

With quasi-likelihood approach, we use our own variance function (e.g., for count data, $v(\mu) = c\mu$ with unknown constant c estimated from data), typically to permit overdispersion.

When we misspecify the variance function, the actual asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$var(\hat{\boldsymbol{\beta}}) \approx \mathbf{V}\Big[\sum_{i=1}^{K}(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}})'\frac{var(y_i)}{\big[v(\mu_i)\big]^2}(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}})\Big]\mathbf{V}$$

In practice, true $var(y_i)$ is unknown. Thus, one can estimate $var(\hat{\beta})$ by sample analog (sandwich estimator), replacing μ_i by $\hat{\mu}_i$ and $var(y_i)$ by $(y_i - \hat{\mu}_i)^2$.

For the multivariate QL (GEE),

$$var(\hat{\boldsymbol{\beta}}) \approx K\Big[\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\Big]^{-1} M \Big[\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\Big]^{-1}$$

with

$$M = \Big[\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{V}_i^{-1}var(\mathbf{y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i\Big]$$

where

$$\mathbf{D}_i = \frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}} \text{ and } \mathbf{V}_i \text{ the working covariance matrix.}$$

To obtain estimated covariance matrix, we replace the parameters by their estimates and we replace $var(\mathbf{y}_i)$ by $(\mathbf{y}_i - \boldsymbol{\mu_i})(\mathbf{y}_i - \boldsymbol{\mu_i})'$ to get an empirical sandwich covariance matrix that yields more robust SE values.

To estimate $\boldsymbol{\beta}$, Liang and Zeger (1986) proposed solving the estimating equations

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{K}(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}})'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{2}$$

where $\mathbf{V}_i$ is the variance-covariance matrix for $\mathbf{Y}_i$, noted by $\mathbf{V}_i = \varphi\mathbf{A}_i^{\frac{1}{2}}\mathbf{R}_i(\mathbf{a})\mathbf{A}_i^{\frac{1}{2}}$. Let $\mathbf{A}_i = Diag(v(\mu_{i1}), ..., v(\mu_{in_i}))$, while the working correlation/association structure $\mathbf{R}_i(\mathbf{a})$ describes the correlation pattern of observations within-subject with $\mathbf{a}$ as a vector of association parameters specifying the correlation structure. Several types of correlation structure can be used depending on the occasion, including independent, exchangeable and autoregressive structure. As Ming Wang et al (2015) mention, the estimation of $\mathbf{a}$ is based on an iterative fitting process using the Pearson residual $e_{ij} = \frac{y_{ij}-\hat{\mu}_{ij}}{\sqrt{v(\hat{\mu}_{ij})}}$. Additionally, the scale parameter φ is estimated by

$\hat{\varphi} = \frac{1}{N-p} \sum_{i=1}^{K} \sum_{j=1}^{n_i} e_{ij}^2$ with the total number of observations $N = \sum_{i=1}^{K} n_i$.

The GEE method yields asymptotically consistent $\hat{\boldsymbol{\beta}}$ even when the correlation structure is misspesified. Under mild regularity conditions (the parameter space is an open set and the GEE function $\mathbf{U}(\boldsymbol{\beta})$ is continuously differentiable) and given the true value of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_t$, $\boldsymbol{\beta}$ is asymptotically normally distributed with a mean its true value and a covariance matrix estimated based on the "sandwich" estimator by

$$V_{LZ} = \left(\sum_{i=1}^{K} (\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}})' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^{-1} M_{LZ} \left(\sum_{i=1}^{K} (\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}})' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^{-1} \tag{3}$$

with

$$M_{LZ} = \sum_{i=1}^{K} (\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}})' V_i^{-1} Cov(\mathbf{Y}_i) \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \tag{4}$$

where $Cov(\mathbf{Y}_i) = \hat{\mathbf{r}}_i \hat{\mathbf{r}}_i{}'$ with $\hat{\mathbf{r}}_i = \mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i$ an estimator of the variance-covariance matrix of $\mathbf{Y}_i$ and $\mathbf{a}$, $\boldsymbol{\beta}$, φ can be replaced with their consistent estimates. The specification of the covariance matrix is not always necessary to be correct and for this reason we usually refer to V_i as "working" covariance matrix. The GEE solution will be consistent as long as $E(Y_i - \mu_i) = 0$, which indicates the importance of the correct specification of the mean. Additionally, this "sandwich" estimator is robust and consistent even if the correlation structure is misspecified, a fact that makes its use more appropriate. A consistent estimator for the covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by $\left(\sum_{i=1}^{K} (\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}})' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^{-1}$ which is also referred to as the model-based variance estimator.

Alternative estimators of the covariance matrix have been proposed by Paik (1988) as well as by Mancl and DeRouen (2001) using Jackknife estimators for samples with small sizes and by Pan (2001) under the assumption of a common correlation matrix across the subjects.

Generalized Estimating Equations are based on quasi-likelihood method of estimation. In addition to previously mentioned assumptions (expectation of outcome variable to be a linear function of covariates and that variance is a known function of the mean), one needs to specify the working correlation structure between the repeated measurements for each subject. The general idea is to incorporate the correlation structure between repeated measurements to get consistent estimators of coefficients

and of their variances (Zeger & Liang 1986).

The next section will give the general idea of quasi-likelihood estimator of which GEE is based on.

2.3.2 Quasi-likelihood estimator

This section is a short overview of quasi-likelihood used in GEE based on Zeger & Liang 1986. Quasi-likelihood is a methodology for regression that requires the specification of relationships between mean response and covariates and between mean response and variance. Thus it does not assume a probability distribution as in the case of full likelihood.

Let $\mathbf{Y}_i$ be the response variable for each subject $i = 1, \ldots, K$ and $\mathbf{X}_i$ be $p \times 1$ vector of covariates. Let $\boldsymbol{\beta}$ be $p \times 1$ vector of regression parameters to be estimated. Define $\boldsymbol{\mu}_i = E(\mathbf{Y}_i|\mathbf{X}_i)$ to be the conditional expectation of Y_i and a function of covariates and regression parameters, so that $\boldsymbol{\mu}_i = h(\mathbf{X}_i'\boldsymbol{\beta})$. The inverse of h is the link function which relates the mean response to the linear predictor $\mathbf{X}_i'\boldsymbol{\beta}$. For quasi-likelihood, variance of each $\mathbf{Y}_i$, denoted as u_i, is a known function of the expectation $\boldsymbol{\mu}_i$, so that $u_i = f(\boldsymbol{\mu}_i)\phi$. The scale parameter ϕ is treated as a nuisance parameter. The quasi-likelihood estimator is the solution to the equations:

$$S_k(\boldsymbol{\beta}) = \sum_{i=1}^{N} (\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}_k}) u_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0, \quad \text{for k = 1, 2, ..., p}$$

Estimators of regression parameters, $\hat{\boldsymbol{\beta}}$, are obtained by iteratively reweighted the least squares method.

2.3.3 Marginal Models

A standard GEE is known as a marginal model. Marginal models extend generalized linear models to longitudinal data and are typically used when the inference is population-based, rather than individual-based. The term "marginal" means that in the model specification the expected value of the response variable Y, depends only on covariates (fixed effects) and does not depend on subject specific random effects nor directly on previous responses of the subject. Since the purpose is to describe the

changes in population mean rather than changes within subjects, within-subject correlation is regarded as a nuisance characteristic. Regression parameters and within-subject correlation is modelled separately (Fitzmaurice et al. 2004).

Let's introduce some notation for the repeated measurements. We have K subjects who are measured repeatedly. Y_{ij} denotes the response variable for the i^{th} subject on the j^{th} measurement occasion. A realisation of each Y_{ij} is observed at time t_{ij}. The response variable can be continuous, binary, multinomial or a count. We assume the data are unbalanced (the number of repeated measurements can be different for subjects and/or they can be measured at different occasions) and that there are n_i repeated measurements for the i^{th} subject.

The response variable is a $n_i \times 1$ vector

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})', \; i = 1, 2, ..., K$$

Y_i are assumed to be independent, but observations within the subject are not assumed to be independent. Associated with each response at a given time point j, there is a $p \times 1$ vector of covariates

$$\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, ..., X_{ijp})', \; i = 1, ..., K, \; j = 1, ..., n_i$$

They can be either time-invariant or time-dependent. Time-invariant variable is fixed within a subject at the same value irrespective of time point j, whereas time-dependent variable is varying with time for each subject.

The GEE requires the following specifications for a marginal model:

(1) Conditional mean of Y_{ij} is related to covariates by a known link function,

$$g(\mu_{ij}) = n_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta}$$

where $\mu_{ij} = E(Y_{ij}|\mathbf{X}_{ij})$ is a conditional expectation (or mean) of the response variable and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters.

(2) The conditional variance of each Y_{ij} may depend on the mean response, given the effects of covariates, as

$$var(Y_{ij}) = \phi v(\mu_{ij})$$

where $v(\mu_{ij})$ is a known variance function of the mean and ϕ is a scale parameter that may be known or may need to be estimated (Davis, 2002; Fitzmaurice et al., 2004). When the response is a continuous variable, then variance of each Y_{ij} does not depend on mean response and is $var(Y_{ij}) = \phi v(\mu_{ij}) = \phi$. Note that this assumes homogeneity of variance over time, which is often too strong of an assumption.

(3) Correlation among repeated measurements is a function of the means, μ_{ij} , and a set of parameters, $\boldsymbol{\alpha}$, which characterize the within-subject correlation and need to be estimated. The "working" covariance matrix for Y_i is given by

$$\mathbf{V}_i = \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{\frac{1}{2}} / \phi$$

Correlation matrices $\mathbf{R}_i(\boldsymbol{\alpha})$ can be different for subjects, however, it is fully specified by $\boldsymbol{\alpha}$, which is the same for all subjects. $\mathbf{A}_i$ is a $n_i \times n_i$ diagonal matrix with $g(\mu_{ij})$ on the diagonal. "Working" covariance means that we do not know the true correlation structure between repeated measurements and are not assuming we are specifying it correctly. We would like to get consistent estimates of regression parameters regardless of the chosen structure (Fitzmaurice et al. 2004, Zeger & Liang 1986).

2.3.4 Working Correlation Structures

In this section we present five basic working correlation structures;

Independence

Independence is the most basic structure where each observation within a subject is uncorrelated with another observation.

$$Cor(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ 0 & otherwise \end{cases},$$

for instance, a 4×4 correlation matrix with independence structure is the following

$$R_{IN} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Exchangeable

In exchangeable working correlation structure, responses are assumed to be equally correlated within an individual.

$$Cor(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha & otherwise \end{cases},$$

$$R_{EX} = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}$$

Autoregressive AR-1

Two observations closer in time are more correlated than two observations more further in time. This structure is often used in longitudinal designs. Note that n_i in this example is 4.

$$Cor(Y_{ij}, Y_{i,j+t}) = \alpha^t,\ t = 0, ..., n_i - j\ ,$$

$$R_{AR-1} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}$$

Toeplitz

Correlation is the same for any two observations that have the same distance in time. Note that n_i in this example is 4.

$$Cor(Y_{ij}, Y_{i,j+t}) = \begin{cases} 1 & t = 0 \\ \alpha_t & t = 1, ..., n_i - j \end{cases},$$

$$R_{TOEP} = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ \alpha_3 & \alpha_2 & \alpha_1 & 1 \end{pmatrix}$$

Unstructured

There is no assumption made about any two observations within a subject, so correlation can take a value between -1 and 1. This type of correlation is the most flexible one, but the number of parameters can become too high very quickly.

$$Cor(Y_{ij}, Y_{i,k}) = \begin{cases} 1 & j = k \\ \alpha_{jk} & otherwise \end{cases},$$

$$R_{UN} = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} \\ \alpha_{14} & \alpha_{24} & \alpha_{34} & 1 \end{pmatrix}$$

2.4 Modified Variance Estimators of Generalized Estimating Equations with small samples

This chapter outlines a main problem that arises when the sample size is small. Due to the fact that the fitted value $\hat{\mu}_i$ tends to be closer to Y_i than the true value μ_i , the term $\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i{}'$ in V_{LZ} is biased for estimating $E(\mathbf{e}_i\mathbf{e}_i')$ and the bias tends to be larger when the sample size is much smaller. Additionally, the hypothesis testing tends to be too liberal and the resulting confidence interval is narrow. We present the eight variance modifications as they were proposed by Wang et. al (2015).

The first modified variance estimator is denoted by V_{MK} and provides a degrees-of-freedom adjustment of "sandwich" variance estimator proposed by MacKinnon (1985). This specific estimator seems quite simple, as it incorporates the factor of $\frac{K}{K-p}$, where K is the number of subjects being measured. The formula becomes

$$V_{MK} = \frac{K}{K-p} V_{LZ} \tag{1}$$

When $K \to \infty$, then $V_{MK} \to V_{LZ}$. V_{MK} corrects the bias but increases the variability.

The second modified estimator which is denoted by V_{KC} and proposed by Kauermann and Carroll (2001) is a bias-corrected "sandwich" variance estimator under the assumption of the correct specification of the correlation structure. This estimator is given by

$$V_{KC} = \left(\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1} M_{KC} \left(\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1} \tag{2}$$

with $\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ and

$$M_{KC} = \sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}(\mathbf{I}_i - \mathbf{H}_{ii})^{-1/2}\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i{}'(\mathbf{I}_i - \mathbf{H}_{ii}')^{-1/2}\mathbf{V}_i^{-1}\mathbf{D}_i \tag{3}$$

where $\mathbf{I}_i$ is an $n_i \times n_i$ identity matrix and subject leverage $\mathbf{H}_{ii}$ is a diagonal matrix with the leverage of the i^{th} subjects, which can be calculated by

$$\mathbf{H}_{ii} = \mathbf{D}_i(\textstyle\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i)^{-1}\mathbf{D}_i'\mathbf{V}_i^{-1}$$

Furthermore, V_{PAN} is the third modified variance estimator which was proposed by Pan (2001) with two additional assumptions. Firstly, the conditional variance of Y_{ij} given X_{ij} has to be correctly specified. Secondly, a common correlation structure R_c has to exist across all subjects. This modified variance estimator is given by

$$V_{PAN} = \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} M_{PAN} \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} \tag{4}$$

with

$$M_{PAN} = \sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \left\{ \mathbf{A}_i^{1/2} \left(\frac{1}{K} \sum_{i=1}^{K} \mathbf{A}_i^{-1/2} \hat{\mathbf{r}}_i \hat{\mathbf{r}}_i' \mathbf{A}_i^{-1/2} \right) \mathbf{A}_i^{1/2} \right\} \mathbf{V}_i^{-1} \mathbf{D}_i \tag{5}$$

V_{PAN} performs more efficiently as it pools data across all subjects in estimating $Cov(\mathbf{Y}_i)$.

We continue with the forth modified variance estimator which is denoted by V_{GST} and was proposed by Gosho et al (2014). V_{GST} made an additional modification on Pan's estimator by incorporating the bias of the term $\mathbf{A}_i^{1/2}(\frac{1}{K}\sum_{i=1}^{K} \mathbf{A}_i^{-1/2} \hat{\mathbf{r}}_i \hat{\mathbf{r}}_i' \mathbf{A}_i^{-1/2}) \mathbf{A}_i^{1/2}$ for small K. This estimator is written as

$$V_{GST} = \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} M_{GST} \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} \tag{6}$$

with

$$M_{GST} = \sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \left\{ \mathbf{A}_i^{1/2} \left(\frac{1}{K-p} \sum_{i=1}^{K} \mathbf{A}_i^{-1/2} \hat{\mathbf{r}}_i \hat{\mathbf{r}}_i' \mathbf{A}_i^{-1/2} \right) \mathbf{A}_i^{1/2} \right\} \mathbf{V}_i^{-1} \mathbf{D}_i \tag{7}$$

The V_{GST} estimator has a similar bahavior with the V_{PAN} estimator, as it also pools data across all subjects in estimating $Cov(\mathbf{Y}_i)$. Particularly, V_{GST} approximately equals to V_{PAN} when K is large enough and $K >> p$.

The fifth modified variance estimator which is denoted by V_{MD} and was proposed by Mancl and DeRouen (2001) is another bias-corrected "sandwich" variance

estimator which is written as

$$V_{MD} = \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} M_{MD} \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} \tag{8}$$

with

$$M_{MD} = \sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \hat{\mathbf{r}}_i \hat{\mathbf{r}}_i' (\mathbf{I}_i - \mathbf{H}_{ii}')^{-1} \mathbf{V}_i^{-1} \mathbf{D}_i \tag{9}$$

This estimator, unlike V_{KC} does not assume a correctly specified correlation structure while $\mathbf{I}_i$ and $\mathbf{H}_{ii}$ are defined the same as V_{KC}.

Moreover, it is worthmentioned that Mancl and DeRouen in order to correct the bias in finite samples, relied on the approximate identity

$$E(\hat{\mathbf{r}}_i \hat{\mathbf{r}}_i{}') \approx (\mathbf{I}_i - \mathbf{H}_{ii}) Cov(\mathbf{Y}_i) (\mathbf{I}_i - \mathbf{H}_{ii})'$$

ignoring the term $\sum_{j \neq i} \mathbf{H}_{ij} \mathbf{Cov}(\mathbf{Y}_i) \mathbf{H}_{ij}^T$ from its first-order Taylor expansion leading to overcorrection (Wang et. al, 2015; Mancl and DeRouen, 2001).

We continue our analysis with the sixth modified variance estimator that was proposed by Fay and Graubard (2001) and is denoted by V_{FG}. This estimator introduces a further adjustment on V_{MD} for a simple bias correction. The formula is given by

$$V_{FG} = \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} M_{FG} \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} \tag{10}$$

with

$$M_{FG} = \sum_{i=1}^{K} \mathbf{n}_i^{-1} \mathbf{D}_i' \mathbf{V}_i^{-1} \hat{\mathbf{r}}_i \hat{\mathbf{r}}_i{}' \mathbf{V}_i^{-1} \mathbf{D}_i \mathbf{n}_i'^{-1} \tag{11}$$

where $\mathbf{n}_i = \mathbf{I}_p - \mathbf{N}_i$. The jj^{th} diagonal value of $\mathbf{n}_i^{-1/2}$ is equal to $(1 - min(b, \{N_i\}_{jj}))^{-1}$, where $\mathbf{N}_i = \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i (\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1}$ and b is prespecified subjectively to avoid extreme adjustments when $\mathbf{N}_i$ is close to 1.

The seventh modified variance estimator, V_{MBN}, provides a bias correction to the "sandwich" variance estimator and was suggested by Morel et al (2003). V_{MBN} incorporates correlation on the residual cross-products and sample size and is given by

$$V_{MBN} = \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} M_{MBN} \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} \tag{12}$$

with

$$M_{MBN} = \sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} (k \hat{\mathbf{r}}_i \hat{\mathbf{r}}_i{}' + \delta_m \xi \mathbf{V}_i) \mathbf{V}_i^{-1} \mathbf{D}_i \tag{13}$$

where $k = \frac{N-1}{N-p} \frac{K}{K-1}$, $\delta_m = \begin{cases} \frac{p}{K-p} & K > (d+1)p \\ \frac{1}{d} & otherwise \end{cases}$

and

$$\xi = max\left(r, \frac{trace\left(\left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} M_{LZ}\right)}{p}\right) \text{ with } 0 \leq r \leq 1.$$

Morel et al. (2003) mentioned that k is a factor to adjust the bias of empirical variance estimator of $Cov(\mathbf{Y}_i)$ and δ_m can be bounded by $1/d$. The default value for d is 2 and for r is 1, respectively.

Last but not least, we introduce the final among the most recent variance estimators, which is a combination of V_{PAN} and V_{MD} for pooling information from all subjects and also reducing the bias of the estimate for $\mathbf{e}_i \mathbf{e}_i'$. This estimator is denoted by V_{WL} and is recommended by Wang and Long (2011), while it is written as

$$V_{WL} = \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} M_{WL} \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} \tag{14}$$

with

$$M_{WL} = \sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{A}_i^{1/2} \left\{ \sum_{i=1}^{K} \mathbf{A}_i^{-1/2} (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \hat{\mathbf{r}}_i \hat{\mathbf{r}}_i' (\mathbf{I}_i - \mathbf{H}_{ii}')^{-1} \mathbf{A}_i^{-1/2} / K \right\} \mathbf{A}_i^{1/2} \mathbf{V}_i^{-1} \mathbf{D}_i \tag{15}$$

As this estimator is the combination of the strength of both previous estimators V_{PAN} and V_{MD}, it was supposed to perform as well as or better than those two. The additional assumptions specified in the notation of V_{PAN} also need to be satisfied in V_{WL}.

Since we have presented the eight most recent adjustments and corrections on the classical "sandwich" variance estimator, we continue our analysis by comparing theoretically those variance estimators. One can notice that all these estimators share the same two outside terms, that is $(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1}$. As a consequence, the middle matrix, M, differentiates itself among the eight corrections. Wang and Long (2011) have shown that the modifications through the degrees-of-freedom adjustment or bias-correction are mostly applied when the sample size is small, as V_{LZ} tends to underestimate the variance. Additionally, V_{PAN}, V_{GST} and V_{WL} incorporate the efficiency gain by pooling data across all subjects in order to improve the estimator of $Cov(\mathbf{Y}_i)$ instead of using only data from the i^{th} subject.

V_{WL} is the only estimator which takes into consideration both bias correction and efficiency improvement. Thus, it is expected to outperform the other alternatives when the two assumptions mentioned before are satisfied.

In the following tables we provide a summary of the eight modified variance estimators that we introduced above, based on Wang et.al (2015).

Table 1: Summary of eight modified variance estimators for GEE with small sample.

Variance estimator	Modification	Reference
V_{MK}	Degrees-of-freedom adjustment	MacKinnon (1985)
V_{KC}	Bias correction	Kauermann and Carroll (2001)
V_{PAN}	Efficiency improvement	Pan (2001)
V_{GST}	Efficiency improvement	Gosho et al. (2014)
V_{MD}	Bias correction	Mancl and DeRouen (2001)
V_{FG}	Bias correction	Fay and Graubard (2001)
V_{MBN}	Bias correction	Morel et al. (2003)
V_{WL}	Bias correction and efficiency improvement	Wang and Long (2011)

Table 2: Covariance matrix of the middle parts from nine variance estimators for GEE.

Matrix M	Covariance matrix of vec(M)
M_{LZ}	$\sum_{i=1}^{K} \mathbf{S}_i\mathbf{T}_i\mathbf{S}_i'$
M_{MK}	$\sum_{i=1}^{K} \frac{K^2}{(K-p)^2}\mathbf{S}_i\mathbf{T}_i\mathbf{S}_i'$
M_{KC}	$\sum_{i=1}^{K} \mathbf{S}_i\mathbf{F}_i\mathbf{T}_i\mathbf{F}_i'\mathbf{S}_i'$
M_{PAN}	$\sum_{i=1}^{K} \mathbf{S}_i\big[\mathbf{E}_i(\sum_{j=1}^{K} \frac{1}{K^2}\mathbf{E}_j^{-1}\mathbf{T}_j\mathbf{E}_j^{-1})\mathbf{E}_i\big]\mathbf{S}_i'$
M_{GST}	$\sum_{i=1}^{K} \mathbf{S}_i\big[\mathbf{E}_i(\sum_{j=1}^{K} \frac{1}{(K-p)^2}\mathbf{E}_j^{-1}\mathbf{T}_j\mathbf{E}_j^{-1})\mathbf{E}_i\big]\mathbf{S}_i'$
M_{MD}	$\sum_{i=1}^{K} \mathbf{S}_i\mathbf{G}_i\mathbf{T}_i\mathbf{G}_i'\mathbf{S}_i'$
M_{FG}	$\sum_{i=1}^{K} \mathbf{H}_i\mathbf{T}_i\mathbf{H}_i'$
M_{MBN}	$\sum_{i=1}^{K} \mathbf{S}_i\mathbf{N}_i\mathbf{S}_i'$
M_{WL}	$\sum_{i=1}^{K} \mathbf{S}_i\big[\mathbf{E}_i(\sum_{j=1}^{K} \frac{1}{K^2}\mathbf{E}_j^{-1}\mathbf{G}_j\mathbf{T}_j\mathbf{G}_j'\mathbf{E}_j^{-1})\mathbf{E}_i\big]\mathbf{S}_i'$

$\mathbf{T}_i = Cov(vec(\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i{}'))$; $\mathbf{S}_i = (\mathbf{D}_i'\mathbf{V}_i^{-1}) \otimes (\mathbf{D}_i'\mathbf{V}_i^{-1})$;

$\mathbf{F}_i = (\mathbf{I}_i - \mathbf{H}_{ii})^{-1/2} \otimes (\mathbf{I}_i - \mathbf{H}_{ii})^{-1/2}$;

$\mathbf{G}_i = (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \otimes (\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$; $\mathbf{E}_i = \mathbf{A}_i^{1/2} \otimes \mathbf{A}_i^{1/2}$;

$\mathbf{H}_i = (\mathbf{n}_i^{-1}\mathbf{D}_i'\mathbf{V}_i^{-1}) \otimes (\mathbf{n}_i^{-1}\mathbf{D}_i'\mathbf{V}_i^{-1})$;

$\mathbf{N}_i = Cov(kvec(\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i{}') + vec(\delta_m\xi\mathbf{V}_i))$.

3 Simulation Study

This chapter aims to compare the performance of the original "sandwich" variance estimator for small samples with the eight modified variance estimators through simulation studies. The Wald test and t-test are used for hypothesis testing in order to calculate the type I error rate for each estimator. We generated data sets with equal cluster sizes. Three models, one for each type of response repeated outcome (continuous, count and binary) are applied.

The models we used for data generation are the following:

$$Y_{ij} = \beta_0 + \beta_1 \times x_{ij} + b_i + \epsilon_{ij} \tag{1}$$

$$log(u_{ij}|b_i) = \beta_0 + \beta_1 \times x_{ij} + b_i \tag{2}$$

$$logit(u_{ij}|b_i) = \beta_0 + \beta_1 \times x_{ij} + b_i \tag{3}$$

The null hypothesis is $\beta_0 = 0$ and $\beta_1 = 0$ for $i = 1, 2, ..., K$ with sample size $K = 10, 20, 30, 40, 50$ and $j = 1, 2, ..., n$ with equal number of observations within-subject (i.e., cluster size) $n = 5, 10$. The covariate x_{ij} follows the standard normal distribution $N(0, 1)$ and is independent and identical distributed. The subject-level random effects b_i's are also independent and identical distributed from the normal $N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.45$ and the random errors are also i.i.d. from the normal distribution $N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 0.8$.

More specifically, for the case with continuous outcomes, b_i and ϵ_{ij} are independent with each other with the correlation parameter $a = \frac{\sigma_b^2}{\sigma_b^2+\sigma_\epsilon^2} \approx 0.4$.

For the case with count outcomes, according to Guo et al. (2005), the correlation parameter is $a \approx \frac{\sigma_b^2}{1+\sigma_b^2} \approx 0.3$.

Last but not least, for the case with binary outcomes the correlation parameter is given by

$a \approx \frac{\sigma_b^2/16}{E(\frac{1}{1+exp(-b_i)})[1-E(\frac{1}{1+exp(-b_i)})]} \approx 0.1$ according to Guo et al. (2005).

In Generalized Estimating Equations, the Wald test as well as the score test are the most commonly used for the hypothesis testing. However, the Wald test gives bigger than we expected type I error when the sample size is small and the score test has smaller test size than the nominal level. In order to avoid these problems, two modified tests have been proposed; the t-test and the modified score test. Supposing that the parameter of interest is denoted by β and for the simple univariate case, the null hypothesis is given by $H_0 : \beta = 0$ against the alternative hypothesis $H_1 : \beta \neq 0$. The test statistic for the Wald test is $z = \frac{\hat{\beta}}{\sqrt{\hat{V}(\hat{\beta})}}$, where $\hat{V}(\hat{\beta})$ can be replaced by any of the modified variance estimators. We denote k as the estimated mean and v as the estimated variance of $V(\hat{\beta})$. The distribution of $\frac{\hat{V}(\hat{\beta})}{c}$ is approximated with a chi-square distribution X_d^2 where the scale parameter is given by $c = \frac{v}{2k}$ and the degrees of freedom by $d = \frac{2k^2}{v}$. Moreover, the t-test has the following test statistic; $t = \frac{\hat{\beta}/\sqrt{k}}{\sqrt{\hat{V}(\hat{\beta})/cd}}$ which is similar to that of Wald statistic with the degrees of freedom $d \approx 2\hat{V}(\hat{\beta})^2/\widehat{Var}(\hat{V}(\hat{\beta}))$ (Wang and Long, 2011; Pan, 2001). This approximation incorporates the variability of the variance estimator and as a result, it performs better compared with that proposed by Li and Redden (2015) which depends only on the number of clusters.

Under certain regularity conditions, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ has approximately in large samples a multivariate normal distribution with mean equal to the true parameter value and variance-covariance matrix given by the inverse of the information matrix, so that $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \mathbf{I}^{-1}(\boldsymbol{\beta}))$. The regularity conditions include the following: the true parameter value $\boldsymbol{\beta}$ must be interior to the parameter space, the log-likelihood function must be thrice differentiable, and the third derivatives must be bounded. This result provides a basis for constructing tests of hypotheses and confidence regions. For instance, under the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ for a fixed value $\boldsymbol{\beta}_0$, the quadratic form

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' var^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

has approximately in large samples a chi-squared distribution with p degrees of freedom.

The simulation consists of 1000 Monte Carlo iterations for each model, where the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are calculated with all nine variance estimates. Three types of "working" correlation structures are used: independence, exchangeable and

AR-1, while the Wald test and t-test are applied for hypothesis testing and empirical type I error is calculated at the significance level of 0.05 . The true variance of the regression coefficient estimate, $\hat{\beta}_1$, was obtained by the variance of $\hat{\beta}_1$'s from the 1000 Monte Carlo data sets. Moreover, the degrees of freedom for t-distribution vary across different variance estimators, indicating the variability influence of variance estimators.

The following table was computed in R and provides information about the performance of all nine variance estimators. For brevity and ease in comparing, we present a small part of it.

Table 3: Simulation results for normal distributed responses Y_{ij} with the underlying true correlation coefficient $\alpha = 0.2$ and 95% nominal level.

n	K	Variance estimator	$V(\hat{\beta}_1)(SD)$	$CR_Z(CR_T)$
			Exchangeable	
5	10	True	0.029	
		V_{LZ}	0.022(0.014)	0.89(0.92)
		V_{PAN}	0.024(0.008)	0.92(0.94)
		V_{MD}	0.032(0.022)	0.93(0.95)
		V_{WL}	0.028(0.011)	0.95(0.95)
	20	True	0.013	
		V_{LZ}	0.012(0.005)	0.92(0.94)
		V_{PAN}	0.011(0.003)	0.93(0.94)
		V_{MD}	0.014(0.007)	0.94(0.95)
		V_{WL}	0.013(0.003)	0.94(0.95)

$V(\hat{\beta}_1)$ is the average estimated variance of $\hat{\beta}_1$;
SD is the Monte Carlo standard deviation of the estimated variance of $\hat{\beta}_1$;
CR is the Monte Carlo coverage rate of Wald confidence interval for β_1;
Z: Wald-test; T: t-test.

The results from all different types of outcomes (continuous, count and binary) are similar to those in the above table and can be summarized as follows: (1) The V_{LZ} estimator tends to underestimate the true sampling variance of $\hat{\beta}_1$ and the resulting coverage rates fall far short of nominal levels.

(2) For moderate sample size, all modifications achieve similar performance in terms of coverage rates.

(3) The coverage rates of confidence intervals based on t-tests are higher than those using Wald tests and are closer to nominal levels in most cases.

(4) The V_{WL} estimator exhibits smaller bias and leads to coverage rates closer to nominal levels comparing to the other variance estimators.

We continue by presenting the following Figures which were conducted in R and then we summarize our results. We note that the figures based on the "independent" working correlation structure are omitted, as they present similar trend to that of AR-1.

https://doi.org/10.26219/heal.aueb.6669

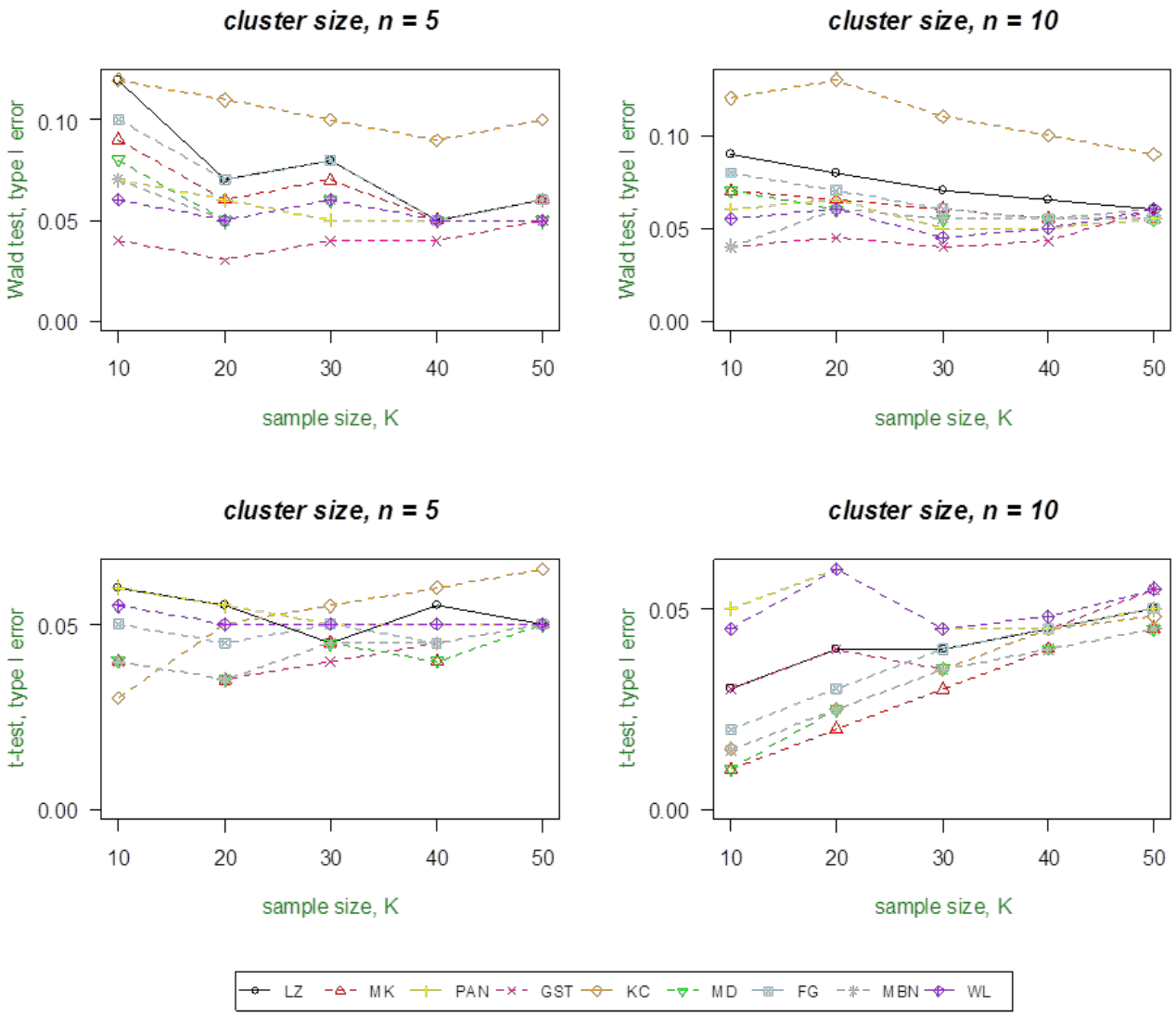

Figure 1: Type I errors based on Wald test and t-tests for **continuous outcomes** with the **exchangeable** correlation structure.

https://doi.org/10.26219/heal.aueb.6669

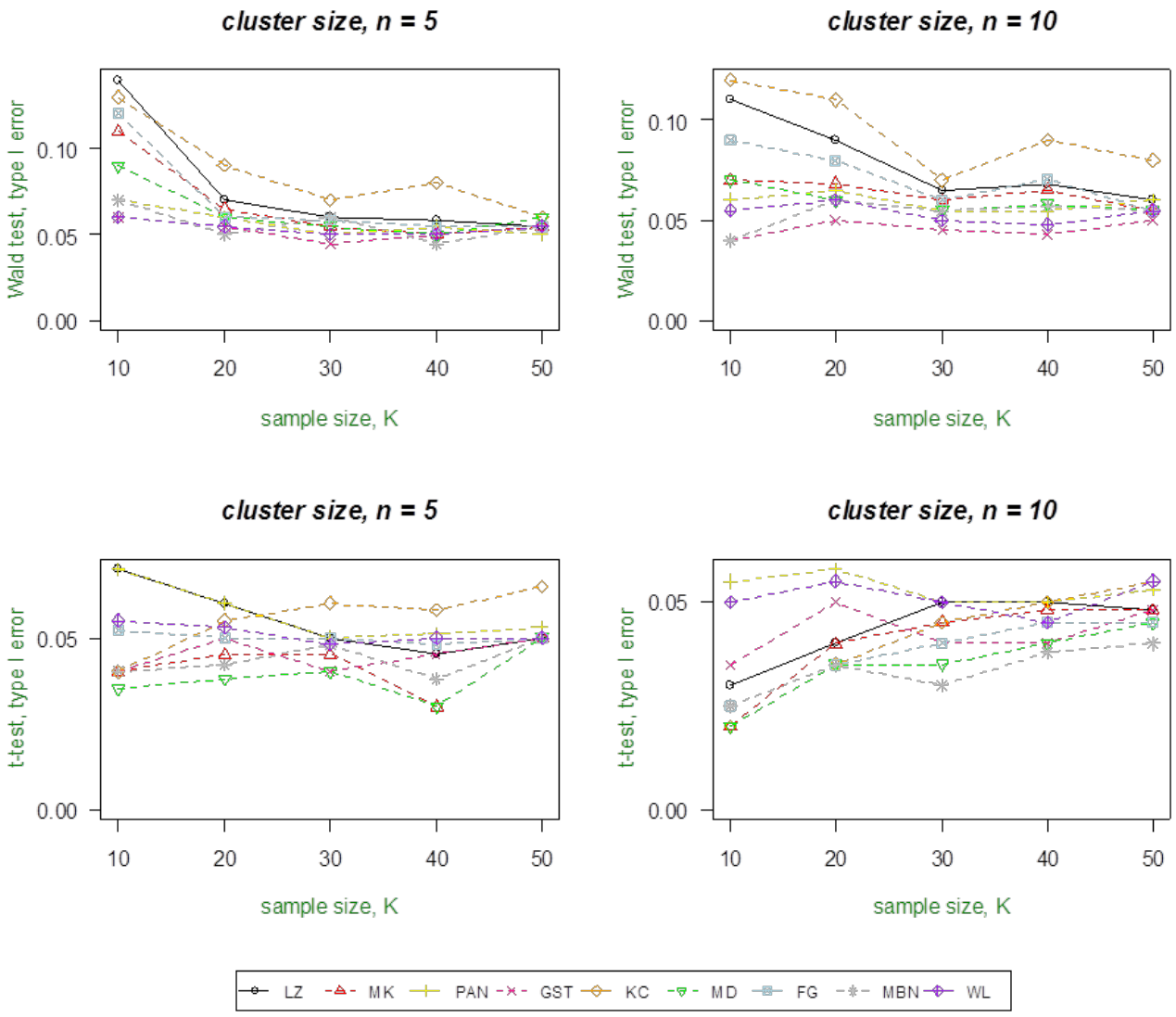

Figure 2: Type I errors based on Wald test and t-tests for **continuous outcomes** with the **AR-1** correlation structure.

https://doi.org/10.26219/heal.aueb.6669

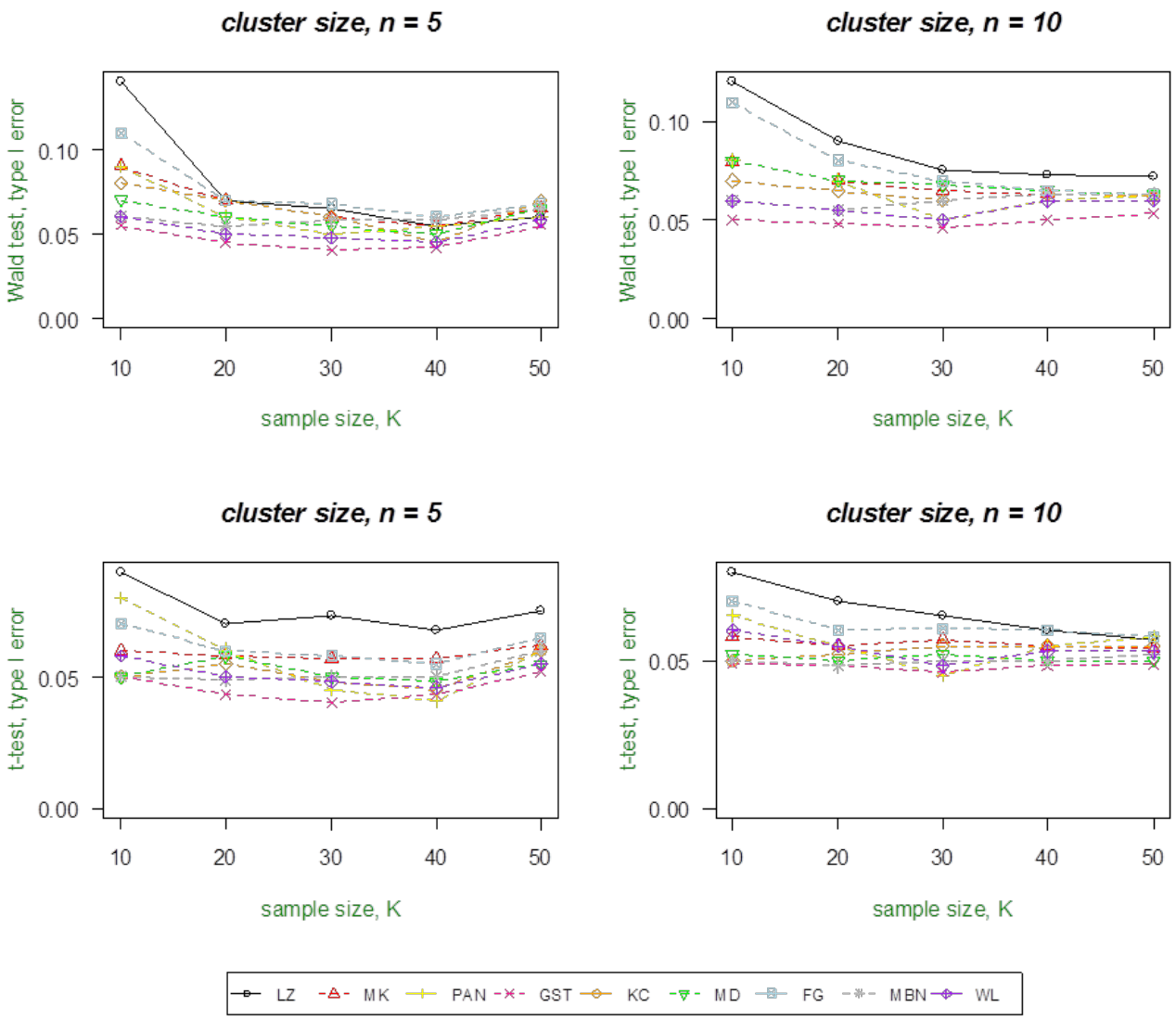

Figure 3: Type I errors based on Wald test and t-tests for **count outcomes** with the **exchangeable** correlation structure.

https://doi.org/10.26219/heal.aueb.6669

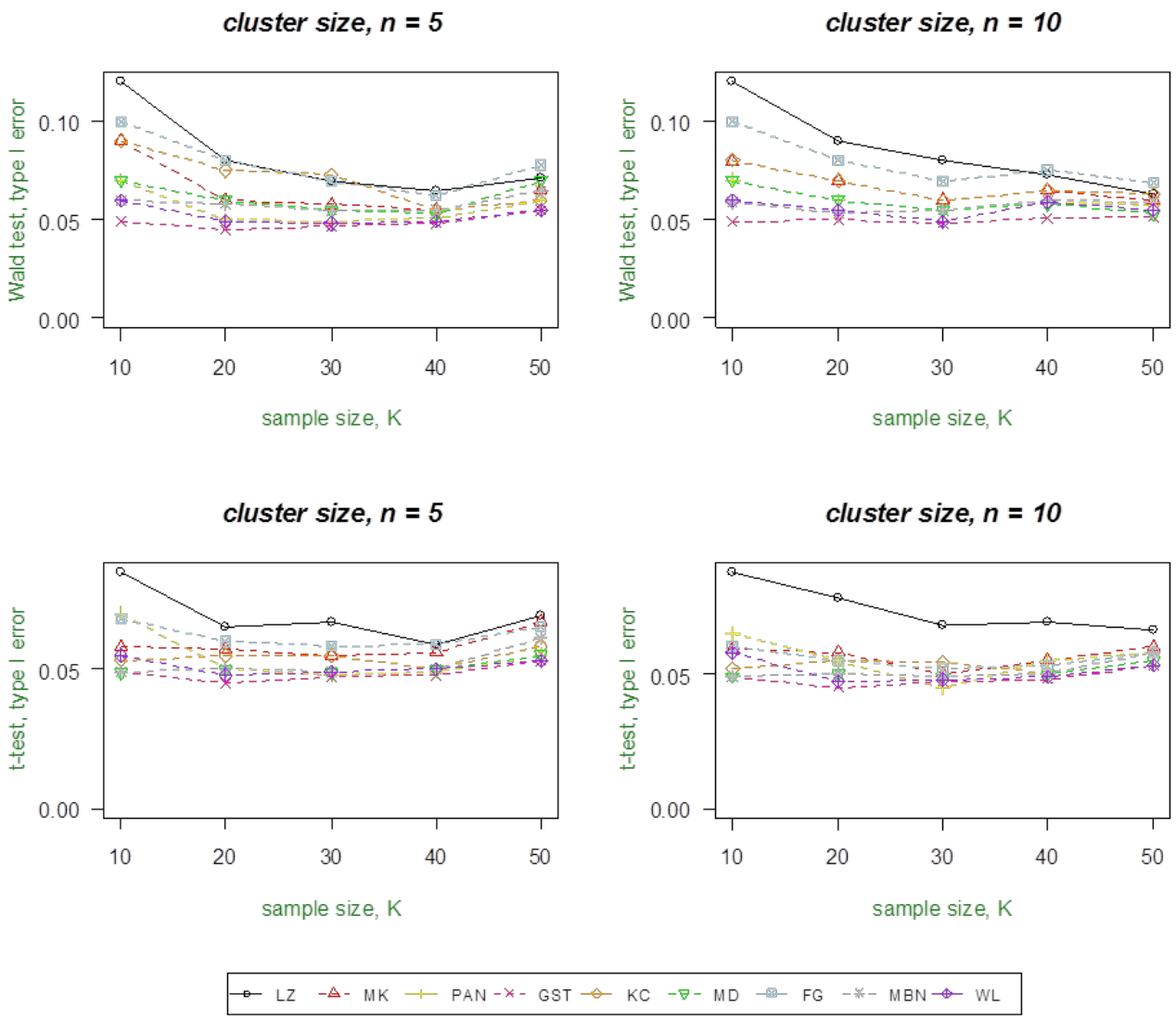

Figure 4: Type I errors based on Wald test and t-tests for **count outcomes** with the **AR-1** correlation structure.

https://doi.org/10.26219/heal.aueb.6669

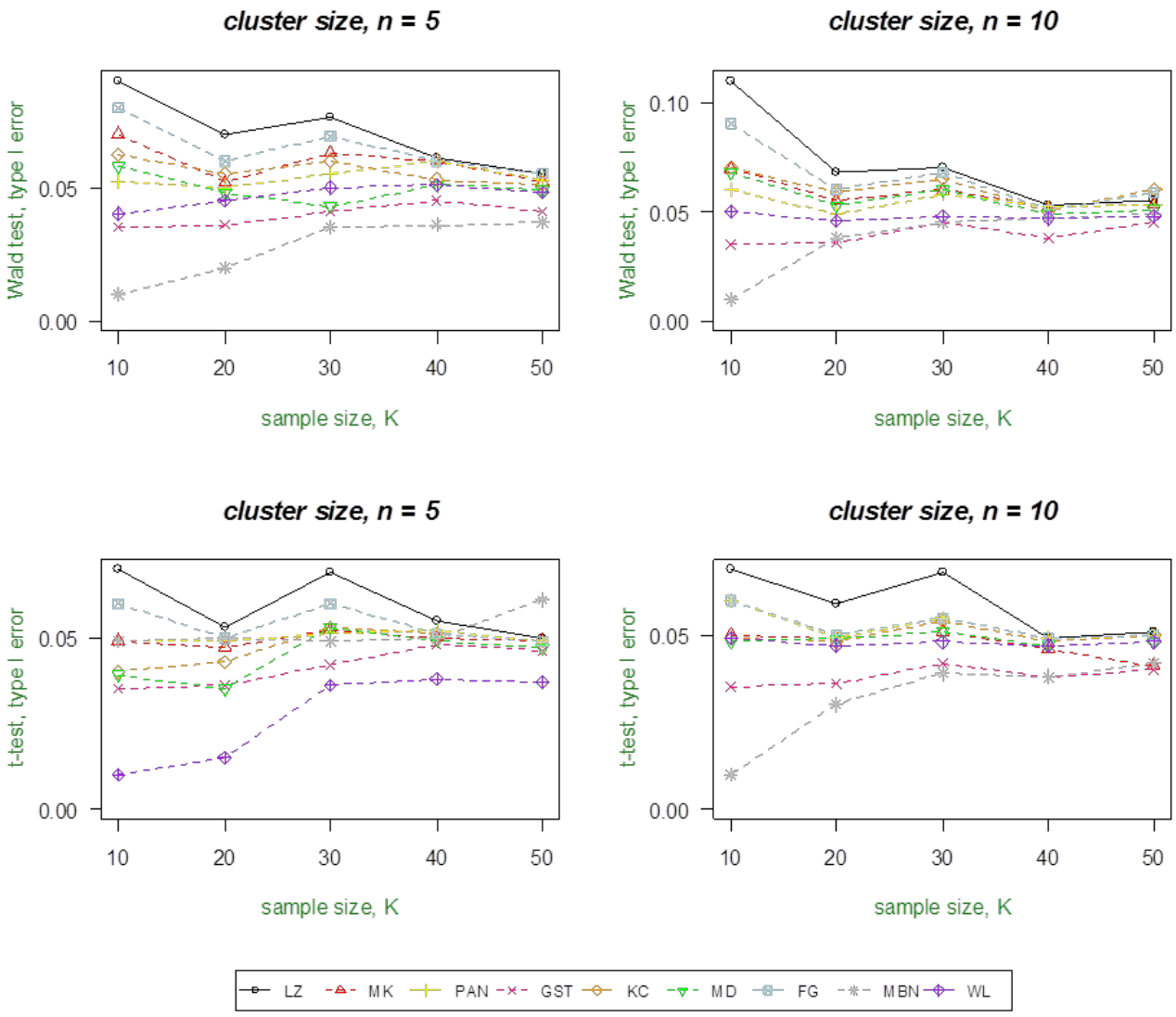

Figure 5: Type I errors based on Wald test and t-tests for **binary outcomes** with the **exchangeable** correlation structure.

https://doi.org/10.26219/heal.aueb.6669

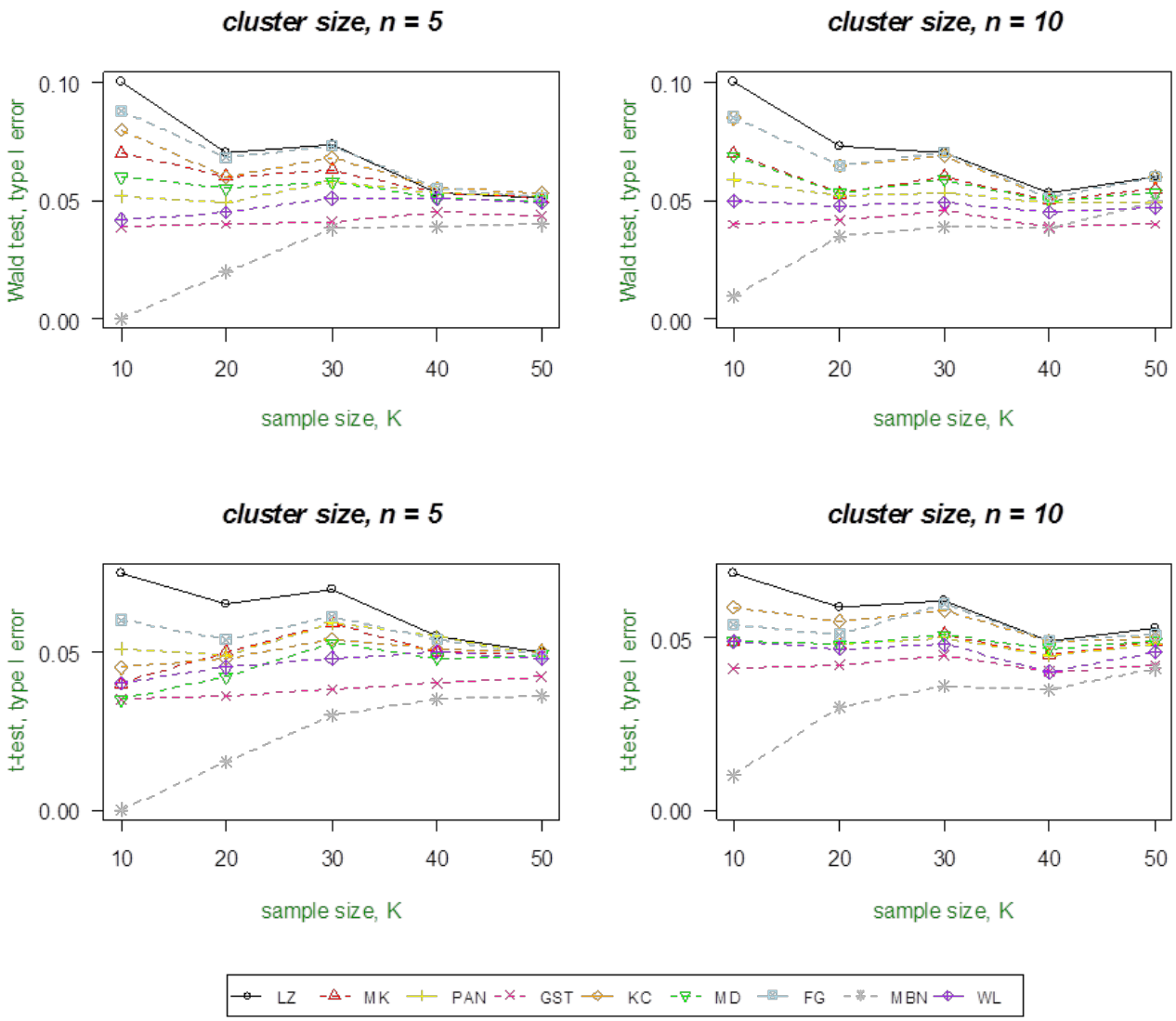

Figure 6: Type I errors based on Wald test and t-tests for **binary outcomes** with the **AR-1** correlation structure.

The most interesting results from our graphical study are described below:

(1) The results based on Wald test confirm our expectation and show that the V_{WL} variance estimator performs better than the others. The use of V_{LZ} robust variance estimator always results in inflated type I error, particularly when the sample size is small, i.e. ≤ 50 . Moreover, the other estimators lead also to inflated type I error but the degrees of freedom are smaller.

(2) The t-test for hypothesis testing performs better than the Wald test regarding the control of type I error across all estimators, while the V_{LZ} estimator still leads to patched type I error. However, V_{LZ} performs satisfactorily when the "working" correlation structure is specified correctly, even when the sample size is small (e.g. 10).

(3) As one can realize from the Figures, the sample size K plays a really important role to the performance of variance estimators for t-tests; the bigger the cluster size becomes, the more conservative results we gain.

(4) V_{KC} estimator attains worse performance than V_{LZ} based on Wald tests as indicated by greater inflation on type I error, but improves itself when the cluster size increases.

(5)There are some modified variance estimators that present a really conservative performance when the sample size is small, such as the estimators V_{GST} and V_{MBN}.

(6) Last but not least, V_{WL} performs better among all nine variance estimators, thus it is the most preferable estimator for the GEE methodology even when the sample size is as small as 10.

4 Data example

4.1 Background Study

Analyzing the growth curves of individuals over time or determining the effects of the continued administration of treatments over time are examples that longitudinal studies are required. One of the most widely known examples of growth curve analysis is that of Potthoff and Roy (1964) data set, which will be analytically discussed and alanyzed in this chapter. Their data consist of measurements obtained during a dental study from 11 girls and 16 boys at the ages of 8, 10, 12 and 14. The response measure is the distance between the pituitary and pterygomaxillary fissure for each child and the purpose of this study is to examine growth of this structure over time and to determine if there are significant differences between girls and boys. A simple approach to analyzing these data would be to conduct a two sample t-test between the measurements from the girls and the measurements from the boys. This approach, although easy to implement, would be invalid and would ignore the time effect in the data. This is because more than one observation from each individual would be included in the data, thereby violating the assumption of independent observations.

Under the assumption of having data that are normally distributed and continuous, one could perform multiple t-tests (Crowder & Hand, 1990; Davis, 2002). Therefore, t-tests would be performed between the measurements of the girls and boys at each occasion. The difficulty using this approach would be in deciding on an overall conclusion, since some of the tests may show significant differences and others may not, leading to the possibility of subjective conclusions. Alternatively, a t-test could be performed on the data from the final measurement occasion only, but this would result in a huge amount of data waste. In particular, this method would not allow for an analysis of growth trends.

To compare the measurements at different time points, paired t-tests could be performed between the data at two different ages. All possible paired combinations of ages could be considered. Because the test comparing *time 1* to *time 2* will be related to the test comparing *time 2* to *time 3* and *time 1* to *time 3*, these tests are not independent, and this can cause the probability of finding at least one test significant to

increase spuriously (Crowder & Hand, 1990; Davis, 2002).

Subject, gender and time could be included in an analysis of variance (ANOVA) approach to analyzing the data, resulting in the model $y_{ij} = \beta_0 + \beta_1\delta_i + \beta_{2i} + \beta_{3j} + \epsilon_{ij}$ where δ_i is an indicator for gender and β_{2i} , β_{3j} are adjustments to the mean response for the i^{th} individual and the j^{th} measurement occasion respectively, while the ϵ_{ij} is the error term. Alternatively, time can be included as a continuous covariate, changing this to an analysis of covariance (ANCOVA). Since subject is included in the mean structure of this model, this approach would imply that the subjects included were the only subjects of interest and inference could not be made beyond these individuals. It also does not allow for the inclusion of variability arising from the random sampling process, and therefore underestimates the variability in the data (Allison, 2005).

A different approach could be to summarise the vector of measurements for each individual into one summary measure (Crowder & Hand, 1990). For this method to be effective, a summary measure needs to be chosen that will adequately describe the subjects' data (Crowder & Hand, 1990; Davis, 2002). This method is referred to as response feature analysis. Examples of response features include the mean, maximum rate of increase, time to reach maximum rate of increase, half-life, or the slope of the least squares regression line. Then, the model simplifies to $y_i = \beta_0 + \beta_1\delta_i + \epsilon_i$, where the term y_i is the response feature and ϵ_i is the random error of the response feature for subject i.

These methods require the assumption that the variance of the derived response feature be homoscedastic. This would be violated if there are different numbers of observations being summarised for each individual, implying that this can only be achieved when there are no missing values and the number and sequence of measurements are the same for each individual (Fitzmaurice et al., 2004).

All of the methods discussed so far result in information loss and make very strong assumptions about the data, such as homogeneity of variance (Crowder & Hand, 1990; Fitzmaurice et al., 2004). None of these methods consider the covariance between repeated measures on the same individual, which may contain much information about the total response of an individual. Therefore, in order to take full advantage of the longitudinal study design, methods of analysis which explicitly include the covariance

between repeated measures should be used.

4.2 Problems related to using simple techniques

As it is mentioned above, there are many methods that can be used in longitudinal analysis. Despite the fact that these methods are really simple in use and can be useful for exploring data, one must be very careful as an overly simple analysis for repeated measurements may result in efficiency loss i.e. increasing the variability while not capitalising on the information available in the data, as well as biasing the results (Weiss, 2005).

Loss of efficiency can result from omitting subjects, e.g. because they contain missing data or from omitting observations in order to accommodate a certain method of analysis.

Bias can be introduced into the analysis in a number of ways, e.g. by means of inappropriate experimental designs, inappropriate analysis or leaving out subjects for reasons related to the study. If the design of a study leads to subjects being sampled so that the true sampled population is different to the intended population of interest, then the results of the analysis will be biased in favour of the subset of the population that was sampled. Therefore appropriate randomisation is important to avoid bias.

Moreover, if there exist groups with differences in a longitudinal study, the result can be the same using a simple statistical method. For instance, two groups that have different means may have the same slope over time or the slopes could be very different or in both cases the same difference in means may be found. Therefore simple analyses are very limited in the types of conclusions that can result in.

Alternatively, it is also very possible that two groups with very different responses over time can result in a non-significant result. For example, two groups may have the same average over time, but their slopes could be very different. Therefore these groups respond differently over time, but their averages do not convey this information (Weiss, 2005; Fitzmaurice et al., 2004). In that case, means of analysis such as repeated measures ANOVA is too restrictive in the compound symmetry assumption for the covariance structure, which assumes equal covariance between all repeated

measures, and can lead to overly conservative conclusions (Fitzmaurice et al., 2004).

Much of the loss of information resulting from overly simple methods of analysis is due to the disregard of the covariance between observations. Only by incorporating the covariance into the analysis is it possible to make predictions of the subjects' responses through time (Weiss, 2005).

4.3 Potthoff and Roy dataset analysis

In this section, we present the above results using one real data example in order to campare the finite performance of different variance estimators under small sample size. The dataset of Potthoff and Roy (1964) is a classic example of growth curve analysis. The data are related to a dental study of orthodontic measurements on children, which includes 11 girls and 16 boys repeatedly measured at the ages of 8, 10, 12 and 14. This study was conducted by researchers at the University of North Carolina Dental School. The response variable is the distance, calculated in millimeters, from the center of the pituitary to the pterygomaxillary fissure, while the covariates of interest are age (in years) and gender (male, female). Let y_{ij} , $i = 1, 2, ..., K$ and $j = 1, 2, ..., n_i$ denote the length between the pituitary and the pterygomaxillary fissure for the i^{th} individual at the j^{th} measurement occasion, where there are K individuals and n_i measurement occasions for the i^{th} individual ($n_i = 4$ for all individuals in this example).

The aim is to investigate if there exist statistically significant gender differences in dental growth measurements and their trends as age increases.

In particular, we are interested in testing the following hypothesis $H_0 : F_8 = F_{10} = F_{12} = F_{14}$ of no time effect, where F_s denotes the marginal distribution of the distances at age s. As recommended for any statistical analysis, we begin by plotting the data in order to understand the distribution of the data for each age group. The most important relationship to plot for longitudinal data on multiple subjects is the trend of the response over time by subject.

The box plots of Figure 7 show the minimum, first quartile, median, third quartile, and the maximum distance measured for each time point separately. They indicate that the measured distances have a skewed distribution (especially as the age

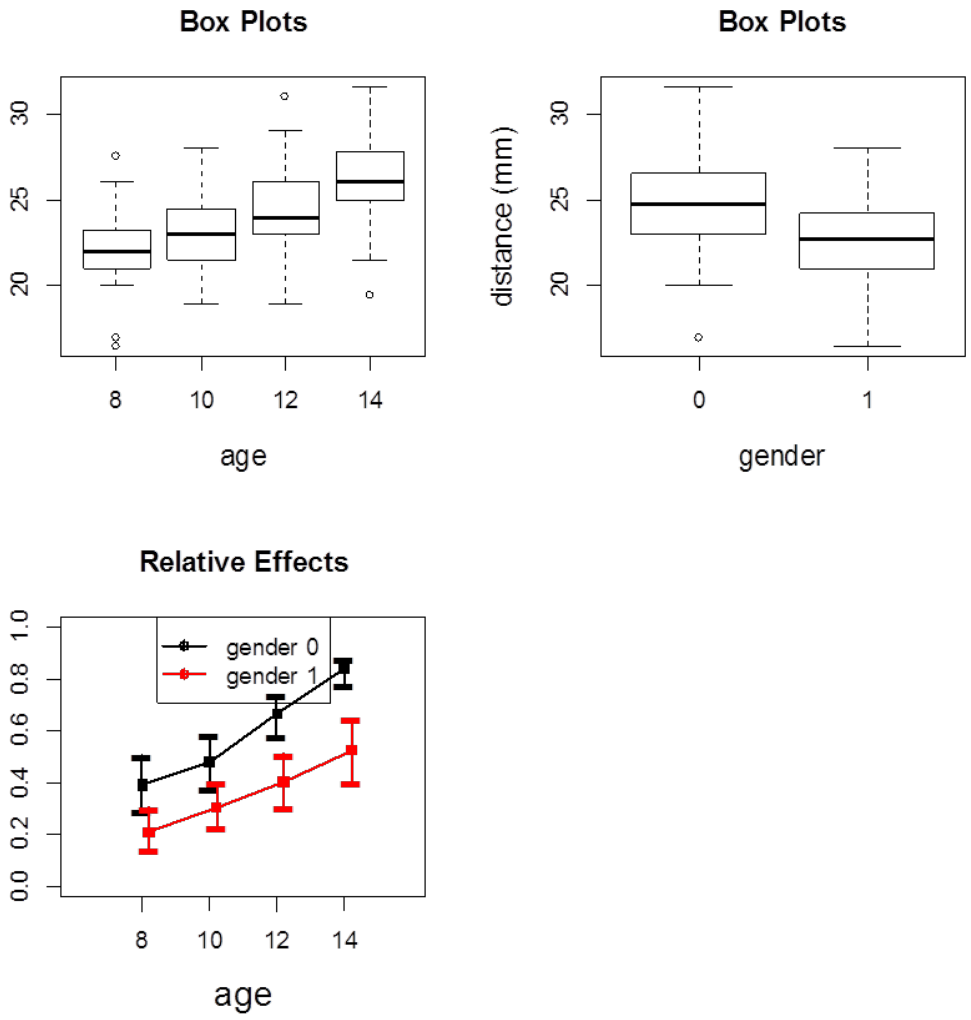

Figure 7: Box plots and 95% confidence intervals for parameters in the dental study.

increases). The increase in median gives rise to a time effect. The 95% confidence intervals at the bottom of Figure 7 present the lower bound, point estimate, and the upper bound for each time point separately. The point estimates increase, meaning the older the children, the larger the observed distances between pituitary and the pterygomaxillary fissure.

```
> summary(f1np)
Model:
F1 LD F1 Model

Call:
distance ~ age + gender

Relative Treatment Effect (RTE):
               RankMeans  Nobs  RTE
gender0        64.79688   64 0.5953414
gender1        39.52273   44 0.3613215
age8           32.98295   27 0.3007681
age10          43.05966   27 0.3940709
age12          58.32812   27 0.5354456
age14          74.26847   27 0.6830414
gender0:age8   42.87500   16 0.3923611
gender0:age10  52.43750   16 0.4809028
gender0:age12  72.65625   16 0.6681134
gender0:age14  91.21875   16 0.8399884
gender1:age8   23.09091   11 0.2091751
gender1:age10  33.68182   11 0.3072391
gender1:age12  44.00000   11 0.4027778
gender1:age14  57.31818   11 0.5260943

Wald-Type Statistc (WTS):
            Statistic  df  p-value
```

```
gender       8.797738  1 3.016043e-03
age        103.424543  3 2.851266e-22
gender:age   4.676974  3 1.970375e-01

ANOVA-Type Statistc (ATS):
            Statistic df      p-value
gender      8.797738 1.00000 3.016043e-03
age        46.191394 2.55914 7.475954e-26
gender:age  1.872467 2.55914 1.412992e-01

Modified ANOVA-Type Statistic for the Whole-Plot Factors:
        Statistic df1 df2     p-value
gender  8.797738   1 17.57258 0.008431029
```

Considering the above summary for each age group s, the rank mean of the overall ranks (RankMeans), the number of observations (Nobs) and the point estimate $\hat{p}_s$ of the relative treatment effect (RTE) are displayed. The obtained result of 0.30 for the age group 8 (time8) can be interpreted, for example, as follows: a randomly chosen observation from the whole dataset results in a smaller value than a randomly chosen observation from the age group 8 with an estimated probability of 30%. Further, since $\hat{p}_8 < \hat{p}_{10} < \hat{p}_{12} < \hat{p}_{14}$, the observations from the age group 8 tend to result in smaller values than those from the age group 10 which, in return, also tends to result in smaller values than the measurements from the age groups 12 and 14, respectively. Thus, an increase in the effect seems to indicate the increase in the measured distances. To test the hypothesis H_O of no time effect, Wald-Type Statistic (WTS) and Anova-Type Statistic (ATS) can be applied, which are also displayed in the output of the model summary. The column degrees of freedom (df) for ATS is the numerator degrees of freedom of the F distribution as the denominator degrees of freedom is set to infinity. Both WTS and ATS yield highly statistically significant p-values of 2.85×10^{-22} and 7.48×10^{-26}, respectively, indicating that the null hypothesis of no time effect is to be rejected. To investigate the question about which of the four distribution functions differ, we can

apply multiple comparisons with the Bonferroni adjustment as described below:

Table 4: Multiple comparisons against the control in the dental study with Bonferroni adjustment.

Comparison	Hypothesis	p-value	Adjusted p-value
Time 8 vs Time 10	$H_0 : F_8 = F_{10}$	0.2204	0.6612
Time 8 vs Time 10	$H_0 : F_8 = F_{12}$	<0.0001	<0.0001
Time 8 vs Time 10	$H_0 : F_8 = F_{14}$	<0.0001	<0.0001

The results are presented in Table 3, where, for brevity, only the p-values obtained from ATS are reported. In Table 3, the Bonferroni-adjusted p-value of 0.6612, obtained for testing the age group 8 against the age group 10 (Time 8 vs. Time 10), is calculated by multiplying the original p-value of 0.2204 by 3. Similar calculations are also performed for the other pairwise comparisons. From the results, we can conclude that the distance between the center of the pituitary and the pterygomaxillary fissure significantly increases over time by observing the p-values of $\ll 0.0001$ from both WTS and ATS. In addition, we notice significant differences between the distributions of the measured distances for the age groups 8 and 12 and age groups 8 and 14, respectively. To compare the obtained results and conclusions with parametric methods, we further reanalyze the data with the lme() function in the R package nlme (Pinheiro et al. 2012).

We obtain an overall significant time effect (p-value $\ll 0.0001$). Regarding the multiple comparisons against age group 8 and multiplying the original p-value by 3, we obtain the adjusted p-value of 0.4395 for the comparison "Time 8 vs. Time 10", as well as the p-values of 0.0009 and $\ll 0.0001$ for "Time 8 vs. Time 12" and "Time 8 vs. Time 14", respectively. Thus, both parametric and nonparametric procedures result in similar conclusions in this example, which is not surprising since the data exhibit only a minor degree of skewness as indicated by the box plots.

We continue our analysis with the following graph:

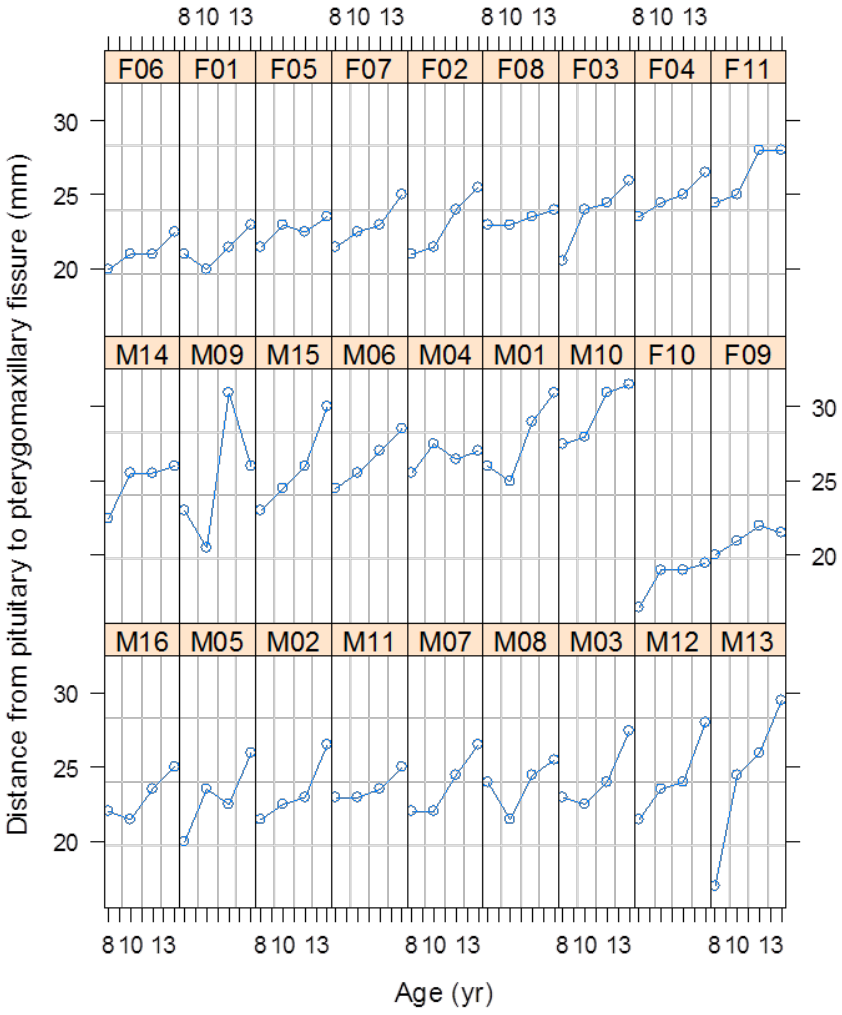

Figure 8: A lattice plot (groupedData) of the average distance (mm) versus age (years) by subject for the Potthoff and Roy data set.

In this plot in which the data for different subjects are shown in separate panels with the axes held constant for all the panels, allows for examination of the time-trends within subjects and for comparison of these patterns between subjects. Through the use of small panels in a repeating pattern Figure 8 conveys a great deal of information, the individual time trends for 27 subjects all of them being examined at the age of 8, 10, 12 and 14 years.

As stated above, all the panels have the same vertical and horizontal scales, allowing us to evaluate the pattern over time for each subject and also to compare patterns between subjects. It is provided to enhance our ability to discern patterns in both the slope (the typical change in distance per year of examination for that particular subject) and the intercept (the average distance for the subject).

The aspect ratio of the panels (ratio of the height to the width) has been chosen, according to an algorithm described in Cleveland (1993), to facilitate comparison of slopes. The panels have been ordered (from left to right starting at the bottom row) by increasing intercept. Because the subject identifiers, shown in the strip above each panel, are unrelated to the response it would not be helpful to use the default ordering of the panels, which is by increasing subject number. If we did so our perception of patterns in the data would be confused by the, essentially random, ordering of the panels. Instead we use a characteristic of the data to determine the ordering of the panels, thereby enhancing our ability to compare across panels. For example, a question of interest to the experimenters is whether a subject's rate of change in distance is related to the subject's initial distance. If this was the case we would expect that the slopes would show an increasing trend (or, less likely, a decreasing trend) in the left to right, bottom to top ordering.

There is little evidence in Figure 8 of such a systematic relationship between the subject's initial distance and their rate of change in distance per year of measurement. We do see that for each subject, the distance increases, more-or-less linearly, with the increase of the age. However, there is considerable variation both in the initial distance and in the annual rate of increase in distance. We can also see that these data are balanced, both with respect to the number of observations on each subject, and with respect to the times at which these observations were taken. This can be confirmed by

cross-tabulating subject and years.

In cases like this where there are several observations (4) per subject and a relatively simple within-subject pattern (more-or-less linear) we may want to examine coefficients from within-subject fixed-effects fits. However, because the subjects constitute a sample from the population of interest and we wish to drawn conclusions about typical patterns in the population and the subject-to-subject variability of these patterns, we will eventually want to fit a model.

We proceed our analysis by presenting some more graphs in order to provide a further exploration of our data and then we fit the model using the R "gee" package.

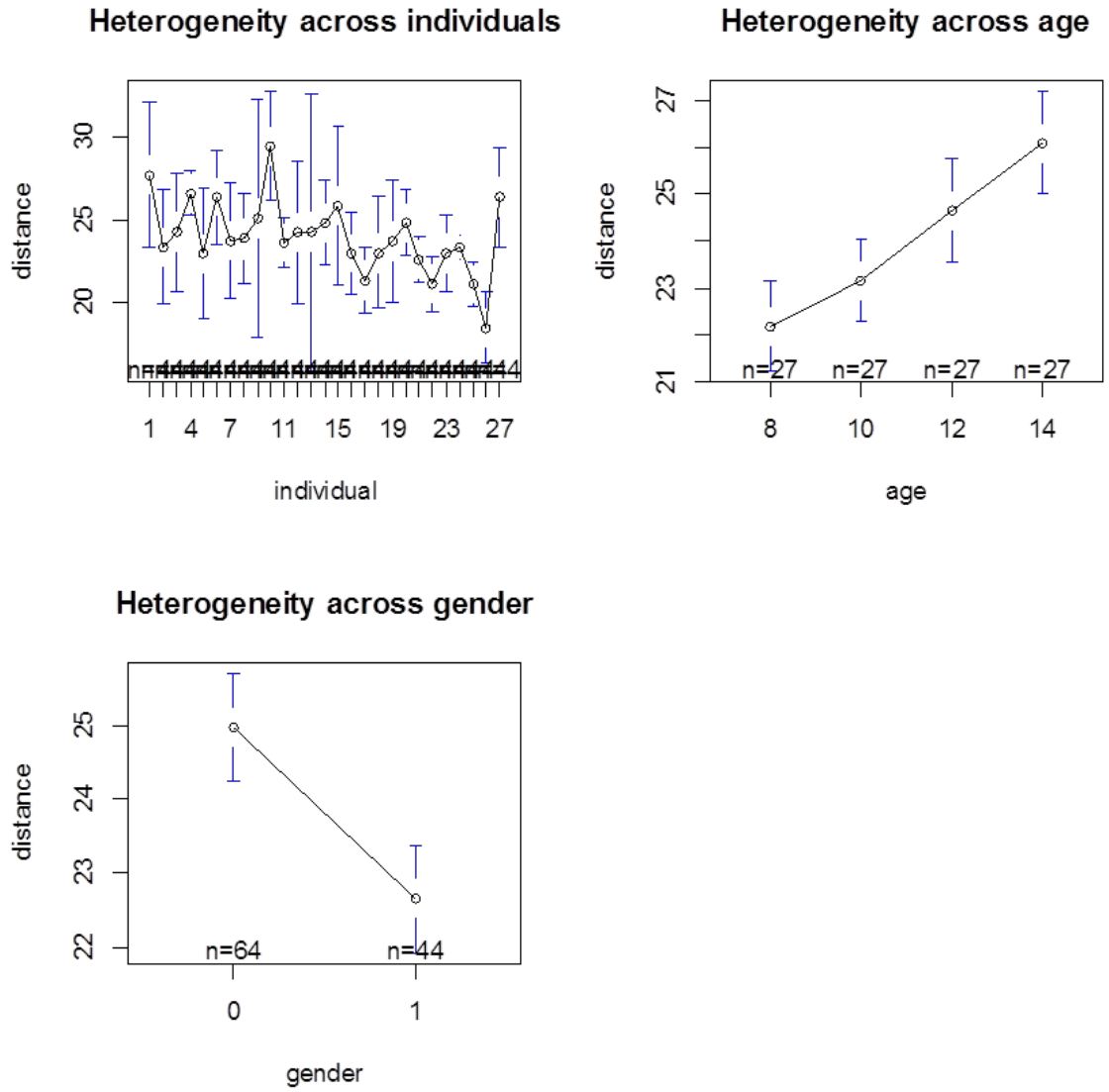

Figure 9: Plots which indicate the heterogeneity across individuals, across age and across gender for the Potthoff and Roy dataset.

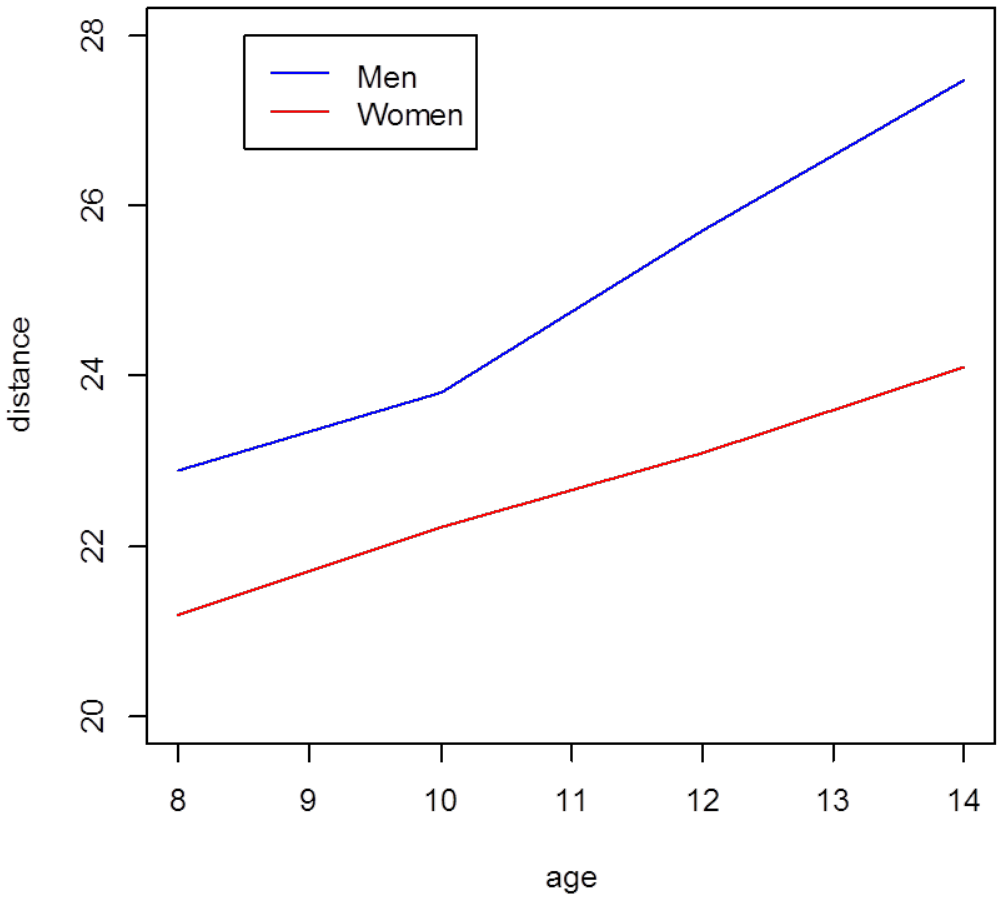

Figure 10: Plot which indicates the relationship between the distance and the age for Men and Women for the Potthoff and Roy data set.

From the above plot one can observe that men have higher measurements of the distance between the pituitary and the pterygomaxillary fissure compared to women. Additionally, mens' slope presents a sharper increase after the age of 10 while the distance of both genders increases as age grows up.

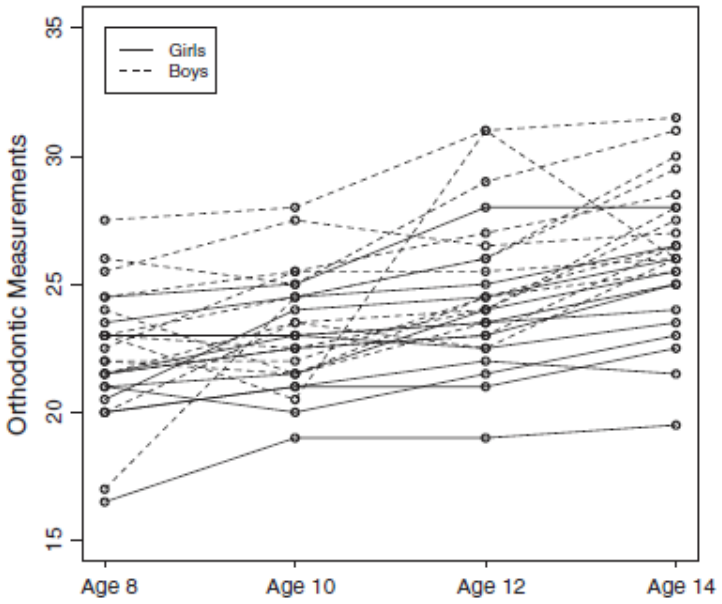

Figure 11: Orthodontic measurements by subject over time.

The scatter plot of orthodontic measurements is shown in Figure 11. One can notice that the boys have higher measurements than the girls on average and the measurements tend to increase with age.

The same results are also shown in Table 4.

Table 5: The mean Distance for Men and Women.

Age	Distance (Men)	Distance (Women)
8	22.87500	21.18182
10	23.81250	22.22727
12	25.71875	23.09091
14	27.46875	24.09091

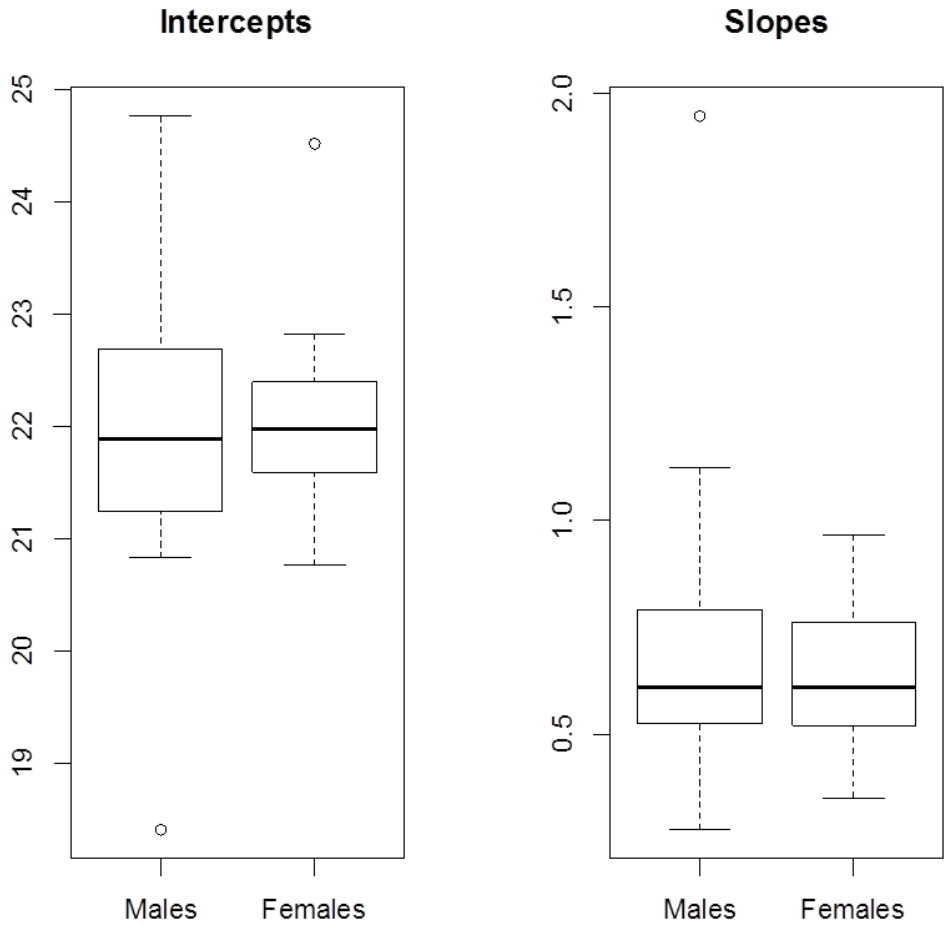

Figure 12: Boxplots for Intercepts and Slopes for Males and Females of the Potthoff and Roy data set.

From the above boxplots one can observe that both, intercept and slope, present bigger variance in men compared to women but they have almost the same mean, which is approximately near to 22 for the intercept and 0.7 for the slope.

After examining and exploring the data set of orthodontic measurements on children, we continue with the application of the model. The outcome variable of interest is dental growth measurements of the distance (in millimeters) from the center of the pituitary gland to the pterygomaxillary fissure, which was repeatedly measured at ages 8, 10, 12 and 14 for each child. Age (in years) and gender (female or male) are the primary covariates of interest. As the distribution of age was skewed, a square-root

transformation yielded a distribution closer to the normal. The mean model took the following form:

$$E(y) = \beta_0 + \beta_1 \times \sqrt{age} + \beta_2 \times gender$$

We fitted the above model and we estimated the regression parameters and their variance using the nine variance-estimators discussed in Chapter 2. We used the complete data set of 27 subjects in order to perform the hypothesis testing. The results are provided in Table 5.

Table 6: Parameter and variance estimates for case study on orthodontic measurements.

	$\hat{\beta}$	V_{LZ}	V_{MK}	V_{PAN}	V_{GST}	V_{KC}	V_{MD}	V_{FG}	V_{MBN}	V_{WL}
Independence										
Complete										
Interc.	6.077	3.462	3.894	3.704	4.167	3.675	3.905	15.593	4.397	4.127
$\sqrt{age}$	4.319	0.213	0.239	0.213	0.239	0.221	0.229	0.236	0.329	0.229
gender	2.321	0.562	0.632	0.537	0.604	0.612	0.666	0.699*	0.645	0.629
Exchangeable										
Complete										
Interc.	6.077	3.462	3.894	3.704	4.167	3.675	3.905	6.022	4.121	4.127
$\sqrt{age}$	4.319	0.213	0.239	0.213	0.239	0.221	0.229	0.220	0.248	0.229
gender	2.321	0.562	0.632	0.537	0.604	0.612	0.666	0.699**	0.669	0.629
AR1										
Complete										
Interc.	5.999	3.689	4.150	3.994	4.493	3.582	4.160	8.476	4.558	4.445
$\sqrt{age}$	4.249	0.230	0.259	0.230	0.259	0.223	0.249	0.276	0.289	0.249
gender	2.410	0.569	0.640	0.539	0.605	0.613	0.674	0.707*	0.661	0.632
Unstructured										
Complete										
Interc.	5.999	3.300	3.712	3.676	4.136	3.313	3.722	5.925	3.961	4.093
$\sqrt{age}$	4.270	0.217	0.245	0.217	0.245	0.266	0.234	0.224	0.257	0.234
gender	2.220	0.533	0.599**	0.508	0.572	0.613**	0.632**	0.663*	0.630**	0.596

*Not significant on either test at the significance level of 0.01.

**Significant based only on Wald tests at the significance level of 0.01.

The results we obtained are consistent with our findings from the simulation study. More specifically, both Wald-test and t-tests with the significance levels of 0.01 and 0.05 are applied for hypotheses testing.

(1) All nine variance estimators provide comparable results on hypotheses testing of $\sqrt{age}$ with the Wald tests.

(2) t-tests at the significance level of 0.01 provide different conclusions for gender.Thus, the choice of the small sample adjustment is significant for the statistical results.

(3) All covariate estimates are statistically significant at a level of $a = 0.01$ as well as $a = 0.05$ using Wald test or t-test except those with the marks.

5 Conclusions and discussions

In this Master thesis, we analytically presented the theory concerning Generalized Estimating Equations as long as the theory of marginal models and mixed effects models. We continued by presenting the robust "sandwich" variance estimator and the eight most recent variance modifications for GEE in order to improve the sample properties especially in the case of small sample size. We implemented one simulation study for three different types of response variables (continuous, count and binary) and we confirmed our results using the very known dataset of Potthoff and Roy for orthodontic measurements between 16 boys and 11 girls. The "geesmv" R package was proved really useful in our numerical study. In addition, we emphasized two important types of hypothesis testing for GEE, especially when the sample size is small, Wald test and t-test. The simulation study showed that t-tests based on the variance estimator V_{WL} perform well.

Despite the fact that there is a great range of bibliography about the recent developments that concern several modified variance estimators, there is still plenty of space in order to develop methods about improving the efficiency and the robustness of parameter estimates. Moreover, our simulation analysis based on equal cluster sizes, so a very interesting task for the future would be to discover how the simulation study would be without this limitation. Additionally, greater emphasis could be given on other issues, such as evaluating the type II error or selecting the appropriate model or even handling the missing data under the condition of small sample size.

Another issue that could be really interesting and challenging in addition to modified variance estimators and test statistics is the power analysis (Shih WJ, 1997). Shih relied on Wald tests using the estimates of regression parameters and robust variance estimators in order to provide the power calculations. However, these calculations have two important and necessary conditions that must be fulfilled; (1) the $V(\hat{b})$ has to be unbiased and (2) asymptotic normality has to be satisfied. However, when the sample size is small, the estimated power tends to be overestimated. Thus, a modification on the power estimation must be applied incorporating the variance estimators which were discussed in Chapter 2 for improving the efficiency.

Appendix

1.Code used for the geesmv package in R and the 9 variance estimators for the Capter 3.

```
### Get necessary information (i.e., the number of clusters, cluster sizes)
### of the data set.
cluster.size(individual)

### 1 (Fay and Graubard, 2001)
data_alt <- reshape(dental, direction="long", timevar="Time",
varying=names(dental)[3:6], v.names="response", times=c(8,10,12,14))
data_alt <- data_alt[order(data_alt$subject),]
data_alt$gender <- as.numeric(data_alt$gender)
data_alt$Time <- sqrt(data_alt$Time)

formula <- response~Time+gender
fg.ind <- GEE.var.fg(formula,id="subject",family=gaussian,
data_alt,corstr="independence") ##Independence correlation structure;
fg.exch <- GEE.var.fg(formula,id="subject",family=gaussian,
data_alt,corstr="exchangeable") ##Exchangeable correlation structure;
fg.ar1 <- GEE.var.fg(formula,id="subject",family=gaussian,
data_alt,corstr="AR-M") ##AR-1 correlation structure;
fg.unstr <- GEE.var.fg(formula,id="subject",family=gaussian,
data_alt,corstr="unstructured") ##Unstructured correlation structure;
fg.ind
fg.exch
fg.ar1
```

```
fg.unstr


### 2 (Gosho et al., 2014)
formula <- response~Time+gender
gst.ind <- GEE.var.gst(formula,id="subject",family=gaussian,data_alt,corstr="independ
gst.exch <- GEE.var.gst(formula,id="subject",family=gaussian,
data_alt,corstr="exchangeable") ##Exchangeable correlation structure;
gst.ar1 <- GEE.var.gst(formula,id="subject",family=gaussian,
data_alt,corstr="AR-M") ##AR-1 correlation structure;
gst.unstr <- GEE.var.gst(formula,id="subject",family=gaussian,
data_alt,corstr="unstructured") ##Unstructured correlation structure;
gst.ind
gst.exch
gst.ar1
gst.unstr


### 3 (Kauermann and Carroll, 2001)

formula <- response~Time+gender
kc.ind <- GEE.var.kc(formula,id="subject",family=gaussian,
data_alt,corstr="independence") ##Independence correlation structure;
kc.exch <- GEE.var.kc(formula,id="subject",family=gaussian,
data_alt,corstr="exchangeable") ##Exchangeable correlation structure;
kc.ar1 <- GEE.var.kc(formula,id="subject",family=gaussian,
data_alt,corstr="AR-M") ##AR-1 correlation structure;
kc.unstr <- GEE.var.kc(formula,id="subject",family=gaussian,
data_alt,corstr="unstructured") ##Unstructured correlation structure;
kc.ind
kc.exch
```


```
kc.ar1
kc.unstr


### 4 (Liang and Zeger, 1986)
formula <- response~Time+gender
lz.ind <- GEE.var.lz(formula,id="subject",family=gaussian,
data_alt,corstr="independence") ##Independence correlation structure;
lz.exch <- GEE.var.lz(formula,id="subject",family=gaussian,
data_alt,corstr="exchangeable") ##Exchangeable correlation structure;
lz.ar1 <- GEE.var.lz(formula,id="subject",family=gaussian,
data_alt,corstr="AR-M") ##AR-1 correlation structure;
lz.unstr <- GEE.var.lz(formula,id="subject",family=gaussian,
data_alt,corstr="unstructured") ##Unstructured correlation structure;
lz.ind
lz.exch
lz.ar1
lz.unstr


### 5 (Morel, Bokossa and Neerchal, 2003)
formula <- response~Time+gender
mbn.ind <- GEE.var.mbn(formula,id="subject",family=gaussian,
data_alt,corstr="independence",d=2,r=1) ##Independence correlation structure;
mbn.exch <- GEE.var.mbn(formula,id="subject",family=gaussian,
data_alt,corstr="exchangeable",d=2,r=1) ##Exchangeable correlation structure;
mbn.ar1 <- GEE.var.mbn(formula,id="subject",family=gaussian,
data_alt,corstr="AR-M",d=2,r=1) ##AR-1 correlation structure;
mbn.unstr <- GEE.var.mbn(formula,id="subject",family=gaussian,
data_alt,corstr="unstructured",d=2,r=1) ##Unstructured correlation structur;
mbn.ind
```


```
mbn.exch
mbn.ar1
mbn.unstr


### 6 (Mancl and DeRouen, 2001)
formula <- response~Time+gender
md.ind <- GEE.var.md(formula,id="subject",family=gaussian,
data_alt,corstr="independence") ##Independence correlation structure;
md.exch <- GEE.var.md(formula,id="subject",family=gaussian,
data_alt,corstr="exchangeable") ##Exchangeable correlation structure;
md.ar1 <- GEE.var.md(formula,id="subject",family=gaussian,
data_alt,corstr="AR-M") ##AR-1 correlation structure;
md.unstr <- GEE.var.md(formula,id="subject",family=gaussian,
data_alt,corstr="unstructured") ##Unstructured correlation structure;
md.ind
md.exch
md.ar1
md.unstr


### 7 (Mackinnon, 1985)
formula <- response~Time+gender
mk.ind <- GEE.var.mk(formula,id="subject",family=gaussian,
data_alt,corstr="independence") ##Independence correlation structure;
mk.exch <- GEE.var.mk(formula,id="subject",family=gaussian,
data_alt,corstr="exchangeable") ##Exchangeable correlation structure;
mk.ar1 <- GEE.var.mk(formula,id="subject",family=gaussian,
data_alt,corstr="AR-M") ##AR-1 correlation structure;
mk.unstr <- GEE.var.mk(formula,id="subject",family=gaussian,
data_alt,corstr="unstructured") ##Unstructured correlation structure;
```


```
mk.ind
mk.exch
mk.ar1
mk.unstr


### 8 (Pan, 2001)
formula <- response~Time+gender
pan.ind <- GEE.var.pan(formula,id="subject",family=gaussian,
data_alt,corstr="independence") ##Independence correlation structure;
pan.exch <- GEE.var.pan(formula,id="subject",family=gaussian,
data_alt,corstr="exchangeable") ##Exchangeable correlation structure;
pan.ar1 <- GEE.var.pan(formula,id="subject",family=gaussian,
data_alt,corstr="AR-M") ##AR-1 correlation structure;
pan.unstr <- GEE.var.pan(formula,id="subject",family=gaussian,
data_alt,corstr="unstructured") ##Unstructured correlation structure;
pan.ind
pan.exch
pan.ar1
pan.unstr


### 9 (Wang and Long, 2011)
formula <- response~Time+gender
wl.ind <- GEE.var.wl(formula,id="subject",family=gaussian,
data_alt,corstr="independence") ##Independence correlation structure;
wl.exch <- GEE.var.wl(formula,id="subject",family=gaussian,
data_alt,corstr="exchangeable") ##Exchangeable correlation structure;
wl.ar1 <- GEE.var.wl(formula,id="subject",family=gaussian,
data_alt,corstr="AR-M") ##AR-1 correlation structure;
wl.unstr <- GEE.var.wl(formula,id="subject",family=gaussian,
```

```
data_alt,corstr="unstructured") ##Unstructured correlation structure;
wl.ind
wl.exch
wl.ar1
wl.unstr
```

2.Code used for the boxplots in Chapter 4.

```
library("nparLD")
par(mfrow=c(2,2))
boxplot(distance ~ age, data = mydata, lwd = 1, xlab = "age",
font.lab = 1.2, cex.lab = 1.2, main = "Box Plots")

boxplot(distance ~ gender , data = mydata, lwd = 1, xlab = "gender",
ylab = "distance (mm)", font.lab = 1.2, cex.lab = 1.2, main = "Box Plots")

f1np <- nparLD(distance ~ age + gender, data = mydata, subject = "individual",
description = FALSE)
plot(f1np)

### more information
f1np <- nparLD(distance ~ age + gender, data = mydata, subject = "individual",
description = TRUE)
plot(f1np)
```

3.Code used for exploring the data set.

```
coplot(distance ~ age|individual, type="b", data=mydata)  ###points and lines

par(mfrow=c(2,2))
plotmeans(distance ~ individual, main="Heterogeneity across individuals", data=mydata
### plotmeans draws a 95% confidence interval around the means
```

```
plotmeans(distance ~ age, main="Heterogeneity across age", data=mydata)
plotmeans(distance ~ gender, main="Heterogeneity across gender", data=mydata)
```

4.Code used for analyzing the data set (gee).

```
mydata_order<- order(as.integer(mydata$individual))
mydata1 <- mydata[mydata_order,]
mydata1

fit.gee1 <- gee(distance ~ age + gender + age:gender, id=individual, family=gaussian,
corstr="independence", data=mydata1)
summary(fit.gee1)


>summary(fit.gee1)

GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link:                      Identity
Variance to Mean Relation: Gaussian
Correlation Structure:     Independent

Call:
gee(formula = distance ~ age + gender + age:gender, id = individual,
data = mydata1, family = gaussian, corstr = "independence")

Summary of Residuals:
Min          1Q         Median      3Q        Max
```

```
-5.6156250 -1.3218750 -0.1681818  1.3299006  5.2468750


Coefficients:
             Estimate   Naive S.E. Naive z   Robust S.E. Robust z
(Intercept) 16.3406250  1.4162242 11.538163  1.17148092 13.9486906
age          0.7843750  0.1261673  6.216945  0.09834755  7.9755416
gender       1.0321023  2.2187969  0.465163  1.37778506  0.7491025
age:gender  -0.3048295  0.1976661 -1.542143  0.11686730 -2.6083390

Estimated Scale Parameter:  5.093818
Number of Iterations:  1


Working Correlation
      [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
>

coef(summary(fit.gee1))
##### get the P values using a normal approximation for the distribution of z

> 2 * pnorm(abs(coef(summary(fit.gee1))[,5]), lower.tail = FALSE)
(Intercept)          age       gender   age:gender
3.204341e-44 1.517141e-15 4.537954e-01 9.098279e-03


fit.gee2 <- gee(distance ~ age + gender + age:gender, id=individual, family=gaussian,
```

```
corstr="exchangeable", data=mydata1)
summary(fit.gee2)


> summary(fit.gee2)

GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link:                      Identity
Variance to Mean Relation: Gaussian
Correlation Structure:     Exchangeable

Call:
gee(formula = distance ~ age + gender + age:gender, id = individual,
data = mydata1, family = gaussian, corstr = "exchangeable")

Summary of Residuals:
Min          1Q     Median          3Q         Max
-5.6156250 -1.3218750 -0.1681818  1.3299006  5.2468750


Coefficients:
             Estimate  Naive S.E.  Naive z    Robust S.E.  Robust z
(Intercept) 16.3406250 0.98813100 16.5369015  1.17148092 13.9486906
age          0.7843750 0.07879034  9.9552182  0.09834755  7.9755416
gender       1.0321023 1.54810375  0.6666881  1.37778506  0.7491025
age:gender  -0.3048295 0.12344073 -2.4694405  0.11686730 -2.6083390

Estimated Scale Parameter:  5.093818
```

```
Number of Iterations:  1


Working Correlation
       [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.6100109 0.6100109 0.6100109
[2,] 0.6100109 1.0000000 0.6100109 0.6100109
[3,] 0.6100109 0.6100109 1.0000000 0.6100109
[4,] 0.6100109 0.6100109 0.6100109 1.0000000
>


coef(summary(fit.gee2))

> 2 * pnorm(abs(coef(summary(fit.gee2))[,5]), lower.tail = FALSE)
(Intercept)          age       gender   age:gender
3.204341e-44 1.517141e-15 4.537954e-01 9.098279e-03


fit.gee3 <- gee(distance ~ age + gender + age:gender, id=individual, family=gaussian,
corstr="AR-M", data=mydata1)
summary(fit.gee3)


> summary(fit.gee3)

GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link:                      Identity
```


```
Variance to Mean Relation: Gaussian
Correlation Structure:     AR-M , M = 1

Call:
gee(formula = distance ~ age + gender + age:gender, id = individual,
data = mydata1, family = gaussian, corstr = "AR-M")

Summary of Residuals:
Min          1Q     Median          3Q         Max
-5.7502655 -1.3670055 -0.1914044  1.2205495  5.1719079


Coefficients:
            Estimate    Naive S.E.  Naive z    Robust S.E. Robust z
(Intercept) 16.5946122  1.3530104 12.2649556   1.2788086 12.9766190
age          0.7694567  0.1166041  6.5988813   0.1049699  7.3302593
gender       0.7266739  2.1197599  0.3428095   1.4968683  0.4854628
age:gender  -0.2856919  0.1826835 -1.5638623   0.1223804 -2.3344571

Estimated Scale Parameter:  5.099523
Number of Iterations:  3

Working Correlation
      [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.6135308 0.3764201 0.2309453
[2,] 0.6135308 1.0000000 0.6135308 0.3764201
[3,] 0.3764201 0.6135308 1.0000000 0.6135308
[4,] 0.2309453 0.3764201 0.6135308 1.0000000
>

coef(summary(fit.gee3))
```

```
> 2 * pnorm(abs(coef(summary(fit.gee3))[,5]), lower.tail = FALSE)
(Intercept)          age       gender   age:gender
1.660501e-38 2.297079e-13 6.273481e-01 1.957180e-02
```

```
fit.gee4 <- gee(distance ~ age + gender + age:gender, id=individual, family=gaussian,
corstr="unstructured", data=mydata1)
summary(fit.gee4)
```

```
> summary(fit.gee4)

GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link:                      Identity
Variance to Mean Relation: Gaussian
Correlation Structure:     Unstructured

Call:
gee(formula = distance ~ age + gender + age:gender, id = individual,
data = mydata1, family = gaussian, corstr = "unstructured")

Summary of Residuals:
Min         1Q     Median         3Q        Max
-5.6285551 -1.3572403 -0.1781935  1.3128169  5.2189881


Coefficients:
```


```
            Estimate    Naive S.E. Naive z    Robust S.E. Robust z
(Intercept) 16.3236414 1.00582208 16.2291539   1.1701159 13.9504481
age          0.7881142 0.08500322  9.2715804   0.0982681  8.0200410
gender       1.0736105 1.57582036  0.6813026   1.3762246  0.7801128
age:gender  -0.3100200 0.13317445 -2.3279243   0.1172035 -2.6451442

Estimated Scale Parameter:  5.094256
Number of Iterations:  3

Working Correlation
      [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.5009582 0.7363481 0.5148767
[2,] 0.5009582 1.0000000 0.5552694 0.6208238
[3,] 0.7363481 0.5552694 1.0000000 0.7788356
[4,] 0.5148767 0.6208238 0.7788356 1.0000000
>


coef(summary(fit.gee4))

> 2 * pnorm(abs(coef(summary(fit.gee4))[,5]), lower.tail = FALSE)
(Intercept)          age       gender   age:gender
3.126351e-44 1.057099e-15 4.353245e-01 8.165610e-03
```

Bibliography

[1] Ming Wang, Lan Kong, Zheng Li and Lijun Zhang *Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples*. Wiley Online Library, 2015

[2] Wang, Y.-G., Lin, X. and Zhu, M. (2005) *Robust Estimating Functions and Bias Correction for Longitudinal Data Analysis*. Biometrics, 61: 684–691.

[3] Fitzmaurice G, Laird NM, Ware JH. *Applied Longitudinal Data*. John Wiley & Sons: New York, 2004.

[4] Wedderburn RWM. *Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method*. Biometrika 1974; 61:439–447.

[5] McCullagh P, Nelder JA. *Generalized Linear Models. Chapman and Hall*. London, 1989.

[6] Liang KY, Zeger SL. *A comparison of two bias-corrected covariance estimators for generalized estimating equations*. Biometrika 1986; 73:13–22.

[7] Gunsolley JC, Getchell C, Chinchilli VM. *Small sample characteristics of generalized estimating equations*. Communications in Statistics-Simulations 1995; 24:869–878.

[8] Wang M, Long Q. *Modified robust variance estimator for generalized estimating equations with improved small-sample performance*. Statistics in Medicine 2011; 30(11):1278–1291.

[9] M.J. Crowder, D.J. Hand. *Analysis of Repeated Measures*. Chapman & Hall/CRC: New York, 1990.

[10] Guo X, Pan W, Connett JE, Hannan PJ, French SA. *Small-sample performance of the robust score test and its modifications in generalized estimating equations.* Statistics in Medicine 2005; 24:3479–3495.

[11] Mancl LA, DeRouen TA. *A covariance estimator for GEE with improved small-sample properties.* Biometrics 2001;57:126–134.

[12] MacKinnon JG. *Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties.* Journal of Econometrics 1985; 29:305–325.

[13] Kauermann G, Carroll RJ. *A note on the efficiency of sandwich covariance matrix estimation.* Journal of the American Statistical Association 2001; 96:1387–1398.

[14] Gosho M, Sato Y, Takeuchi H. *Robust covariance estimator for small-sample adjustment in the generalized estimating equations: a simulation study.* Science Journal of Applied Mathematics and Statistics 2014; 2(1):20–25.

[15] Fay MP, Graubard BI. *Small-sample adjustments for Wald-type tests using sandwich estimators.* Biometrics 2001;57:1198–1206.

[16] Morel JG, Bokossa MC, Neerchal NK. *Small sample correction for the variance of GEE estimators.* Biometrical Journal 2003; 45(4):395–409.

[17] Kimihiro Noguchi, Yulia R. Gel, Edgar Brunner, Frank Konietschke. *nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments.* Journal of Statistical Software, September 2012.

[18] Shih WJ. *Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimating equations.* Biometrical Journal 1997; 39:899–908.

https://doi.org/10.26219/heal.aueb.6669