QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

ON VARIABLE SELECTION WITH THE PRESENCE OF MISSING DATA IN LONGITUDINAL PANEL STUDIES

BY

RADWA M. ISMAIL MOHAMED

A Thesis Submitted to

the College of Arts and Sciences

in Partial Fulfillment of the Requirements for the Degree of

Masters of Science in Applied Statistics

June 2022

COMMITTEE PAGE

The members of the Committee approve the Thesis of
Radwa M. Ismail Mohamed defended on 12/05/2022.

Abdel-Salam Gomaa Abdel-Salam
Thesis/Dissertation Supervisor

Saddam Akber Abbasi
Thesis/Dissertation Co-supervisor

Approved:

Ahmed Elzatahry, Dean, College of Arts and Sciences

ABSTRACT

ISMAIL, RADWA, M., Masters : June : [2022:], Applied Statistics

Title: On Variable Selection with the Presence of Missing Data in Longitudinal Panel Studies

Supervisor of Thesis: Abdel-Salam, G, Abdel-Salam.

Longitudinal data are valuable in various disciplines because they provide helpful developmental patterns over time. However, frequently, it is challenging to have a high dimension of covariates and ubiquitous missing values in longitudinal data due to individual nonresponse and drop out.

Response measurements in longitudinal studies are correlated within-subjects, where this challenge needs to be adequately handled using the linear mixed model (LMM) to get valid inferences and standard errors. LMMs provide an effective and flexible way to accommodate two types of parameters for between-subject correlation and within-subject variation. The powerful two-stage adaptive LASSO method for variable selection adopted provided promising results in LMMs. The joint modeling multiple imputations for handling missingness provided a consistent estimation of parameters and variance components.

Several researchers discussed the variable selection criteria and missing data handling in longitudinal studies separately. Hence, the thesis proposed a computationally efficient combining algorithm of multiple imputations and penalized variable selection using the stacked (homogeneous) approach. The homogeneous algorithm showed better estimation and selection properties.

DEDICATION

It is with genuine gratitude that I dedicate this thesis to my family and instructors/supervisors who have been a constant source of support.

ACKNOWLEDGMENTS

First and foremost, my appreciation to my work supervisor and thesis instructor, Dr. Abdel-Salam G. Abdel-Salam, where he inspired me to be an independent researcher and motivated me to investigate new areas in statistics. He was always there to give me sound advice whenever I had hard time with research and career.

I would like to express my deep gratitude to my co-advisor, professor Saddam Akber Abbasi. I am truly grateful for all the time and effort he has to put into monitoring me over the past years.

My sincere thanks must also go to Dr. Khalifa Al-Hazaa for all his contributions of guidance and encouragement to make my masters and work experience productive and stimulating.

I gratefully acknowledge Dr. Galal M. Abdella because he generously gave his time to offer me valuable feedback and suggestions toward improving my work. I would like also to extend my appreciation to Dr. Abdul Haq (Quaid-i-Azam University, Pakistan) and Dr. Md. Hamidul Huque (The University of New South Wales, UNSW Sydney, Australia) for providing helpful contribution that helped in the methodology.

I would like to acknowledge the support of Statistics Department and Graduate Studies Department in Qatar University for providing all the needs to achieve the requirements of the Master's degree.

Last but not the least, my family and friends deserve endless gratitude for their love and support. To my family, I give everything, including this. I am also indebted to my colleagues Waleed Rahmat Ullah (Senior Research Analyst, Student Experience Department) and Dr. Ahmed Ben Said (Senior Data Analyst, Student Experience Department) for their encouraging words, continued guidance, and tremendous support. Words cannot express how fortunate I am, thank you all for the strength you gave me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Motivation of Thesis

High instability and unbiasedness may result when the estimation model includes all the candidate variables, especially those highly correlated (Greenland, 1989). In multivariable analysis, redundant or irrelevant predictors add noise to other quantities that the model is interested in estimating, cause multicollinearity, and increase the cost and time for measuring trivial predictors. Based on the principle of parsimony, having fewer variables in a simple model is preferred over complex models in terms of computational time, cost, and interpretation (Chowdhury & Turin, 2020). Hence, it can be valuable to select a limited subset of the most relevant predictors (i.e., covariates) to the response variable to include it in the statistical model. The terminology "variable selection" is a special case of "model selection" and is often used when the competing models agree on the mathematical form of predictors but differ on which predictor should be included. The variable selection also includes choosing the product terms (i.e., the interaction between regressors) to enter the model. Historically, variable selection has occurred directly on linear models because 1) analytic tractability facilitates great insight, and 2) many problems can be represented as linear models. However, the development of computer technology allowed the implementation of richer treatments of the variable selection problem in the general model selection framework (George, 2000).

Stepwise algorithms are the most widely applied techniques that work on two classes of variables. One class of variables enters the initial model and is not subject to deletion (i.e., forced-in variables). At the same time, the other class of variables enters a repeated cycle of selection-deletion by the stepwise regression algorithm (i.e., non-forced variables). The two conventional stepwise strategies are Forward-selection and

Backward-deletion. Any variable selection algorithm must be evaluated for a) validity of the estimates, b) sensitivity of selecting the candidate variables, and c) specificity of screening out the inappropriate candidate variables. Greenland (1989) indicated that conventional stepwise variable selection tools are impaired by poor sensitivity (low power) and may lead to the nonnormality of coefficient estimates. An alternative selection strategy to these conventional algorithms is to keep the outcome variable forced in every fitted model, and the predictor variables are selected based on the changes they impose on the estimated outcome. This algorithm is called change-in-estimate and is implemented in a stepwise fashion. Evidence tends to favor the change-in-estimate algorithm to control the nonnormality in a way that produces the most valid coefficient estimate and standard error.

On the other hand, it is worthwhile to use the prior information concerning the effects of the variables that are not forced to be in the model. This information helps in the construction of the coefficient estimates. Mitchell and Beauchamp (1988) proposed a hybrid Bayesian variable selection approach. It is assumed that during the analysis, some predictors may be deleted from the model, concluding several possible sub-models (2^p), where p is the given predictor variables in the data. A sequence of prior distributions is used for the regression coefficient (β) and the random error term (ε). The distribution of the regression coefficient is uniformly diffused, covering the region of non-eligible likelihood, except for a bit of probability mass concentrated at 0 if the predictor variable is to be deleted. Hence, the selection of the candidate variables is based on a nontrivial limiting set of posterior distributions obtained by Bayes' theorem. However, the difficulties associated with prior and posterior computation arise when the set of variables and candidate models are large.

Moreover, prior distribution requires significant effort. Therefore, Markov

chain Monte Carlo (MCMC) implementation for prior specification has been proposed to develop the Bayesian variable selection approach. MCMC implementation is more versatile and offers improved performance (George, 2000). Other regularization approaches proposed include the ridge regression, least absolute shrinkage, and selection operator (LASSO), Bridge (Fu, 1998), and elastic net (Zou & Hastie, 2005). These methods can handle datasets even when the number of predictor variables is much larger than the number of observations. In this regard, Fan and Lv (2010) reviewed the methods that cope with high and ultra-high dimensionality.

Tibshirani (1996) introduced the LASSO, which does not focus on subsets but it improves the overall prediction accuracy of the ordinary least squares (OLS) estimates by shrinking some regression coefficients and setting others to exactly 0. Also, it facilitates the interpretation of the model by determining a small model that exhibits the strongest effect. Another method proposed by Breiman (1995) is called nonnegative (nn) garotte for doing subset regression. This method eliminates some variables and shrinks others. Unlike subset and ridge regression, nn-garotte is relatively stable and scale-invariant.

On the other hand, subset selection is unstable because small changes in the data can result in a drastic change in the selected model and reduce the prediction accuracy. Ridge regression is more stable in comparison to subset selection. However, ridge regression is not scale-invariant, and the recipe is to standardize the predictor variables before applying ridge. Furthermore, Ridge regression does not set any coefficient to zero, which gives an easily interpretable model. LASSO and nn-garotte are prime competitors to subset and ridge regression. However, LASSO retains the good features of both subset and ridge regression. In the presence of highly correlated variables, LASSO estimators cannot be consistent. Hence, adaptive LASSO was introduced

earlier by Zou (2006) to amend the deficiencies of LASSO (Epprecht et al., 2021).

In statistical practices, datasets are often incomplete, while most variable selection techniques require complete datasets. Missing data is in various settings, including surveys, clinical trials, and longitudinal studies. This complication is commonly encountered while selecting important variables due to nonresponse or individual withdrawal from a study. Liu et al. (2016) stated that the default practice for dealing with missing data with the variable selection approach is listwise deletion. Listwise deletion or complete-case strategy excludes individuals if they are missing any of the variables included in the analysis. However, listwise deletion can introduce bias when the missing completely at random (MCAR) assumption is not satisfied. Performing variable selection with complex missing data patterns and mechanisms raises several new challenges, underscoring the need for adequate statistical models. Zhao and Long (2017) provided a review for all the methods of imputation-based variable selection in datasets prone to MCAR or missing at random (MAR) missingness in linear regression models when the number of predictors (p) is allowed to be smaller or more significant than the sample size (n). The authors illustrated three strategies that combine the variable selection techniques with the imputation methods, which overcome the formerly stated challenges. The first strategy combines the variable selection techniques applied to each imputed dataset. For instance, conduct variable selection on each imputed dataset separately, then choose the final set of the selected variables based on a pre-specified threshold π. The variable selection methodologies that can be conducted under this strategy include: 1) Forward, backward, or stepwise Wald test selection method; 2) likelihood ratio test; 3) Akaike information criteria (AIC); 4) Regularization method (e.g., lasso, ridge, or elastic net); 5) Bayesian variable selection method.

The second strategy applies the variable selection technique on stacked imputed datasets. Variable selection techniques, like backward variable selection and the penalization elastic net, are applied on the multiply imputed *m* datasets stacked to form one large dataset. However, this strategy is prone to standard error underestimation because the sample size is artificially increased. Therefore, one approach is conducted by introducing a fixed weight for all observations, and then a classical variable selection technique is implemented.

The third strategy applies variable selection on imputed datasets combined with resampling techniques. Resampling techniques (e.g., bootstrapping) have several advantages because they are similar to multiple imputations in analyzing each dataset independently, combining the individual analysis results into one final result, and generating multiple datasets that preserve the variations in the data. The proposed variable selection techniques that provide superior performance when applied to the bootstrapping are the different versions of the randomized lasso. The growing literature about combining resampling techniques and multiple imputations into the variable selection process is summarized. Two approaches that demonstrated a good performance compared to all other techniques in this context are the stability selection within bootstrap imputation (BI-SS) and multiple imputation random LASSO (MIRL). In BI-SS, bootstrapping samples are first generated from the incomplete dataset, and multiple imputations are implemented on each bootstrap dataset. Then, a randomized lasso is used on each multiply imputed bootstrap dataset to estimate the unknown parameter, and the final set of variables is determined by stability selection determined by a threshold π. MIRL is the multiple imputation random lasso method that can accommodate the high proportion of missingness and the collinearity between covariates. Bootstrap samples are generated from the multiply imputed datasets, then

lasso-OLS estimates are obtained. Variable selection in MIRL is based on the stability selection determined by the ranks of importance measure.

Yang et al. (2005) introduced two Bayesian variable selection approaches on multiply imputed datasets when the covariates have ignorable missing data through the stochastic search variable selection procedure (an MCMC algorithm). First, "Impute, then Select" (ITS) and the embedded methodology to a single combined imputation and selection processes, "Simultaneously Impute and Select" (SIAS). ITS refers to generating multiply imputed datasets and applying the variable selection to each of the imputed datasets. On the other hand, SIAS refers to a single combined Gibbs sampling process of imputation and the variable selection steps. Findings indicated that SIAS provided smaller Monte Carlo standard errors, better than the ITS in performance. Collinearity between covariates and higher rates of missing data worsens selecting the right variables in ITS and SIAS. ITS is slightly worse but easier in implementation. Both Bayesian strategies showed outstanding performance compared with the case-deletion stepwise regression.

In another study by Chen and Wang (2013), multiple imputation-LASSO variable selection approaches are presented to overcome the difficulty of interpreting the final model and inferences when LASSO is applied on each imputed dataset. The authors chose to base their study on sequential regression multiple imputation framework to handle the ignorable missingness with a haphazard pattern in both the response variable and the covariates. Multiple imputations are generally more appealing than the maximum likelihood estimates calculated from the incomplete data. The novel methodology considers that the regression coefficient estimates are treated through the different imputed datasets associated with the same variable. The group LASSO penalty is applied to select the whole group or remove the whole group when

all estimated coefficients are zero. This determines the uncertainty among imputed datasets caused by the missing information and performs consistent variable selection. Chang and Wang's research showed that their proposed approach outperforms the CC-LASSO method performed on complete observations ignoring missingness when data is generated from multivariate normal with compound symmetry covariance structure, and missingness is MAR. On the other hand, the RR-stepwise method applied to multiply imputed datasets showed poor performance, especially with multicollinearity and small sample size with many covariates.

Recently, Pitchiah et al. (2021) introduced an approach that gives a minimum error and good conclusions when dealing with insufficient data that has missingness in critical parameters. The authors considered a thorough analysis of missing data which depends on the form, the cause, and the trend of missingness. Adaptive multiple imputation LASSO (MIAL) is the novel approach proposed by the authors to perform variable selection in settings of data multidimensionality and incorporates issues like nonresponse, mistakes, and system malfunction. The adaptive LASSO algorithm has additional capabilities because of the incorporation of multiple imputations. Under the four major stages of the proposed MIAL method, random missing data under the MAR assumption can be managed, the high-dimensionality case ($p > n$) is considered, and variable selection and predictions are of high precision due to the stability selection criteria used in MIAL. Liu et al. (2016) used the LASSO-OLS on imputed dataset combined with resampling technique (e.g., bootstrap). Alternatively, Pitchiah and others introduced the adaptive LASSO to the resampled imputed datasets.

There is substantial literature where variable selection received attention, especially with the increase of applications and the massive data structure. Investigators have the desire to construct an economic predictive model from thousands of candidate

variables. When the set of candidate variables gets larger, the efficiency of the model decreases. Therefore, selecting the appropriate variables of interest in the model is considered the most crucial and challenging aspect to avoid model underfitting and overfitting. Model misspecification can result when the model is simplistic, where key variables are excluded, or extraneous variables are included. Each additional variable adds to the variance of the model, and using too few variables leads to increased bias. Over the years, the development of variable selection also incorporated variable selection of correlated data due to repeated measurements on the same individual over a specific time interval. One popular aspect of repeated measurements is that missing data are ubiquitous, which is caused by missing items or questionnaires due to individual nonresponse and withdrawal. To select the best variables for correctly analyzing data in longitudinal settings the correlation and missing data need to be acknowledged and considered to provide the best fit for having accurate predictions.

Missing observations impose challenges in model fitting and variable selection. There are three types of assumptions/mechanisms of missing data: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) depending on the factors related to missing probability, whether observed or not. Assumptions about missing data are substantial for choosing the needed handling method, which has considerable advantages in implementing a particular study. Bhaskaran and Smeeth (2014) elucidated the confusion between the two standard missingness mechanisms for implementing the novel missing data handling methodologies, namely, MCAR and MAR. In short, the distribution of missing data is systematically similar to the distribution of observed data when the MCAR mechanism is satisfied. Alternatively, both will have the exact similar pattern/shape of the histogram using the imaginary histogram. While in the MAR mechanism, there is a

systematic difference between the missing and observed data, explained by the differences among other variables in the dataset.

On the other hand, the third well-known NMAR mechanism does not have the advantage of providing unbiased parameter estimates by handling missing data and is not statistically valid unless the missing data model is tailored with caution. Ignorability assumption of missing data exists when the MAR mechanism and other technical conditions are satisfied. However, MAR and ignorability are used interchangeably in real data applications since the technical conditions are most likely to be violated. Technical conditions specify that the parameters responsible for missing data mechanisms are distinct from the parameters in the estimated model.

Missingness is ubiquitously available in multiple patterns across the real-world experiments or surveys: data are missing on all variables for a single observation (i.e., subject), data are missing on a variable for all observations, and data are missing for some variables and some observations. Unlike the MAR-based methods, the ad-hoc methods for handling missing data, for instance, the Listwise Deletion, are insufficient in attempting one or more of the three specified goodness criteria and lack strong "mathematical foundations" (Allison, 2009). He (2010) indicated that complete-case analysis could produce biased point estimates when the missingness mechanism is not MCAR, and conventional imputation methods (e.g., mean imputation) underestimate the standard errors. On the other hand, the superior method of multiple imputations (MI) relies on the weak assumption for missing data mechanism (i.e., ignorability) and follows two well-known approaches for imputing multivariate datasets. Buuren (2007) stated the two imputing approaches; the Joint Modeling (JM) approach, which is, based on "parametric statistical theory," and the more flexible alternative Sequential Regression Multiple Imputation (i.e., Fully Conditional Specification FCS), which is,

based on "semi-parametric statistical theory." To get correct imputation results and unbiased estimates, the imputation model must be compatible with the analysis model, and this may lead to the necessity of generating different imputed datasets compatible with each different analysis model. It is also recommended to use models that impose no structure for the covariance matrix among the repeated measurements to overcome the exchangeability resulting from random intercept mixed models and the complications encountered for getting the right covariance matrix structure (Allison, 2012).

Objectives of Thesis and Outline

Commonly in longitudinal studies, where subjects are measured repeatedly over time, subjects may miss some scheduled visits or drop out before observing a given follow-up endpoint for some reason. Longitudinal studies have an important role in health sciences in understanding the development and persistence of a certain malady. As shown in Figure 1, the literature covered three mechanisms of missingness, three missing data patterns, and three common variable selection classes of techniques in longitudinal settings. Classical methods for variable selection are intensive and computationally unstable with many longitudinal data, for instance, high-dimensional data (Zheng et al., 2018). When the number of candidate variables of interest gets larger, the number of candidate models to be selected in a certain classical model selection strategy increases causing an infeasible computation process. Therefore, the novel variable selection strategies (i.e., penalized and Bayesian variable selection techniques), especially within linear-mixed modeling, are attractive for various longitudinal settings because the coefficients of unrelated covariates are shrunken to zero, and parameters are estimated simultaneously. These algorithms apply to longitudinal data prone to both dropouts (i.e., monotone) and intermittent missing

values under the ignorable assumption. Monotone missingness assumes that there "exists a permutation of the measurement components" (Verbeke & Molenberghs, 2009, p. 215). Therefore, it is necessary to have a balanced study structure (i.e., balanced panel data) to make the pattern of missingness meaningful.

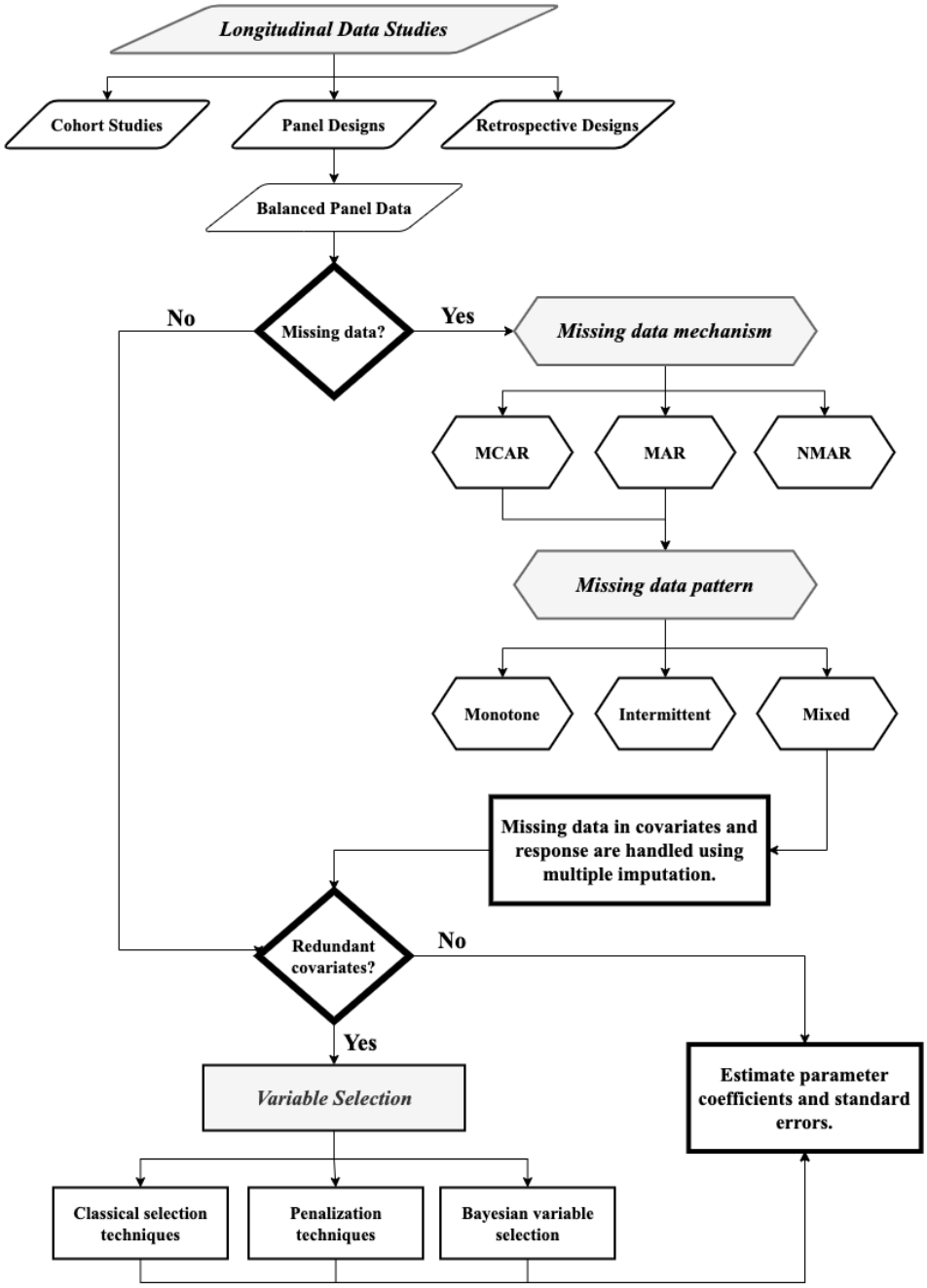

Figure 1. Flowchart of Missing Data and Variable Selection in Longitudinal Studies.

The common assumption for missing data in the real world is missing at random (MAR). However, MAR is a considerably weaker assumption than MCAR. In other words, if the missing data mechanism is assumed to be MCAR, then the missing data is MAR (Allison, 2001). In general, modern missing data handling approaches start from the MAR assumption. Multiple imputation-based statistical inferences are widely studied, but few can effectively perform variable selection with multiply imputed datasets (Shen & Chen, 2013). In longitudinal settings, subjects may have missing observations at once but have observations after a sparse structure of the data (intermittent), or subjects drop out permanently before the study completes (monotone). The third pattern of missingness is the mixing between both intermittent patterns followed by permanent dropouts. MI is one of the modern techniques developed under the assumption that the data are MAR or possess the ignorability mechanism. The correct implementation of the MI approach demands the central feature that MI adopts; the compatibility between the model of interest and the imputation model for preserving the multilevel data structure (Lüdtke, Robitzsch & Grund, 2017).

Multiple imputations with substantive-model-compatible joint modeling (JM-SMC) and substantive-model-compatible joint modeling with random covariance matrix (JM-SMC-het) approaches are recommended when both the covariates and response variable contain missing values (Huque et al., 2020; Goldstein et al., 2014). He (2010) explained that the JM approach partitions the observations based on their missing data patterns then specify a parametric multivariate density suitable to the type of data given using the appropriate prior distribution for the parameters. Imputations are drawn from these multivariate density sub-models. Simulation results in Huque et al. (2020) showed that JM-SMC-het outperforms the other methods, like the FCS

approach, concerning the regression estimation coefficients and variance components. Also, it is observed that JM-SMC-het better estimate subject-specific associations under the random covariance matrices than the JM-SMC. Buuren et al. (2010) stated that FCS is an alternative to JM when no reasonable multivariate distribution of the missing data is provided. However, Quartagno and Carpenter's (2019) results showed that JM performs better than FCS when the latent normal variables are used with the JM as an extension to impute categorically or count data (i.e., non-normal data). Moreover, imputation using FCS cannot handle variables measured at higher levels in multilevel data structures.

In longitudinal studies, Geronimi and Saporta (2017) presented a novel methodology for variable selection with missing data that outperforms the single imputation (i.e., mean imputation) when the missing data rate is small and accepts any correlation structure. However, higher rates of missing data, for example, 60% missingness, can lead to acceptable imputation, but predictors selection is insufficient. Another limitation is that the variable selection interferes when the study suffers from dropouts because the distribution of missing data is not examined under the penalized generalized estimating equations (GEE) approach. Both Marino et al. (2017) and Li et al. (2019) proposed methodologies that showed promising results in multilevel data, but the variable selection focuses only on fixed effects. Also, the methods need to be improved for higher rates of missing data.

Very limited literature jointly addresses variable selection in single-level longitudinal data (i.e., panel data) with missing values in both covariates and response using a mixed-effects model. However, many existing works of literature studied these challenges individually. The main objective of this thesis is to provide a sufficient variable selection methodology that jointly handles missingness and selects the

significant covariates, where the selection of variables is applied to both the random and fixed effects. More specifically, the thesis tends to provide an extension for the sophisticated variable selection technique developed by Pan and Shang (2018) by tailoring it to accommodate missing data assumptions and imputations in longitudinal studies. This sophisticated direction overcame the challenges in the literature while analyzing the longitudinal data hindered by missingness in both covariates and the response variable and by introducing the random effects. The proposed approach is also applicable to high-dimensional settings since the technique is a shrinkage-based method.

The adaptive LASSO regularization technique by Pan and Shang (2018) can be tailored to multiply imputed datasets 1) to accommodate missing data in both covariates and response variables and 2) to select variables from both fixed and random effects when the linear mixed model is employed, implementing profile log-likelihood and Newton-Raphson algorithm. Applications of adaptive LASSO on missing data can be particularly challenging, and methods that combine both complexities are not widely applied for mixed models in longitudinal data. Zou (2006) stated that the LASSO is a regularization approach that uses the tuning parameter λ in the ℓ_1 penalty to perform variable selection and parameter estimation simultaneously. When the tuning parameter of prediction is optimal (i.e., sufficiently large), some coefficients are shrunk to exact zero, indicating that the non-zero coefficients are selected to be in the model. However, the LASSO shrinkage produces biased estimates for large coefficients due to the many noise features included in the predictive model. The Adaptive LASSO is an enhanced version when the LASSO is an inconsistent variable selection technique in specific scenarios. Adaptive LASSO assigns different (data-dependent) adaptive weights for penalizing different coefficients in the ℓ_1 penalty. LASSO and Adaptive LASSO both

are ℓ_1 penalization methods that pose a similar algorithm to compute the estimates.

The remainder of the thesis is organized around six chapters. Chapter 2 discusses the widely known modeling techniques of longitudinal studies. An overview of the variable selection methods that are used for incomplete longitudinal studies is given in Chapter 3. In Chapter 4, we first define the notations of linear mixed models for longitudinal data that will be used in the remaining sections of the thesis. Then we employ the multiple imputations using joint modeling to handle the missingness. Finally, the two-stage adaptive LASSO approach needs adaptation when applied to multiply imputed longitudinal data. In this stage, we combine multiple imputations and adaptive LASSO by the stacking (homogeneous) approach. In fact, the way of combining multiple imputation and variable selection techniques provided a challenge in the literature and there is a lack of tools to address this challenge in a principled way (Zhao & Long, 2017; De et al., 2020).

In Chapter 5, we present simulation studies and comparisons that illustrate the effectiveness of the proposed algorithms. As a motivating example, we consider a study of the association between country-level food security and economic growth. Global food security issues consider multiple aspects: affordability, availability, quality and safety, and natural resources. According to the Rome Declaration on the 1996 World Food Summit, "Food security exists when all people, at all times, have physical and economic access to sufficient, safe and nutritious food to meet their dietary needs and food preferences for an active and healthy life." Food security and resilience significantly benefit human beings and are also essential in achieving sustainable economic growth. Thus, food security is more than a single sector issue; it requires combined coordination between the actions in finance, agriculture, health, and other sectors (Torero, 2014). Further details of this data and the associated indicators that

measure the drivers of food security across selected developed countries are described in Chapter 6. Finally, we conclude the thesis and offer some additional remarks in Chapter 7. Future research recommendations are also provided in this chapter. Figure 1 summarizes the gap in literature, research objectives and the structure of the manuscript.

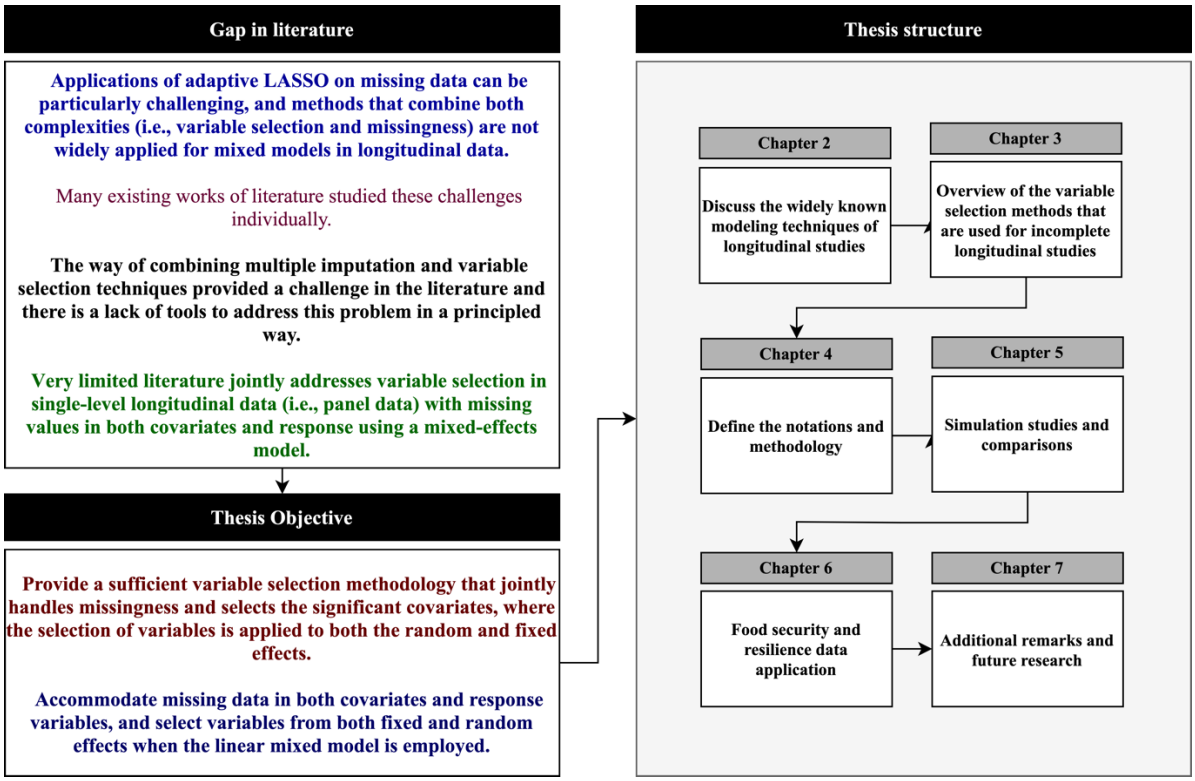

Figure 2. Gap in literature, objectives, and structure of the thesis.

CHAPTER 2: MODELING TECHNIQUES IN LONGITUDINAL DATA STUDIES

Overview

Longitudinal data is mainly concerned with directly assessing the heterogeneity among individual characteristics and differences in the response variable by repeatedly measuring the individuals throughout the study. In most cases, a fixed number of repeated measurements are observed on all study individuals on a set of occasions. The longitudinal analysis proceeds in two distinct stages; to describe trends in the within-individual changes in the response and to relate these changes to inter-individual differences in selected factors. Also, longitudinal studies can predict how a particular individual changes over time by borrowing information from other individuals to make reliable predictions. One inescapable feature of the longitudinal analysis is obtaining great precision estimates. Any extraneous factors (i.e., not of substantive interest and its influence remain stable throughout the study) that impact the response are blocked when an individual's response variables are compared over time. These noise factors may include gender, genetic factors, and socioeconomic status.

Participants being studied under longitudinal analysis are individuals (e.g., human subjects, animals, regions). Occasions of repeated measurements are either distributed equally or unequally throughout the study. Balanced longitudinal designs occur when all study units have an exact number of observations measured at a common set of time points. In contrast, unbalanced designs are when the sequence of observation times or the repeated measures are not common to all individuals because of mistimed measurements. In other words, balanced longitudinal data is when the timing of observations is defined in terms of one origin and unbalanced when measurement time is defined based on different origins. With repeated observations in longitudinal studies, past observations very likely predict future observations of the same individual. Hence,

the fundamental assumption of many statistical techniques that require random variables independent of one another is no more invalid. As a result, longitudinal studies are positively correlated, and the correlation strength is a decreasing function of time separation. Between-subject heterogeneity, within-subject heterogeneity, and measurement error are the three sources of variability that impact the correlation among repeated measurements on the same individual. The between-subject heterogeneity is the one important source of positive correlation among repeated measures due to the variability in the response variable between individuals in the population or the individuals' underlying propensity to respond. An individual responding with a low value for the response on one occasion is expected to have a low value on the subsequent occasion, so repeated measures from the same individual are expected to be more similar than measurements from randomly selected individuals. As a result, this accounts for the positive correlation among repeated measurements. The other two sources of variability are discussed thoroughly in Fitzmaurice et al. (2012) book.

There are three types of longitudinal study designs: panel, cohort, and retrospective. Panel data, sometimes referred to as longitudinal data, measure two or more variables on the same individuals collected at a regular frequency of occasions and ordered chronologically (i.e., same individuals are measured at every wave). Panel data intends to 1) explain a pattern of change described by the chronological time; and 2) establish the sign, direction, and magnitude of a causal relationship. For instance, longitudinal panel data can be used to examine an individual's historical or developmental characteristic changes over time and age.

For longitudinal analysis, a wide range of techniques is available that is statistically adjusted because the observations are not mutually independent. The choice of a specific analytical approach depends on multiple aspects, for example, the design

of the study, the scientific question of interest, and other statistical considerations. The lack of understanding concerning the appropriate analysis technique leads to inefficient analysis, inaccurate results, and wrong interpretation. A simple regression of the dependent variable on time, repeated-measure analysis of covariance (ANCOVA), and generalized linear models (GLM) with fixed-subject effects are examples of traditional analytical methods used to analyze longitudinal data. However, these methods are disadvantages because they are restricted to balanced data without missingness and few time points. Therefore, these methods are not recommended unless for quick exploratory respects. On the other hand, random-coefficient models (i.e., random-effect modeling), generalized estimating equation (GEE) models, and latent growth curve models are more advanced and leading methods. Some of the most common and developed topics of different modeling techniques in literature are discussed below.

Random-effect Modeling

In random-effect models, some subset of the model parameters varies randomly from one individual to another, accounting for the heterogeneity across the population. The linear mixed-effects model (LMM) assumes that the response variable depends on a combination of population characteristics (i.e., fixed effects) and subject-specific effects (i.e., random effects). Because of this explicit modeling of fixed and random effects, LMMs can analyze both between-subject and within-subject variations in longitudinal study designs. In other words, prediction of individual growth over time is also possible besides predicting response change in population level. Furthermore, LMMs are flexible in terms of dealing with any degree of imbalanced longitudinal data. It is not required to have the same number of individuals observed or take the measurements simultaneously. Cnaan et al. (1997) used a general linear mixed model (GLMM) in their tutorial to illustrate the design in which subjects are allowed to have

an unequal number of observations. An epidemiologic application in the Cnaan et al. (1997) study of pulmonary function growth in children ages six to eighteen indicated different ages at study entry, forming highly unbalanced data.

Generally speaking, studies based on LMMs assume that the data follows a normal distribution. However, the usual assumption of normal random effects is often unrealistic in real data applications. Parzen et al. (2011) considered using a developed approach of GLMM that modifies the random effects to follow a bridge distribution to fit longitudinal data with binary outcomes. The proposed method used correlated random intercepts that led to a marginal logistic regression model. It is comparable to probit-normal marginal models. Parzen et al.'s (2011) method can be 1) generalized to be used with any link function with appropriate bridge distribution, 2) easily implemented using the existing software packages, and 3) used in the wide applications of social and behavioral sciences. Li et al. (2004) proposed a semi-parametric approach that requires no assumption on the random effects distribution. However, the normality of within-subject longitudinal data is assumed and often reasonable to explore. The proposed joint modeling approach assumed that longitudinal data follows LMM, whose random effects are covariates in a GLM. This approach can be applied to any generalized linear model formulation.

Moreover, Arnau et al. (2012) examined the robustness of LMM when the data is showing slight skewness and extreme kurtosis. Distributions that involve absolute values of 1.0 or more in skewness and kurtosis are closer to that data found in real life, that is, the log-normal distributions. The non-normality of the data affects the business of the estimation methods. LMM, in combination with Kenward and Roger (KR) method, is used to control the bias of fixed effects estimation. Recently, Wu and Jones (2021) introduced the proportional likelihood interpretation to the mixed-effects model

(PLRMM) to fit the finite ordinal (discrete) outcomes in clinical rating scale applications.

Incorporating the random effects in the linear predictive model require full distributional assumption for the response variable. In contrast, GEE depends on the correct specification of the first and second moments of the response variable. In the next subsection, GEE is discussed because it is one of the predominantly used methods in longitudinal studies for providing population average (i.e., population-level) inferences.

Marginal Modeling

The marginal model, known as a generalized estimating equation, is a convenient method with no distributional assumption about the observations. Based on the concept of "estimating equations," GEE provides a unified approach for estimating correlated discrete or continuous response variables. In contrast to LMMs, the mean response of GEE depends only on the covariates of interest and accounts for within-subject correlation among the repeated measures.

Due to the challenges presented by the correlation among longitudinal data, Liang and Zeger (1986) introduced the GEE, a class of GLMs that considers more efficiency. Zeger et al. (1988) used the GEE approach to fit subject-specific and population-averaged models with discrete and continuous response variables. Subject-specific models explicitly model the heterogeneity across individuals, while population-averaged models use a function of covariates without explicitly specifying the individual-to-individual heterogeneity. Another study by Miller et al. (1993) extended the marginal modeling approach to accommodate categorical (polytomous) response variables. In addition, a connection between GEE and weighted least squares (WLS) was developed. Both latter studies used a longitudinal clinical trial of respiratory

disease.

GEE approach has an issue with the theory that supports the consistency of the joint distribution of the estimates of regression coefficients β and the nuisance variable α in the working correlation matrix, especially if the correlation structure is misspecified. Thus, the whole estimation procedure is ruined. Chaganty (1997) provided a new method that obtained unique and feasible estimates for the correlation (i.e., nuisance) parameter. Chaganty's approach is an extension of generalized least squares where the covariance matrix elements are functions of the regression parameter. In this study, a balanced longitudinal application is employed.

Various real data applications employ GEE to conduct the longitudinal data analysis. For instance, in a prospective study design, Carlier et al. (2013) assessed the relationship between re-employment among unemployed people in the Netherlands and their general health and quality of life. Whereas, in a retrospective cohort longitudinal study, Noda et al. (2015), patient-based GEE is employed to evaluate patients' risk factors for dental implants failures. The smoking habit factor is identified as an impact on the early implant failures. At the same time, some risk factors that affected the late implant failures are the number of remaining teeth, maxillary implant, and having a removable superstructure type.

CHAPTER 3: VARIABLE SELECTION METHODS IN INCOMPLETE LONGITUDINAL STUDIES

Many covariates often need to be minimized in longitudinal studies to determine the most predictive variables to the response. Variable selection methods fall within one of the three categories: 1) classical approaches, 2) penalized shrinkage methods, and 3) Bayesian variable selection. For example, Ni, Zhang, and Zhang (2010) proposed a double-penalized likelihood approach for simultaneous model selection and parameter estimation, which considers the dependent nature of longitudinal data and guard against the data missingness.

Gokalp Yavuz and Arslan (2019) focused on the selection of variables in elliptical LMM with "shrinkage penalty function (SPF)." SPFs simultaneously select variables and estimate the parameters. Several studies implement the full LMM definition with elliptical distributions to help overcome outliers and heavy tailedness within the data. In addition to robust estimation, one of the LMM's severe topics is variable selection. Shrinkage methods have recently emerged as an efficient procedure for the selection of the model. For instance, ridge selection is better than subset regression considering its variance reduction and accuracy as one of the most common shrinkage methods. However, the ridge regression method has its disadvantages, and the subset procedures are not stable when a minor data change may result in a different selection model. Shrinkage methods performing variable selection like LASSO overcome such obstacles. Two different methods for making comparisons are used to conduct simulation studies. These are the ECM algorithm for t-distributed LMM and ECM algorithm for the classical LMM. Using the smoothly clipped absolute deviation (SCAD) approach to expand the ECM algorithm in elliptical LMM enables one to select fixed effects effectively. The simulation results show that the proposed method is better

than classical ECM-SCAD, especially with longer-tailed data. Pan and Shang (2018) have proposed a procedure for variable selection to simultaneously select the random and fixed effects in linear mixed models. They employed an adaptive LASSO penalty and a profile log-likelihood for selecting and estimating variables. Newton-Raphson (NR) optimization algorithm is used to estimate the parameter.

Lee and Chen (2019) looked at the corresponding selection of variables in models with many parameters (p). Their research uses the Bayesian variable selection approach instead of the penalized approaches. The Bayesian method is for variable selection for "finite mixture of linear mixed-effects models (FMLMEMs)" fitting. Their corresponding algorithm can also determine the model's variable importance and classify the observation. The article presented a unified approach for FMLMEMs to predict the component membership and identify the essential fixed and random effects. In the simulation study, the approach outperforms selecting variables and classification, including $p > n$ and multicollinearity problems. Yang et al. (2020) proposed a novel Bayesian approach with a penalized shrinkage distribution for a joint selection of random and fixed effects simultaneously in LMM. The shrinkage effect improves the accuracy and efficiency of the technique. Simulation results provided outperformance of the proposed shrinkage approach.

Recently, Chen and Yin (2022) proposed a method for selecting variables in high-dimensional longitudinal settings when the data is following non-Gaussian distribution, especially ordinal data. the selection of variables is handled using penalized GEE where the penalty function is the nonconvex SCAD. Another study by Taavoni and Arashi (2022) handled longitudinal data with multiple response variables using multivariate linear mixed models. Multivariate linear mixed model is extended in their study to give robust inference against the potential outlying observations

considering joint multivariate-t distribution for the random effects and the within-subject errors. SCAD penalty function is used to select variables and the computationally flexible expectation conditional maximization (ECM) is used for parameter estimation.

The longitudinal data is modeled using two extended popular classes of GLMs: generalized estimating equations and linear mixed-effects model. The GEE includes a semi-parametric design as a particular case to address the MAR missing data mechanism. It relies on estimating equations to address the correlated repeated measurements, while LMM uses random effects. These two classes of models employ different aspects to capture both the between-individual differences and within-subject dynamics. The following subsections present a review for these two model classes (i.e., GEE and LMM) in the context of variable selection for incomplete longitudinal data, constraining the search covering the years 2010 through 2022. The search is conducted using the Advanced Search function in Google Scholar with the following terms: variable selection, missing, missing data, dropout, incomplete, longitudinal. All the articles mentioned down are published within Wiley, SAGE, Springer, Taylor & Francis, and Elsevier databases. The sources identified are then manually examined and checked in the references section of each article for additional relevant resources. Figure 3 lists a summary of the algorithms extracted from the reviewed articles.

GEE Based Methods

This section reviews the recent studies that combine the missing data handling and variable selection in a longitudinal setting with the GEE approach. GEE is a semi-parametric, marginal extension of the GLMs that models the mean response instead of the within-subject covariance matrix in longitudinal data. For parameter estimation, GEE uses quasi-likelihood estimation instead of the maximum likelihood. These

estimators are consistent even if the covariance structure is mis-specified because GEE depends on the first moment (i.e., mean) (Slavkovic, 2018).

The validity of estimation methods in longitudinal studies often relies on complete data and precisely measured covariates. In 2015, Yi et al. proposed a simultaneous variable selection and estimation algorithm that address the features of missing data and covariates measurement error at the same time. Their method extended the ordinary simulation-extrapolation (SIMEX) procedure that accommodates the covariate measurement error by including additional steps for missingness and variable selection. The inverse probability-weighted GEE is developed to accommodate the effects of missingness in the response variable. For variable selection, a penalized quadratic loss function using LASSO type and SCAD type penalty functions is used with the optimal tuning parameter selected based on the smallest $BIC(\lambda)$ value. As mentioned before, using a marginal generalized linear model is attractive because the minimal distribution assumptions are required for the response process. Simulation results showed that the measurement error and sample size affect the variable selection performance. The sensitivity, specificity, and correct fit rate decrease as the measurement error increases when the sample size increases. However, missing data have a more considerable impact on variable selection performance than measurement error.

Geronimi and Saporta (2017) extended Chen and Wang's (2013) methodology (MI-LASSO) to deal with longitudinal studies. Penalized generalized estimation equations (PGEE) is a penalized regression introduced to the GEE to select the most influential variables when the model has many covariates. The new method, MI-PGEE, has the advantage of accommodating missing values scattered haphazardly through the data using multiple imputations by chained equation (i.e., FCS). Also, it integrates the

intra-subject correlation into the variable selection framework. Common ridge penalties and adaptive weights are introduced to the group of estimated regression coefficients in the GEE estimation equation across multiply imputed datasets, and the tuning parameter is selected using the BIC-like criterion. As a result, a unique selection across all imputed datasets is obtained. In this manner, the coefficients are all zero or non-zero over the *m* imputed datasets. Simulation studies indicated that the proposed MI-PGEE method produced lower values of mean square error (MSE) and relative mean square error (RMSE) and higher sensitivity and specificity values in the MAR case with missingness covariates and response as soon as the covariates are correlated. The higher specificity of MI-PGEE means that the method can delete the unimportant covariates. However, as the correlation between covariates increases, the MI-PGEE can not select the important covariates. Therefore, an increase in the missing value rate across the dataset leads to similar results from the MI-PGEE.

In another study by Kowalski et al. (2018), SCAD is integrated with the weighted generalized estimating equation (WGEE) to provide a flexible selection of variables from the main and missing data modules to improve the fit validity of inferences. The WGEE consists of two modules: 1) the main module that models the relationship that involves the response variable, and 2) the missing data module that focuses on the MAR mechanism of missingness. The missing data module assumes that the covariate and the response are missing together. The authors' contribution is the penalized modern variable selection technique with the semi-parametric or distribution-free model (PWGEE). Also, they introduced a new formulation of the WGEE model to enable the joint inference for parameters in the two modules. As a result, WGEE provides a robust and efficient approach that analyzes the response variable when no possible parametric distribution is specified. Also, SCAD is more reliable and selects a

more robust subset of variables than the classical counterparts. Simulation results indicated that this new approach works well with moderate sample size and continuous and count response variables. For binary response, increasing the magnitude of the coefficients will improve the performance. This means that the strength of the relationship between the response and covariates in the main module will determine the performance of penalized weighted generalized estimating equation (PWGEE).

Ignorable dropouts are discussed frequently in the literature, and GEE is developed when the missingness is MAR. However, Wang and Ma (2021) proposed a novel methodology that accommodates the challenge of developing statistical analysis for non-ignorable dropouts, especially in the response variable. The longitudinal data is assumed to be balanced with the same cluster size. The authors' contribution in variable selection is by proposing the penalized empirical likelihood (PEL), where the profile EL is combined with SCAD to incorporate the possible dependence in longitudinal data. The tuning parameter is identified based on three BIC-type information criteria. Simulation findings showed that the PEL efficiently selected the significant variables and estimated parameters simultaneously. The developed methodology is based on quadratic inference function (QIF) and hybrid GEE methods for the non-ignorable monotone missing data pattern. Simulation results implied that the proposed variable selection method is satisfactory, and the selected models are very close to the true model, utilizing the extended BIC of Chen and Chen (2008) for selecting the tuning parameter. This algorithm can be implemented with the unbalanced longitudinal setting, where the data are measured with unequal cluster sizes.

Recently, Chen and Shen (2022) proposed a method under the semi-parametric weighted GEE framework for parameter estimation, called zero-inflated count information criterion (ZICIC). The proposed approach handles longitudinal count data

with an exceptional high percentage of zeros (i.e., zero-inflated) in the response. ZICIC accommodate longitudinal zero-inflated count data with nonmonotone or intermittent missing values in both response and covariates. The authors extended the weighted GEE framework to estimate the model parameters and the missing data model. ZICIC is a developed criterion to select an appropriate subset of covariates and the GEE mean model for the zero-inflated negative binomial models. However, numerical studies showed that when the zero proportion of responses and the missing data proportion increase, the ZICIC criterion select the true model with only 40% proportion. This result is expected due to the much loss of information for the cases under study. In fact, ZICIC criterion showed a robust performance when the missing data model is mis-specified.

LMM Based Methods

Regarding the variable selection with the missing data approach based on the linear mixed-effects models, Marino et al. (2017) indicated that complete case analysis is only applicable when the missing values are in the MCAR case; otherwise, biased results will arise. Hence, the authors proposed a method that fills the gap in the previous literature that performs variable selection when the missing data are multiply imputed. There is also a lack of clear and correct applying guidelines of variable selection on multiply imputed datasets in the two-levels linear mixed model (i.e., multilevel model). The methodology starts by performing *m* imputations to produce complete datasets; then, the *m* imputed datasets are stacked into a single wide complete dataset where *m* imputed variables present each covariate. For fully observed variables where no imputation is required to be performed, the stacked dataset should consist of *m* imputed columns, and the fully observed covariate is represented by only one column. Then a penalized likelihood method is utilized to perform variable selection on multiply imputed datasets via group LASSO. The simulation declared that the proposed

approach performed the best compared to only complete cases and ad hoc procedures. As the number of imputations increases, the smaller the model is produced. This methodology depicted the missingness in covariates, and the selection technique is on fixed effects only. It can also be adapted to panel data as a special case of multilevel data with only one level of clustering.

Incorporating incompleteness in the modeling process requires a good consideration of the nature of the missing data mechanism and its implications on the statistical inferences. The previously reviewed article assumed that the missing data were missing by design. Therefore, authors in this aspect presented powerful algorithms and data imputation combined with computing resources, providing a solution to the ignorable missing data mechanisms. The missing data mechanism can also be NMAR when the unobserved outcomes are the cause of describing the missing data process. A valid analysis is obtained by ignoring the missingness mechanism in the ignorable missing data. However, the missing value process in the NMAR case should explicitly be considered in the analysis for valid inferences.

Li et al. (2019) proposed a variable selection technique when missing data is in covariates and response variables with the NMAR mechanism and intermittent pattern—estimating a large set of nuisance parameters yields non-identifiability and model misspecification, which may lead to invalid inferences under the mixed model with the NMAR mechanism and a complicated missing data pattern. Therefore, a penalized composite likelihood (CL) method is utilized via SCAD, simultaneously estimating the parameters and selecting predictive covariates. The proposed methodology handles multilevel longitudinal data fitted using a GLMM. The developed composite likelihood framework can handle flexible missing data patterns, whether monotone or intermittent. Also, it accommodates missingness in covariates, response

variables, or both. Numerical results in simulation showed that CL yields little biases and has good computation efficiency under a 30% missingness rate.

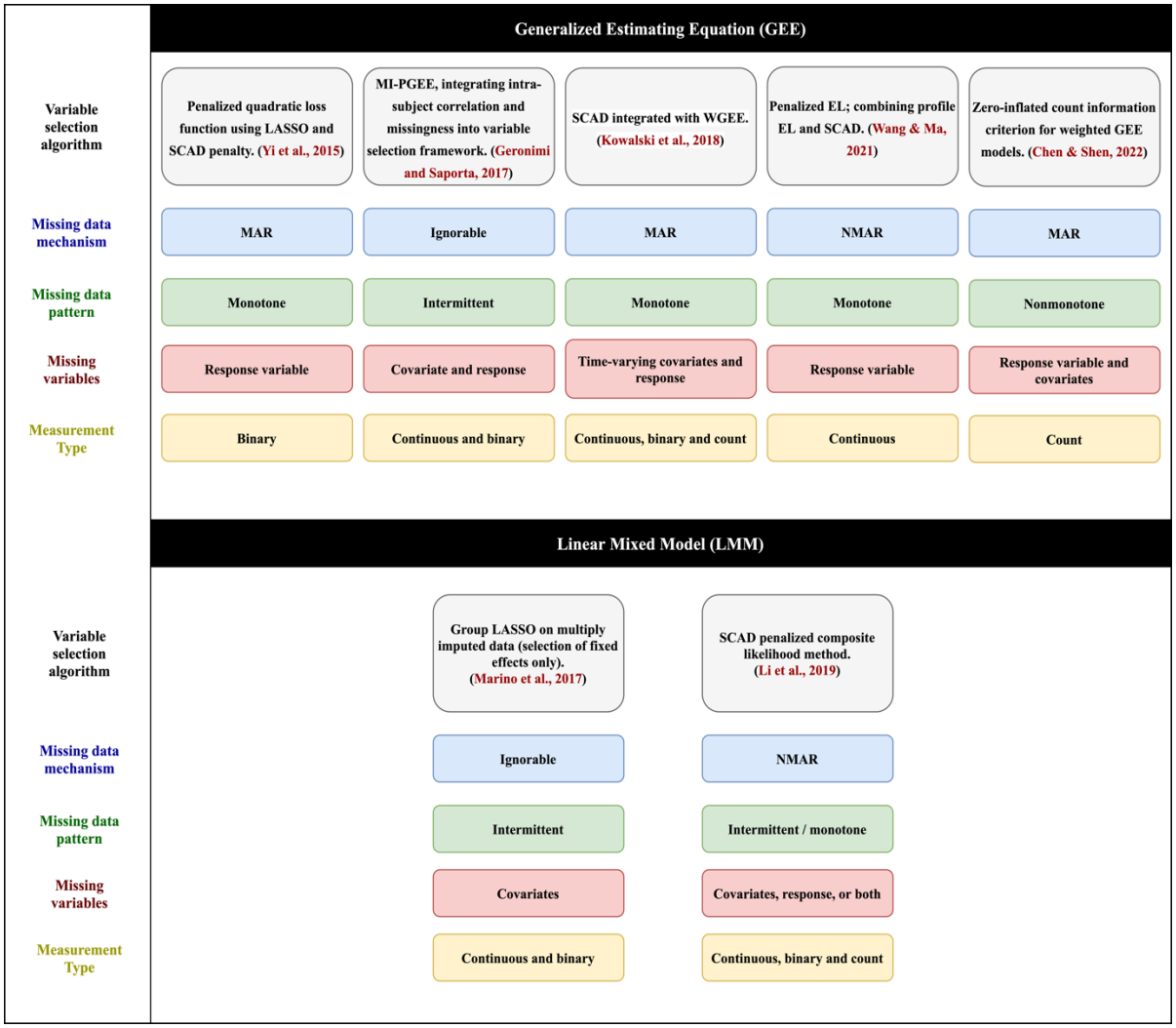

Figure 3. Variable Selection algorithms with incomplete longitudinal data in the literature.

CHAPTER 4: ADAPTIVE LASSO FOR VARIABLE SELECTION WITH MULTIPLY IMPUTED INCOMPLETE LONGITUDINAL DATA

Patients within hospitals, students within schools, or repeated observations on each individual over time arise correlation in analysis modeling. In general, repeated measurements per individual result in correlated errors that violate the assumption of classical regression models. Linear mixed models provide a general and flexible approach for analyzing data in these situations because the correlation patterns are widely varied and explicitly modeled through the utilization of random effects.

The mixed models' term refers to the utilization of both random and fixed effects in the same model of analysis. The variability of the population between a set of treatments in an experiment is explained by the fixed effects, while the random effects represent the variability among study individuals. The levels under the random effects are not of primary interest (in comparison to those under the fixed effects) but are thought to be the random selection from a larger set of levels. In other words, random effects are associated with subjects that are drawn randomly from a population. In some situations, mixed models are very close to hierarchical linear models where the hierarchy arises when individuals are at one level (upper-level units), and the measurements within these individuals are on another level (lower-level units). This multilevel situation can get more complicated as the number of levels becomes more than two. In fact, there is a correlation between lower-level units within the same upper-level unit. Also, these lower-level units have a variety of variance-covariance structures, for instance, diagonal, compound symmetry, Auto-Regressive, Toeplitz, and unstructured.

In other aspects, mixed models are called the model for repeated measurements

when it is used to analyze panel or longitudinal data. In longitudinal data, each time series form an individual cluster, which has two sources of variability: within and between clusters. Measurements on different clusters are not correlated, while measurements on the same cluster (i.e., within a cluster) are correlated with $\rho = \frac{var(b_i)}{var(b_i+\varepsilon_{ij})}$. $var(b_i)$ is the source of variation between clusters (inter), and $var(\varepsilon_{ij})$ is within-cluster (intra) variation. The correlation coefficient stated before indicates that the higher inter variation, the larger the correlation within each cluster. Generally speaking, ignoring the hierarchical or cluster structure can lead to false interpretations. Therefore, linear mixed models are well fitted for modeling repeated (clustered) data with multiple sources of variation (NCSS Statistical Software, 2020a; Demidenko, 2013).

Variable selection in linear mixed models is challenging due to the added random effects. The complexity of the model increases as the mean structure, and the covariance structure should be correctly identified. This challenge increases as the dimension of fixed and random effects increase. Zou (2006) stated that the LASSO

$$\hat{\beta}(LASSO) = \arg min_{\beta} ||\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x_j}\beta_j||^2 + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\ell_1 \text{penalty}} \quad (1)$$

is a regularization approach that uses the tuning parameter λ in the ℓ_1 penalty to perform variable selection and parameter estimation simultaneously. When the tuning parameter of prediction is optimal (i.e., sufficiently large), some coefficients are shrunk to exact zero, indicating that the non-zero coefficients are selected to be

in the model. However, the LASSO shrinkage produces biased estimates for large coefficients due to the many noise features included in the predictive model. The Adaptive LASSO is an enhanced version when the LASSO is an inconsistent variable selection technique in specific scenarios. Adaptive LASSO assigns different adaptive (data-dependent) weights for penalizing different coefficients in the ℓ_1 penalty.

$$\hat{\beta}(Adaptive\ LASSO) = \arg min_{\beta} ||\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x_j}\beta_j||^2 \quad + \underbrace{\lambda \sum_{j=1}^{p} w_j|\beta_j|}_{\ell_1 \text{penalty}} \qquad (2)$$

LASSO and Adaptive LASSO both are ℓ_1 penalization methods that pose a similar algorithm to compute the estimates.

The adaptive LASSO algorithm proposed by Pan and Shang (2018) outperforms other model selection methodologies, such as the Akaike Information Criterion, Generalized Information Criterion, and Mallows' Cp, because it is a computationally feasible and large dimension of parameters is involved.

In addition to the oracle properties possessed by the results generated from Pan and Shang (2018) adaptive LASSO method, this method is outperforming because of the separate selection of random and fixed effects that accommodate the distinct properties between random and fixed effects. The estimators used in the proposed approach are more robust to outliers (i.e., penalized restricted log-likelihood). Also, the selection of variables criteria 1) catch primary information, 2) require fewer iterations, 3) have simpler derivatives, and 4) convergence is computationally feasible and accurate.

Linear-Mixed Effects Model for Longitudinal Panel Data

In this part, the basic notations for linear mixed models that will be used for the subsequent sections will be provided. Response measurements in longitudinal studies are correlated within-subjects, where this challenge needs to be handled properly to get valid inferences and standard errors. Standard regression models assume independent observations that produce invalid parameter estimates in longitudinal settings. However, a complete model that includes the intra-subject correlation assumption should be adopted. The linear mixed model is a natural extension of the general linear models (GLMs) by the added random effects. Mixed models have several advantages in dealing with longitudinal data over the general linear models.

First, this approach can fit intricate covariance patterns by specifying the residual component's best variance-covariance structure and random effects, providing precise fixed effects estimates and standard errors. Moreover, LMMs assume that the variation in observations is caused by 1) variation within a subject and 2) variation between subjects. The within-subject variation measures the distance between repeated measurements taken on the same subject over time. On the other hand, the between-subject variation is the distance between measurements on different subjects. It is assumed that the between-subject variation is greater than the within-subject variation.

Let

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\varepsilon_i} \tag{3}$$

represents the general expression of the linear mixed-effects model with an unknown p-dimensional vector of fixed effects $\boldsymbol{\beta}$, and an unknown q-dimensional

vector of random (subject-specific) effects $\boldsymbol{b_i}$ following $N(\mathbf{0}, \sigma^2 \boldsymbol{D})$. Where $\boldsymbol{D}$ denotes the $(q \times q)$ general covariance matrix. The random variable $\boldsymbol{y_i}$ denotes the n_i-repeated-measurements vector of the continuous response for subject $i \in (1, \dots, n)$ measured at fixed time points $t_i = (t_{i1}, t_{i2}, \dots, t_{in_i})'$, that is $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \dots, y_{in_i})'$. The total number of observations is $N = \sum_{i=1}^{n} n_i$. The residual component n_i-dimensional vector $\boldsymbol{\varepsilon_i}$ follows independent $N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n_i})$. The LMM assumes that the residual components $\boldsymbol{\varepsilon_i}$ and the random effects $\boldsymbol{b_i}$ are independent of each other; $Cov(\boldsymbol{b_i}, \boldsymbol{\varepsilon_i}) = 0$. The $\boldsymbol{X_i}$ is a $(n_i \times p)$ design matrix of fixed effects and $\boldsymbol{Z_i}$ is a $(n_i \times q)$ design matrix of random effects for subject i. Generally, $\boldsymbol{Z_i}$ is chosen to be a sub-vector of $\boldsymbol{X_i}$, thus $q < p$. Conditional on the random effect, the response $\boldsymbol{y_i}$ follows $N(\boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i}, \sigma^2 \boldsymbol{I}_{n_i})$. Furthermore, the response $\boldsymbol{y_i}$ has a marginal distribution that follows multivariate $N(\boldsymbol{X_i\beta}, \sigma^2 \boldsymbol{V_i}(\theta))$, where $\boldsymbol{V_i}(\theta) = \boldsymbol{Z_i D Z_i}' + \boldsymbol{I}_{n_i}$. Linear mixed models utilized for the longitudinal situation have two subtypes, which depend on modeling the longitudinal data; the Covariance Pattern Model and the Random Coefficients Model. The following expression is the Random Coefficients model

$$\boldsymbol{y_i} = \beta_0 + \beta_1 \boldsymbol{x_i} + \beta_2 \boldsymbol{t_i} + \underbrace{b_{0i} + b_{1i}\boldsymbol{t_i}}_{between-subject} + \underbrace{\boldsymbol{\varepsilon_i}}_{within-subject} \quad (4)$$

where the measurement time is included in the LMM as a covariate with a corresponding slope, the slope will change with subjects; therefore, the model should fit each subject's separate intercept and slope. The subject intercept and slope are not independent and included as random effects in the model; that is $\boldsymbol{b_i} = (\boldsymbol{b_{0i}}, \boldsymbol{b_{1i}}) = (\begin{matrix} subject_i \\ (subject \times time)_i \end{matrix})$. The Random Coefficient models are usually

utilized when the relationship with time is of interest (Van Belle et al., 2004; Molenberghs & Verbeke, 2000; NCSS Statistical Software, 2020a; NCSS Statistical Software, 2020b; Huque et al., 2020).

Multiple Imputation

Frequently, longitudinal studies are riddled with missing data. However, statistical power, analytical options, confidence in results, and generalizability of the linear mixed model findings and variable selection may be compromised if longitudinal studies have missing values. Determining the appropriate method to address missing data depends on the missing data rate and assumption. Methods that address missing data in longitudinal studies are either non-stochastic or stochastic. Non-stochastic methods, like mean replacement, last value carried forward, regression imputation, and hot-deck assume the missing data pattern follows MCAR or MAR. While stochastic methods involve multiple imputation, creating new datasets where missingness is imputed with plausible values (Roberts et al., 2017).

In longitudinal settings, subjects may have missing observations at one time but have observations subsequent to a sparse structure of the data (intermittent), or subjects dropout permanently before the study completes (dropouts or attrition). The third pattern of missingness is mixing both the intermittent pattern followed by the permanent dropouts. MI is one of the modern techniques developed under the assumption that the data are missing at random (MAR) or possess the ignorability mechanism. The correct implementation of the MI approach demands the central feature that MI adopts; the compatibility between the model of interest and the imputation model for preserving the multilevel data structure (Lüdtke, Robitzsch & Grund, 2017).

The methodology starts by adopting the MI based on the newly available joint

modeling (JM) approach for incomplete multilevel data presented in Huque et al. (2020), which is extensively discussed by Goldstein et al. (2014). In specific, substantive-model-compatible joint modeling (JM-SMC) approach is adopted when both the covariates and response variable contain missing values. He (2010) explained that the joint modeling approach partition the observations based on their missing data patterns then specify a parametric multivariate density suitable to the type of data given using the appropriate prior distribution for the parameters. Imputations are drawn from these multivariate density sub-models. Simulation results in Huque et al. (2020) showed that the JM-SMC provides consistent estimates of the random intercepts and consistent estimates of random slopes when both the covariates and response variable have missingness under the assumption of normality; Gaussian random effects. Buuren et al. (2010) stated that FCS is an alternative to JM when no reasonable multivariate distribution of the missing data is provided. However, Quartagno and Carpenter's (2019) results showed that JM performs better than FCS when the latent normal variables are used with the JM as an extension to impute categorically or count data (i.e., non-normal data). Moreover, imputation using FCS cannot handle variables measured at higher levels in multilevel data structures.

Goldstein et al. (2014) extended a more efficient and easy parametric approach of multiple imputation; the joint modeling (JM) using Gibbs sampling through the Markov Chain Monte Carlo (MCMC) methods. The MI by joint modeling is carried out for any covariate with missing values by setting up an imputation model where the covariate is treated as a response. JM uses the joint posterior distribution for all the variables that have missingness and assumes the normality of variables. Non-normal variables can be handled using the latent normal approach through MCMC that links the different data types through a multivariate normal distribution or transformed into

normality before the imputation occurs. The original non-normal variables' imputed values will then be obtained on their original scale before the analysis. Therefore, Huque et al. (2020) used the JM approach to preserve congeniality. The JM-SMC joint imputation model is given by

$$\binom{\boldsymbol{x_i}}{\boldsymbol{y_i}} \Big| \boldsymbol{t_i} = (\boldsymbol{y_i}|\boldsymbol{x_i}, \boldsymbol{t_i}) \times (\boldsymbol{x_i}|\boldsymbol{t_i}) \tag{5}$$

which is compatible with the analysis model in (4), where $(\boldsymbol{x_i}|\boldsymbol{t_i}) = \beta_{0(x)} + \beta_{(x)}\boldsymbol{t_i} + \boldsymbol{b}_{0(x)i} + \boldsymbol{b}_{1(x)i}\boldsymbol{t_i} + \boldsymbol{\varepsilon_{(x)i}}$ is the marginal distribution of the covariates.

The Selection of Random Effects in the Linear-Mixed Models

The methodology here uses Adaptive LASSO for variable selection in the linear mixed model in a separate two-stage procedure, implementing profile log-likelihood and Newton-Raphson algorithm in each stage. Formulas in the following steps are based on the assumption that the random effects $\boldsymbol{b_i} \sim N(0, \sigma^2\boldsymbol{D})$, error term $\boldsymbol{\varepsilon_i} \sim N(0, \sigma^2\boldsymbol{I_{n_i}})$ and the response variable $\boldsymbol{y_i} \sim N(\boldsymbol{X_i\beta}, \sigma^2\boldsymbol{V}_i(\boldsymbol{\theta}))$, where $\boldsymbol{V}_i(\boldsymbol{\theta}) = \boldsymbol{I_{n_i}} + \boldsymbol{Z_i D Z_i'}$. It is further assumed that $\boldsymbol{\theta}$ is the vector of $q(q+1)/2$ unique variance components in the covariance matrix $\boldsymbol{D}$.

The restricted log-likelihood is utilized in selecting the random effects because it is preferable when the variance components' estimates are of interest. Generally, the restricted maximum likelihood (REML) estimators are less biased than the maximum likelihood (ML) estimators. Therefore, the restricted log-likelihood is given by

$$\ell_R(\boldsymbol{\theta}, \sigma^2_{REML}) = \ell_F\left(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\theta}, \sigma^2_{ML}\right) - \frac{1}{2}\log\left|\frac{1}{\sigma^2_{REML}}\sum_{i=1}^{n}\boldsymbol{X_i'V_i^{-1}X_i}\right| \tag{6}$$

Where $\ell_F(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\theta}, \sigma^2_{ML}) = -\frac{1}{2}\sum_{i=1}^{n} \log|\sigma^2_{ML}\boldsymbol{V_i}| - \frac{1}{2\sigma^2_{ML}}\sum_{i=1}^{n}(\boldsymbol{y_i} - \boldsymbol{X_i}\widetilde{\boldsymbol{\beta}})'\boldsymbol{V_i}^{-1}(\boldsymbol{y_i} - \boldsymbol{X_i}\widetilde{\boldsymbol{\beta}})$ is the log-likelihood function, $\widetilde{\boldsymbol{\beta}} = (\sum_{i=1}^{n} \boldsymbol{X_i'V_i^{-1}X_i})^{-1}(\sum_{i=1}^{n} \boldsymbol{X_i'V_i^{-1}y_i})$ is the ML estimator of $\boldsymbol{\beta}$. Maximizing equation (6) will produce the REML estimator of $\boldsymbol{\theta}$. By substituting the σ^2_{REML} and σ^2_{ML} in equation (6) by the estimators $\hat{\sigma}^2_{REML} = \frac{1}{N-p}\sum_{i=1}^{n}(\boldsymbol{y_i} - \boldsymbol{X_i\beta})'\boldsymbol{V_i}^{-1}(\boldsymbol{y_i} - \boldsymbol{X_i\beta})$ and $\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{i=1}^{n}(\boldsymbol{y_i} - \boldsymbol{X_i\beta})'\boldsymbol{V_i}^{-1}(\boldsymbol{y_i} - \boldsymbol{X_i\beta})$, the restricted profile log-likelihood is derived from the restricted log-likelihood as

$$P_R(\boldsymbol{\theta}) = -\frac{1}{2}\log|\sum_{i=1}^{n} \boldsymbol{X_i'V_i^{-1}X_i}| - \frac{1}{2}\sum_{i=1}^{n} \log|\boldsymbol{V}_i| - \frac{1}{2}(N-p)\log[\sum_{i=1}^{n}(\boldsymbol{y_i} - \boldsymbol{X_i}\widetilde{\boldsymbol{\beta}})'\boldsymbol{V_i}^{-1}(\boldsymbol{y_i} - \boldsymbol{X_i}\widetilde{\boldsymbol{\beta}})] \quad (7)$$

N is the total number of observations, and p is the total number of fixed parameters. Zou (2006) indicated that the adaptive LASSO asymptotically selects the true model. Hence, maximizing the penalized restricted profile log-likelihood and factorizing $\boldsymbol{\theta}$ as $(\boldsymbol{d}, \boldsymbol{\alpha})$, the penalized restricted profile log-likelihood function is given below after the adaptive LASSO is employed. $\boldsymbol{d}$ is the vector of the diagonal element of the covariance matrix **D,** and $\boldsymbol{\alpha}$ is the vector of upper off-diagonal elements of **D**.

$$Q_R(\boldsymbol{\theta}) = P_R(\boldsymbol{\theta}) - \lambda_{1n}\sum_{j=1}^{q} \omega_{1j}|\boldsymbol{d_j}| \quad (8)$$

λ_{1n} is the tuning parameter which controls the model complexity, $\boldsymbol{d_j}$ is the jth element in $\boldsymbol{d}$ and the adaptive weights $\boldsymbol{w_1} = (\omega_{11}, \omega_{12}, \dots, \omega_{1q})' = 1/|\widetilde{\boldsymbol{d}}|$. $\widetilde{\boldsymbol{d}}$ is the root-n

consistent estimator of $\boldsymbol{d}$. The values of $\widetilde{\boldsymbol{d}}$ will reflect the relative importance of the random effects and penalize any $\boldsymbol{d_j}$ to zero will control the inclusion and exclusion of the random effects. Newton-Raphson algorithm is then employed to maximize $Q_R(\boldsymbol{\theta})$ in equation (8) until convergence. The $\widehat{\boldsymbol{\theta}}$ is the penalized restricted profile log-likelihood estimator where the estimator $\widehat{\boldsymbol{V}}$ can be determined based on it.

The selection of the optimal tuning parameter λ is carried out using the BIC selection criterion because BIC showed outperformance in simulation studies and consistency compared to other selection criteria. The λ that has the minimum BIC is chosen as the optimal. BIC-type criteria for selecting λ_{1n} is given by

$$BIC_R = -2 \times P_R\left(\widehat{\boldsymbol{\theta}}\right) + \log(N) \times df_R \tag{9}$$

The Selection of Fixed Effects in the Linear-Mixed Models

The correct selection of the random effects leads to the appropriate selection and estimation of the fixed effects. For this aim, the proper selection of the covariance structure of the random effects is critical. Fixed effects are selected with the utility of the penalized log-likelihood, then the final model is determined. For the general linear mixed model equation (1), the log-likelihood is given by

$$\ell_F(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n} \log|\sigma^2 \boldsymbol{V_i}| - \frac{1}{2\sigma^2}\sum_{i=1}^{n} (\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})' \boldsymbol{V_i}^{-1} (\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta}) \tag{10}$$

By substituting the ML estimator of σ^2, $\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{i=1}^{n} (\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})' \boldsymbol{V_i}^{-1} (\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})$, in the log-likelihood, the profile log-likelihood can be obtained as

$$P_F(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log|\boldsymbol{V_i}| - \frac{N}{2}\log\{\sum_{i=1}^{n}\left(\boldsymbol{y_i} - \boldsymbol{X_i}\widetilde{\boldsymbol{\beta}}\right)'\boldsymbol{V}_i^{-1}\left(\boldsymbol{y_i} - \boldsymbol{X_i}\widetilde{\boldsymbol{\beta}}\right)\} \quad (11)$$

where $\widetilde{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ as given earlier. Dropping the constant terms in (11) after estimating the covariance matrix of the random effects $\widehat{\boldsymbol{V}}$, the profile log-likelihood is as follows

$$P_F(\boldsymbol{\beta}) = -\frac{N}{2}\log\{\sum_{i=1}^{n}\left(\boldsymbol{y_i} - \boldsymbol{X_i}\widetilde{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y_i} - \boldsymbol{X_i}\widetilde{\boldsymbol{\beta}})\} \quad (12)$$

Therefore, the penalized profile log-likelihood is given by

$$Q_F(\boldsymbol{\beta}) = P_F(\boldsymbol{\beta}) - \lambda_{2n}\sum_{j=1}^{p}\omega_{2j}|\beta_j| \quad (13)$$

The weights vector is data-dependent and chosen to be $w_2 = 1/|\widetilde{\boldsymbol{\beta}}|$ as it possesses optimal properties, where $\widetilde{\boldsymbol{\beta}}$ is the ML estimator of $\boldsymbol{\beta}$ using the estimated covariance matrix $\widehat{\boldsymbol{V}}$, $\widetilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n}\boldsymbol{X_i'}\widehat{\boldsymbol{V}}_i^{-1}\boldsymbol{X_i}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X_i'}\widehat{\boldsymbol{V}}_i^{-1}\boldsymbol{y_i}\right)$. λ_{2n} is the tuning parameter for the fixed effects selection, and optimal tuning parameter is carried out using the BIC-type criteria

$$BIC_F = -2 \times P_F\left(\widehat{\boldsymbol{\beta}}\right) + \log(N) \times df_F \quad (14)$$

Large penalties are assigned for less important covariates and small penalties for the most important covariates. Newton-Raphson algorithm is used to maximize the penalized equation in (14) until the converged value $\widehat{\boldsymbol{\beta}}$ is achieved, and the LMM can finally be identified.

For further details of this methodology, the dissertation of Pan (2016) provided the derivatives and procedures for the two-stage variable selection method.

Stacked (Homogeneous) Adaptive LASSO for Linear Mixed Model with Multiply Imputed Data

When there is missing data, variable selection is performed only on the complete cases in most cases. However, this approach is insufficient as it may bias Type I error rates if the missingness assumption is rather than MCAR. Under the assumption of MAR and MCAR, a large number of variable selection techniques in the presence of missingness are developed based on the combination of variable selection with the inverse probability weighting (IPW) or augmented IPW to handle missing data. Another group of statistical techniques is based on combining variable selection methods with the observed data likelihood for handling missing data.

This thesis adopts the combination of variable selection with imputation methods for handling missing data. Specifically, a combination of the joint modeling multiple imputation and the adaptive LASSO variable selection approach is introduced here in a longitudinal setting. Imputing missing data is attractive and straightforward because the process is separate from the subsequent variable selection analysis on imputed datasets. The performance of variable selection methods relies heavily on the choice of imputation methods and the careful construction of imputation models. Moreover, Zhao and Long (2017) declared that the literature encountered challenges of combining multiple imputation and variable selection techniques in a principled

framework to attain the final results of variable selection.

Hence, the aim is to adapt the existing adaptive LASSO via profile log-likelihood proposed by Pan and Shang (2018) to accommodate multiply imputed data by joint modeling multiple imputations. Wood et al. (2008) discussed a variety of adaptation procedures for the variable selection methods when applied to multiply imputed data. One of the approaches that showed a good performance is to use variable selection on each imputed dataset separately, then select predictors that appear in at least half of the models and combine the parameter estimates across the selected predictors. Another approach is to stack the multiply imputed data sets yielding a single large dataset of length $N \times m$, where N is the total sample size and m is the number of imputed datasets. Fitting models or applying variable selection on the artificially enlarged dataset yields valid parameter estimates but small standard errors. A fixed weight o_i is applied to all subjects in order to correct the standard errors. Therefore, the thesis here use weighted profile log-likelihood for the adaptive LASSO variable selection on the stacked imputed data sets. Du et al. (2022) highlighted that the approach of stacking imputed data set is appealing because no ad-hoc pooling is required to identify the final active set of selected variables. Also, stacking allows simultaneous variable selection, estimation, and generating of interpretable parameter estimates.

Stacking is referred to as a homogenous pooled objective function (penalized profile log-likelihood in the proposed method). The optimization is equivalent to fitting the penalized procedure on stacked imputed datasets to enforce uniform estimation and variable selection across all the imputed data sets. The optimization is straightforward and can be implemented using existing software. Stacked methods are fast and have small MSE values.

In the proposed combining approach, the equal weights ($o_i = \frac{1}{m}$) are introduced so the total weight for each subject in the stacked dataset sums up to 1. Wang (2001) and Xue et al. (2021) defined the weighted likelihood function as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(\boldsymbol{X}_i; \boldsymbol{\beta})^{o_i} \tag{15}$$

To be more specific, the weighted likelihood function for the linear mixed model given in equation (3) is given as

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}} \times \frac{1}{\sqrt{|\sigma^2 \boldsymbol{V}_i|}} \times e^{-\frac{1}{2\sigma^2}(\boldsymbol{y_i}-\boldsymbol{X_i\beta})' \boldsymbol{V}_i^{-1} (\boldsymbol{y_i}-\boldsymbol{X_i\beta})}\right)^{o_i} \tag{16}$$

Hence, the weighted restricted log-likelihood and weighted log-likelihood are derived as

$$\ell_{WR}(\boldsymbol{\theta}, \sigma^2_{REML}) = \ell_{WF}\left(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\theta}, \sigma^2_{ML}\right) - \frac{1}{2} \log \left| \frac{1}{\sigma^2_{REML}} \sum_{i=1}^{n} o_i \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X_i} \right| \tag{17}$$

$$\begin{aligned} \ell_{WF}(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = & -\frac{1}{2} \sum_{i=1}^{n} o_i \log|\sigma^2 \boldsymbol{V_i}| \\ & -\frac{1}{2\sigma^2} \sum_{i=1}^{n} o_i (\boldsymbol{y_i} - \boldsymbol{X_i\beta})' \boldsymbol{V}_i^{-1} (\boldsymbol{y_i} - \boldsymbol{X_i\beta}) \end{aligned} \tag{18}$$

The detailed algorithm for the homogeneous adaptive LASSO variable selection on imputed data is presented in Figure 4.

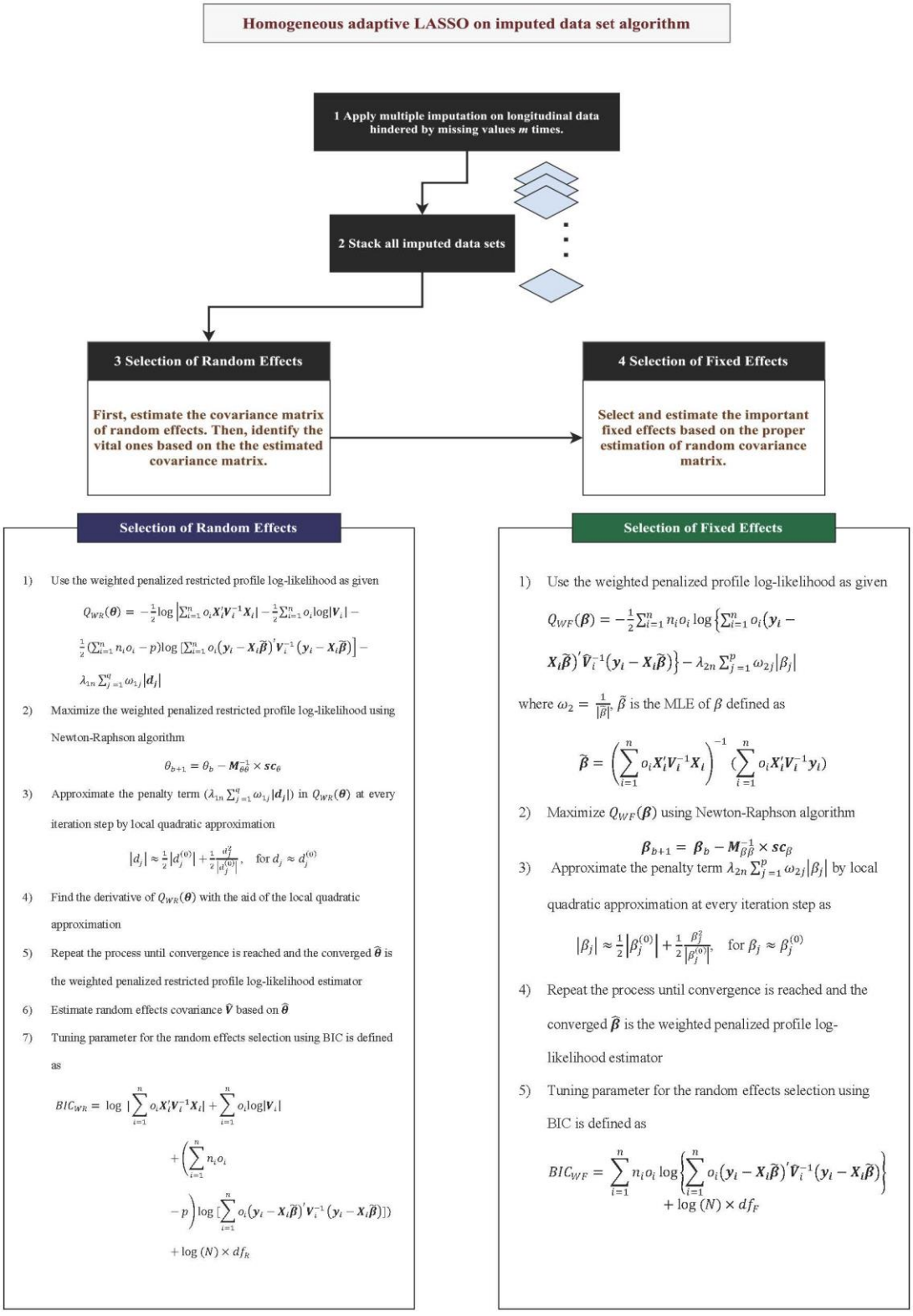

Figure 4. Stacked (homogeneuos) adaptive LASSO for linear mixed model on imputed dataset algorithm

CHAPTER 5: SIMULATION STUDIES

The simulation parts provide a numerical examination of the relative performance of the proposed procedure described in Chapter 4 in the setting of longitudinal data following Pan and Shang (2018) and Huque et al. (2020) examples. The simulation studies performed are particularly interested in comparing the proposed methodology's merits and performance under different design structures. Also, the model performance will be assessed by comparing the simulation results with the gold standard example in the literature in terms of the selection rates of the true model, fixed effects, and random effects.

Simulation 1: Model Performance under Different Scenarios

The simulation study here is based on a true model with $p = 9$ for fixed effects and $q = 2$ for random effects forming longitudinal data collected at equal intervals, where p is associated with the number of fixed effects and q is the number of random effects. The selection of q here is a special scenario drafted from the example given in Pan and Shange (2018). For this setting, the study consists of a large sample, where $n = 60$ independent subjects with $n_i = 10$ (waves) observations per each subject. The true parameter vector is specified to be $\boldsymbol{\beta} = (1, 1, 0, 0, 0, 0, 0, 0, 0)^T$, the true covariance matrix is $\boldsymbol{D} = \begin{pmatrix} 9 & 4.8 \\ 4.8 & 4 \end{pmatrix}$, and the variance σ^2 is assumed to be 1. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are independently generated from a uniform (-2, 2) distribution, except the first column of $\boldsymbol{Z}_i$ consisted of $\mathbf{1}'$s for the subject-specific intercept. The reason behind choosing a large sample here is that joint modeling imputation runs into difficulties with a smaller sample size (Van Buuren, 2018). The following cases are considered here:

Here, the aim is to evaluate the proposed procedure under different missing data proportions. missing values are induced on the fixed effects $\boldsymbol{X}_i$ and outcome variable $\boldsymbol{y}_i$ at each wave under an MAR mechanism with two different proportions: 25% and

35%. The random effects are fully observed because the random slopes in longitudinal settings are only associated with the time variable, which is usually fully observed.

Let R_{ij} indicate the missingness of $\boldsymbol{x}_{ij}$ and y_{ij}, where $R_{ij} = 1$ when the $\boldsymbol{x}_{ij}$ or y_{ij} is observed, and 0 if missing. The missing data indicator R_{ij} is selected from inverse-logistic regression with success probability given by the following models to induce the missing data in fixed effects and outcome variable, respectively,

$$logit\{\Pr(R_{1ij} = 1)\} = \theta_1 + \theta_2 \boldsymbol{z}_{2,ij} + \theta_3 y_{ij} \quad (19)$$

$$logit\{\Pr(R_{2ij} = 1)\} = \theta_1 + \theta_2 \boldsymbol{z}_{2,ij} + \theta_3 y_{ij} \quad (20)$$

$$logit\{\Pr(R_{3ij} = 1)\} = \theta_1 + \theta_2 \boldsymbol{z}_{2,ij} + \theta_3 y_{ij} \quad (21)$$

$$logit\{\Pr(R_{4ij} = 1)\} = \theta_1 + \theta_2 \boldsymbol{z}_{2,ij} + \theta_3 y_{ij} \quad (22)$$

$$logit\{\Pr(R_{5ij} = 1)\} = \theta_1 + \theta_2 \boldsymbol{z}_{2,ij} + \theta_3 y_{ij} \quad (23)$$

$$logit\{\Pr(R_{6ij} = 1)\} = \theta_1 + \theta_2 \boldsymbol{z}_{2,ij} + \theta_3 y_{ij} \quad (24)$$

$$logit\{\Pr(R_{7ij} = 1)\} = \theta_1 + \theta_2 \boldsymbol{z}_{2,ij} + \theta_3 y_{ij} \quad (25)$$

$$logit\{\Pr(R_{8ij} = 1)\} = \theta_1 + \theta_2 \boldsymbol{z}_{2,ij} + \theta_3 y_{ij} \quad (26)$$

$$logit\{\Pr(R_{9ij} = 1)\} = \theta_1 + \theta_2 \boldsymbol{z}_{2,ij} + \theta_3 y_{ij} \quad (27)$$

$$logit\{\Pr(R_{10ij} = 1)\} = \theta_4 + \theta_5 \boldsymbol{z}_{1,ij} + \theta_6 \boldsymbol{z}_{2,ij} \quad (28)$$

The coefficients $\boldsymbol{\theta} = (\theta_1, \dots, \theta_6)^T$ are chosen to control the proportions of missingness for each corresponding variable. Missing data are induced in a dropout and intermittent pattern.

Multiple imputation is applied for the constant covariance matrix described in Chapter 4. The R function `jomo.lmer` of the package `jomo` is used to multiply impute the missing data based on the joint posterior distribution of incomplete variables because it allows the mix of multilevel continuous and categorical variables (Quartagno, Grund, & Carpenter, 2019). Marino et al. (2017) results indicated that one imputation is not sufficient to get reliable results from the variable selection model. Hence, changes in the number of imputations is considered, where each simulated dataset is imputed $m = 5$ and $m = 10$ times before the proposed method is applied to perform variable selection and estimation. The number of between-imputation iterations is set to default, while the number of burn-in iterations is set to be $nburn = 5000$. Because JM imputation uses the MCMC algorithm for model fitting and imputation, it is important to monitor and assess the convergence of the sampler before registering the imputations. The trace plot in Figure 5 shows that a burn-in of 5000 is

reasonable where the sampler clearly converges. The function `jomo.lmer.MCMCchain` captures the state of the sampler as a starting value and provides the sampler a mechanism for the second set of iterations. Before running the imputation model, this dry run is recommended to check the sensible number of burn-in and the number of between imputations for the final imputation process.

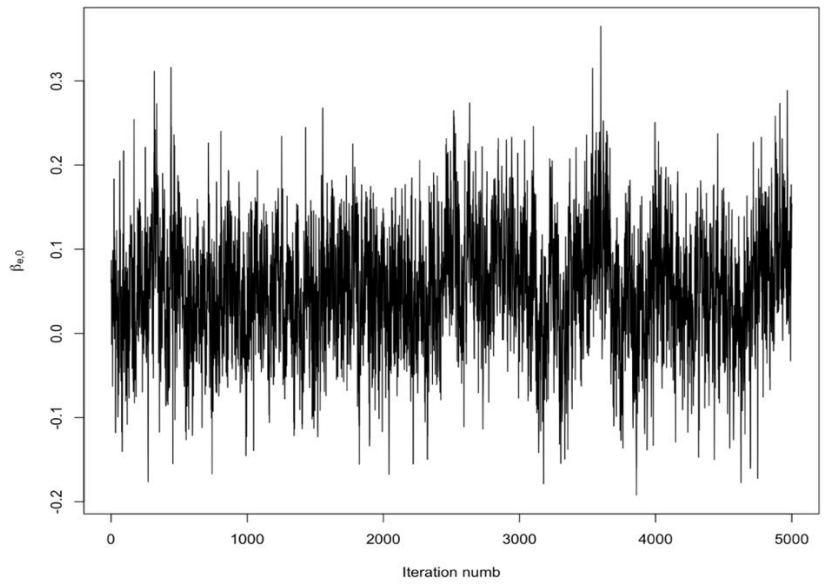

Figure 5. MCMC chain for $\beta_{e,0}$ to decide the number of burn-in and between-imputation iterations

The simulation studies here generated (simulated) 100 datasets and then reported the different measurements used to evaluate the variable selection performance. The performance of the proposed method is evaluated based on the different scenarios of missing data and number of imputed datasets. The measures "CR," "CF," and "C" are presented to evaluate the selection performance. For the selection of random effects, "CR" denotes the frequency (in percentage) that the correct random effects structure is selected. While "CF" shows the percentage that the correct fixed effects are selected, and "C" denotes the percentage that the correct true model is selected given that both

the random and fixed effects are identified correctly. The true values of the aforementioned measures are %CR = 100, %CF = 100, and %C = 100.

The simulated results across the 100 simulated data sets are displayed in Table 1. For fixed effects selection, %CF is dominated by five imputations with 25% missingness in each variable, which has CF equals 80%. It's observed that as the number of imputations increases, the sample size gets enlarged and the model performance rates are much smaller. Therefore, $m = 5$ imputations is adequate in the case of 25% missingness to avoid introducing very large artificial sample size to the variable selection algorithm which may lead to model instability. The %CR is always 100%, meaning the important random effects are always identified using the proposed selection method. The proposed method selects the true model with 78% in the case where the data is prone to 25% missingness in each variable and 10 imputations are generated.

Regarding the higher proportion of missingness (i.e., 35%), results from Table 1 remarked that the proposed variable selection method is performing better with a higher number of multiply imputed datasets. As the m increases, the method selects the correct fixed effects and true model with 50% in comparison to 45% in $m = 5$ scenario.

Table 1. Simulation Results for Simulation 1

Proportion of missingness (%)	Number of imputations	%CR	%CF	%C
25	m = 5	100	80	80
	m = 10	100	78	78
35	m = 5	100	45	45
	m = 10	100	50	50

Simulation 2: Proposed Model Comparison with Existing Literature

In Simulation 2, the selection performance is compared between the proposed approach in Chapter 4 and the variable selection when performed on the full simulated dataset before missing data is induced. Applying variable selection on full dataset is similar to the numerical studies given in Pan and Shange (2018) and is considered the gold standard because the penalized variable selection is implemented on the complete data.

Table 2 indicates that missing values present in any dataset lower the ability of the method to select the correct fixed effects and true model as compared to the benchmark method (i.e., variable selection on full data). However, the thesis proposed algorithm performs well at recovering the correct true model with $m = 5$ imputations.

Table 2. Simulation Results for Simulation 2 (missing data % = 25)

Method	%CR	%CF	%C
Full data without missingness	100	95	95
The proposed method	100	80	80

CHAPTER 6: DATA APPLICATION

Tracking changes over time at regular intervals is very useful way to collect data on complex topics in various research areas. For example, longitudinal genetic studies are more appropriate than measuring a single measurement per individual. Large amount of useful information (e.g., gene-time interactions) in longitudinal gene studies are lost when the existing traditional methodologies are used (Chung & Cho, 2022). As another example, Stamatis et al. (2022) declared that past studies investigating the potential factors for the development of psychological symptoms overtime in response to COVID-19 pandemic relied on cross-sectional designs. This makes it difficult to parse the timeline over which symptoms may unfold in relation to poorer mental health. Therefore, the study of Stamatis et al. (2022) examined the psychological impacts of COVID-19 among U.S. university students during the initial months of the pandemic using a novel variable selection method in longitudinal setting. Predictors used in their study included variables assessing the mental health symptoms, pandemic related experience, and sociodemographic characteristics. Dropouts following the MAR assumption in the data are handled using listwise deletion before implementing the variable selection analytic method, elastic net regression.

In most reviewed literature, longitudinal studies employs variable selection with missing data in health and related epidemiological sciences. Real data applications where the study units are other than real people can be considered too on variable selection with incomplete longitudinal data. For example, similar approaches like in Chapter 3 can be introduced to sustainability and quality control applications where study subjects are regions or appliances.

To further examine the effectiveness of the proposed procedure, the weighted adaptive LASSO penalization method is utilized in a demonstrative example of food

security and resilience (Singh, 2021). Premanandh (2011) stated that “Food is paramount to survival, growth, and reproduction of living organisms.” Food insecurity and hunger remain at unacceptable high levels despite the dramatic increase in production and availability. Under the shadow of the COVID-19 pandemic, world hunger increased in 2020, where between 720 and 811 million people are facing hunger. Of the global population, 12% is severely food insecure in 2020, representing 928 million people (FAO, 2021). Consequently, this leads to a higher rate of child mortality and increased criminal activities in a desperate bid to acquire food.

Food production, supply, and consumption represent one of the most profound environmental implications. In order to develop a sustainable policy, the world must incorporate the production of high-quality food with sufficient quantities that meet the current demand of the market and population.

The idea of food security was found since the First World War, where it defines how each European country has the ability to produce its own food and is vulnerable to political or military boycotts. Because of the food shortage crisis of 1972-1974, food security and sustainable agreements and scientific discussions were made at the World Food Conference in 1974. However, the focus of the conference was on producing safe and adequate food, not as a right for every human to have access to healthy food. This results in hunger and malnutrition due to the access problem, not the production. The right to food has to be put in the context of the right to life. Hence, the Food and Agriculture Organization in 1983 presented the new concept of food security which ensures secure access to the food offered. Additionally, the World Bank in 1986 stated that food security is the realization that everyone has the right to regular and permanent access to quality and sufficient quantity of food (Panzarini et al., 2013).

There is a strong correlation between food security and sustainability. The

projected population growth and the fundamental right for everyone to have sufficient and nutritious food will expand the production area or intensify the production practices. This additional pressure on agriculture delineates the direction of environmentally sustainable development, especially with regard to climate change, irrigation, and soil degradation. Hence, reducing the trade-off between food demand and the environment (Helms, 2004).

Brooks (2016) indicated that the sustainable development goals include linked objectives related to food and environment, which promote food security and sustainable agriculture. To increase the supply of food sustainability, increased investment in agriculture is needed. Food availability, improved nutrition, and lower unit cost will benefit the consumers when there is a broad-based development in agricultural productivity. Premanandh (2011) provided a review of the promising solutions in order to achieve food sustainability. One of the technological solutions is to commercial production of transgenic crops to boost food security and poverty reduction by offering a better yield in a shorter time.

In the State of Qatar, food security became a vital issue of national security. In the Food Systems Summit Dialogue of 2021, Qatar discussed its vision for sustainable food system by 2030. This dialogue shaped the actions and preparations of Qatar to fulfill this vision.

Despite of the harsh climate conditions and scarcity of natural resources, Qatar implemented various initiatives achieving close to a 100% supply for the most critical and essential food items on a day-to-day basis, as well as in times of emergencies and crises. Every year Qatar funds innovative research and projects in Food Security facilitating innovations in local food production and storage. All these actions are aimed to make sure that the food system in Qatar is resilient, equitable, and safe, whilst

protecting and improving the natural resources (Food Systems Summit Dialogues, 2021).

Data Description: 2020 Global Food Security Index (GFSI)

Measuring food security raises evidently the tool of composite indicators. Based on Jacobs (2004), Composite indicators are "useful tools for conveying summary performance information and signaling policy priorities." Missing data in composite indicators should not be ignored, and the variables should be normalized to facilitate comparison. Mainly, food security is based on the pillars: availability, affordability (i.e., access), utilization (i.e., quality and safety), and stability. The Economist Intelligence Unit (EIU) defined a Global Food Security Index (GFSI), one of several measurements of food insecurity at the country level, which will be the focus in this chapter. Other food security measures are also available in the literature; for instance, the International Food Policy Research Institute (IFPRI) developed the Global Hunger Index (GHI) (Izraelov & Silber, 2019).

GFSI is a composite indicator that is produced every year since 2012 to detect the progress towards food security at the national (i.e., country) level. The GFSI data set consists of $n = 110$ countries that are supposed to have a larger population so that the composite indicator would cover as much of regional diversity and economic importance for the world population as possible. The EIU constructs the index on the basis of 59 indicators that measure the issues (i.e., categories) of food security: Availability, Affordability, Quality and Safety, and Natural Resources and Resilience. The Availability category (16 indicators) measures the factors that impact food supply and the ease of access to food. The Affordability category (11 indicators) assesses the capacity of people to pay for food in each country and the cost of food under normal circumstances and at-time of food-related shocks. The Quality and Safety category (12

indicators) measures each country's nutritional quality and food safety environment. Finally, the Natural Resources and Resilience category (20 indicators) explores the country's exposure to the impacts of a changing climate and how the country is adapting to these risks. All the indicators under the four categories of food security are considered as covariates to fit the linear mixed model for the proposed method ($p = 59$). Appendix A gives the list of indicators under each food security category.

In what follows, an investigation of the proposed method on linear mixed model for the data when it is hindered by 25% missingness in each of the covariates and response, aiming to figure out the most appropriate mixed model for describing the factors that affect the continuous response variable of interest; GDP per capita (US$). The data set reported a cohort of $n = 110$ countries measured over $n_i = 9$ −time points, through 2012-2020, thus the total number of observations is $N = 990$.

Results

To fit the linear mixed model, the indicators under each food security category (listed in Appendix A) are considered covariates. The findings generated from the model will be generalized to all countries, thus, country (i.e., subject) is considered as a random factor. The LMM model could be expressed as

$$
\begin{aligned}
GDP \sim 1 + \ & \mathrm{AFF1.1} + \mathrm{AFF1.2} + \mathrm{AFF1.3} + \mathrm{AFF1.4} + \mathrm{AFF1.5.1} \\
& + \mathrm{AFF1.5.2} + \mathrm{AFF1.5.3} + \mathrm{AFF1.5.4} + \mathrm{AFF1.6.1} \\
& + \mathrm{AFF1.6.2} + \mathrm{AFF1.6.3} + \mathrm{AV2.1.1} + \mathrm{AV2.1.2} + \mathrm{AV2.2.1} \\
& + \mathrm{AV2.2.2} + \mathrm{AV2.3.1} + \mathrm{AV2.3.2} + \mathrm{AV2.3.3} + \mathrm{AV2.3.4} \\
& + \mathrm{AV2.4} + \mathrm{AV2.5.1} + \mathrm{AV2.5.2} + \mathrm{AV2.5.3} + \mathrm{AV2.5.4} \\
& + \mathrm{AV2.6} + \mathrm{AV2.7.1} + \mathrm{AV2.7.2} + \mathrm{QS3.1} + \mathrm{QS3.2.1} \\
& + \mathrm{QS3.2.2} + \mathrm{QS3.2.3} + \mathrm{QS3.2.4} + \mathrm{QS3.3.1} + \mathrm{QS3.3.2} \\
& + \mathrm{QS3.3.3} + \mathrm{QS3.4} + \mathrm{QS3.5.1} + \mathrm{QS3.5.2} + \mathrm{QS3.5.3} \\
& + \mathrm{NRR4.1.1} + \mathrm{NRR4.1.2} + \mathrm{NRR4.1.3} + \mathrm{NRR4.1.4} \\
& + \mathrm{NRR4.1.5} + \mathrm{NRR4.2.1} + \mathrm{NRR4.2.2} + \mathrm{NRR4.3.1} \\
& + \mathrm{NRR4.3.2} + \mathrm{NRR4.3.3} + \mathrm{NRR4.4.1} + \mathrm{NRR4.4.2} \\
& + \mathrm{NRR4.5.1} + \mathrm{NRR4.5.2} + \mathrm{NRR4.6.1} + \mathrm{NRR4.6.2} \\
& + \mathrm{NRR4.6.3} + \mathrm{NRR4.6.4} + \mathrm{NRR4.7.1} + \mathrm{NRR4.7.2} \\
& + \mathrm{Year} + (1 + \mathrm{Year} \mid \mathrm{Subject})
\end{aligned}
\tag{29}
$$

Where the subject is the country variable. The random term $(1 + \mathrm{Year} \mid \mathrm{Subject})$ include correlated intercept and slopes for the random intercept and Year.

Before doing the exploratory analysis and fitting the model, covariates in the data are with different measurement scales which needs a transformation to bring them to a comparable metric. Moeller (2015) reported that using standardized values in longitudinal studies is misleading. For example, z-standardization complicates interpretation of differences between groups. Other problems listed in Moeller (2015) article often co-occur because of the complexity of longitudinal data and analysis. As an alternative that do not change the multivariate distribution and covariance matrix of the transformed variables, proportion of maximum scale (POMS) which is also referred

as normalization. The following equation transform each scale to a metric from 0 to 1, where the higher value of a covariate indicates better situation.

$$z = \frac{x - Min(x)}{Max(x) - Min(x)} \tag{30}$$

In order to fit the data with linear mixed model and because the proposed method will be adopted to choose the best mixed model, then checking the (linearity, normality of residuals, and homoscedasticity of residual variance) assumptions is crucial to assure that the data is well-suited to the analysis. The model also assumes that response variable is normally distributed and shows homogeneous variance. It is observed in Figure 6 that the response variable, GDP, is highly skewed to the right. Therefore, the natural log transformation is used to attenuate this skewness (Figure 7) of GDP and is taken for the subsequent analysis rather than the raw GDP.

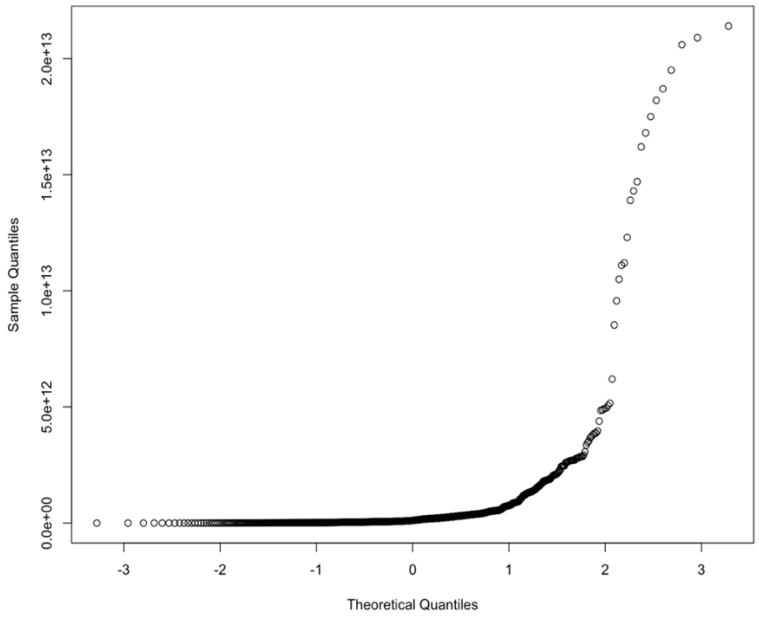

Figure 6. Q-Q plot of GDP

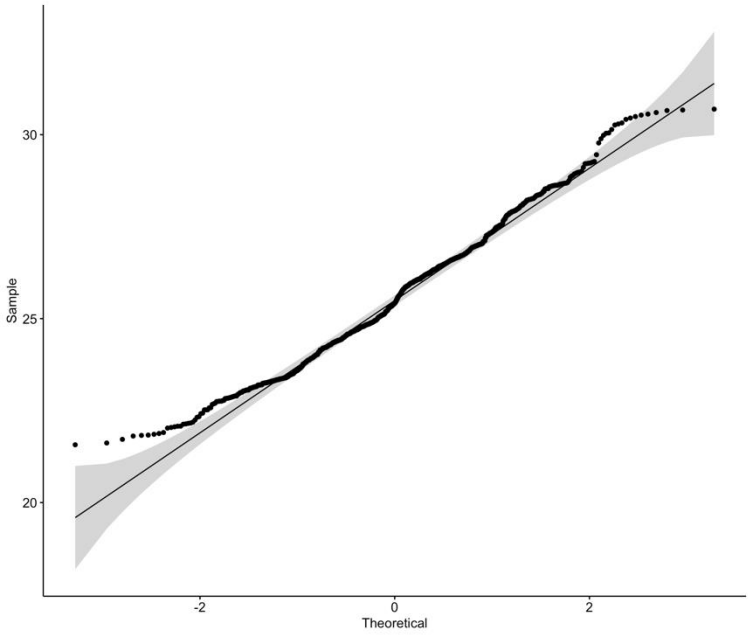

Figure 7. Q-Q plot of log transformation of GDP

To inspect the relationship between the GDP and the 59 covariates plus the time (i.e., year), first a simple mixed effects model is built to predict the GDP and control for the country level only, which is expressed by

$$GDP \sim (1 \mid \text{Subject}) \quad (31)$$

Then, another model is fitted by adding the covariates and year as expressed in equation 29. The log-likelihood ratio test used in Figure 8 to compare the two models (i.e., simple vs. Full) shows that the model with covariates improves on the simple model ($p < .05$). Also, to examine how these factors (i.e., covariates) help in explaining the variance in the response, the pseudo R-squared for the full model is equal to 0.998 (99.8%) revealing that the model is fitting well to the observed data.

```
Likelihood ratio test

Model 1: log(GDP) ~ (1 | Subject)
Model 2: log(GDP) ~ 1 + AFF1.1 + AFF1.2 + AFF1.3 + AFF1.4 + AFF1.5.1 +
    AFF1.5.2 + AFF1.5.3 + AFF1.5.4 + AFF1.6.1 + AFF1.6.2 + AFF1.6.3 +
    AV2.1.1 + AV2.1.2 + AV2.2.1 + AV2.2.2 + AV2.3.1 + AV2.3.2 +
    AV2.3.3 + AV2.3.4 + AV2.4 + AV2.5.1 + AV2.5.2 + AV2.5.3 +
    AV2.5.4 + AV2.6 + AV2.7.1 + AV2.7.2 + QS3.1 + QS3.2.1 + QS3.2.2 +
    QS3.2.3 + QS3.2.4 + QS3.3.1 + QS3.3.2 + QS3.3.3 + QS3.4 +
    QS3.5.1 + QS3.5.2 + QS3.5.3 + NRR4.1.1 + NRR4.1.2 + NRR4.1.3 +
    NRR4.1.4 + NRR4.1.5 + NRR4.2.1 + NRR4.2.2 + NRR4.3.1 + NRR4.3.2 +
    NRR4.3.3 + NRR4.4.1 + NRR4.4.2 + NRR4.5.1 + NRR4.5.2 + NRR4.6.1 +
    NRR4.6.2 + NRR4.6.3 + NRR4.6.4 + NRR4.7.1 + NRR4.7.2 + Year +
    (1 + Year | Subject)
  #Df LogLik Df  Chisq Pr(>Chisq)
1   3 103.92
2  65 498.39 62 788.95  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8. Log-likelihood ratio test for simple model (equation 31) vs. full model (equation 29)

For the model in equation 29, the density plot of the residuals is plotted in Figure 9 and indicates that normality assumption approximately holds with slight remaining skewness in the residuals. Also, the linearity assumption and variance homoscedasticity

are satisfied as shown in Figure 10.

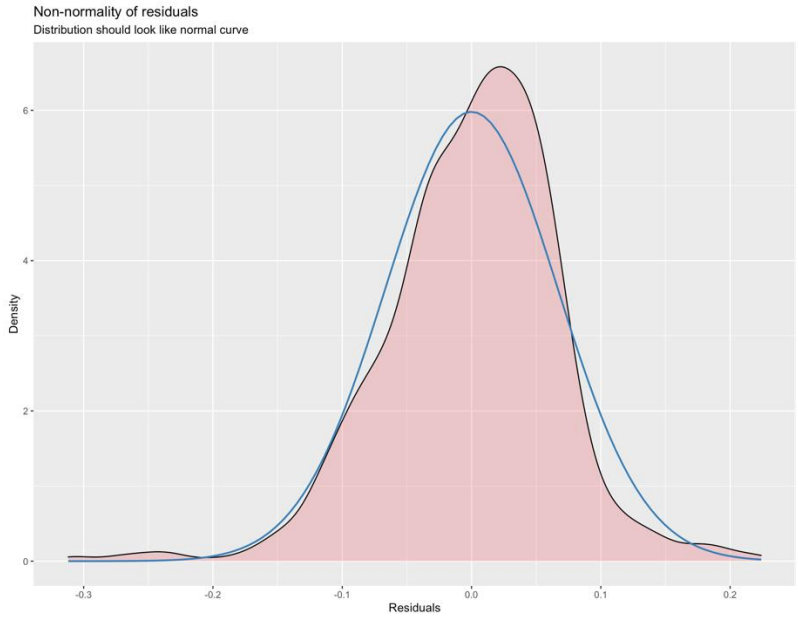

Figure 9. Density plot for the normality assumption of residuals

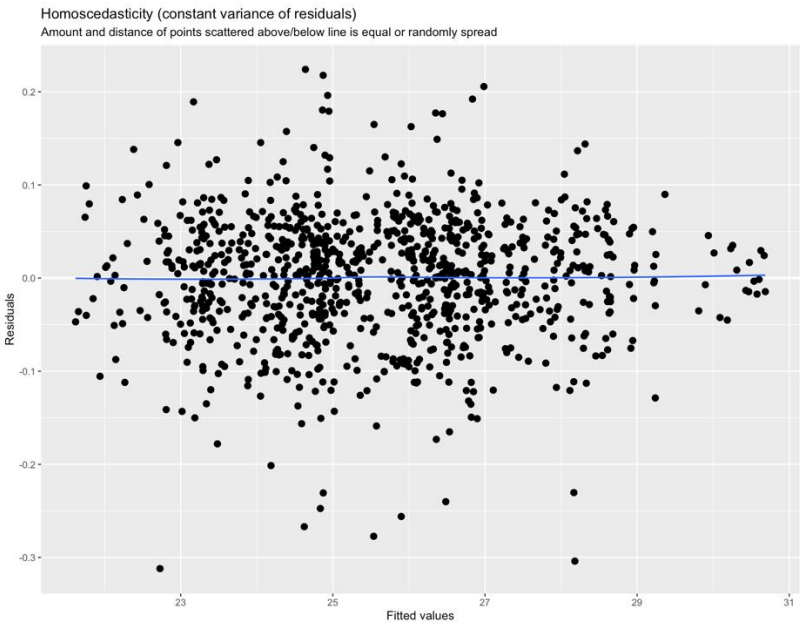

Figure 10. Residual plot against the fitted values for linearity assumption and variance homoscedasticity

The result for the proposed penalized method in Table 3 shows that the coefficients of 25 indicators are penalized to zero, meaning that these covariates are minor and thus excluded from the selected model. Based on the results in Table 3, the model can be shown as

$$\begin{aligned}\log(GDP) \sim \beta_{\circ} + \ & \beta_1 \mathrm{AFF1.2} + \beta_2 \mathrm{AFF1.3} + \beta_3 \mathrm{AFF1.4} + \beta_4 \mathrm{AFF1.5.3} \\ & + \beta_5 \mathrm{AFF1.5.4} + \beta_6 \mathrm{AFF1.6.1} + \beta_7 \mathrm{AV2.1.1} + \beta_8 \mathrm{AV2.3.2} \\ & + \beta_9 \mathrm{AV2.3.3} + \beta_{10} \mathrm{AV2.5.4} + \beta_{11} \mathrm{AV2.6} + \beta_{12} \mathrm{QS3.1} \\ & + \beta_{13} \mathrm{QS3.3.1} + \beta_{14} \mathrm{QS3.3.3} + + \beta_{15} \mathrm{QS3.5.2} \\ & + \beta_{16} \mathrm{QS3.5.3} + \beta_{17} \mathrm{NRR4.1.1} + \beta_{18} \mathrm{NRR4.1.2} \\ & + \beta_{19} \mathrm{NRR4.1.3} + \beta_{20} \mathrm{NRR4.1.4} + \beta_{21} \mathrm{NRR4.1.5} \\ & + \beta_{22} \mathrm{NRR4.2.1} + \beta_{23} \mathrm{NRR4.2.2} + \beta_{24} \mathrm{NRR4.3.1} \\ & + \beta_{25} \mathrm{NRR4.3.2} + \beta_{26} \mathrm{NRR4.3.3} + \beta_{27} \mathrm{NRR4.4.1} \\ & + \beta_{28} \mathrm{NRR4.4.2} + \beta_{29} \mathrm{NRR4.5.1} + \beta_{30} \mathrm{NRR4.5.2} \\ & + \beta_{31} \mathrm{NRR4.6.1} + \beta_{32} \mathrm{NRR4.6.2} + \beta_{33} \mathrm{NRR4.7.1} \\ & + \beta_{34} \mathrm{NRR4.7.2} + \beta_{35} \mathrm{Year} + b_1 + b_2 Year \end{aligned} \tag{30}$$

Table 3. Parameter estimates for indicator (fixed effect) coefficient using the proposed method

Fixed Effects	Coefficient	Penalized (Yes/No)?
AFF1.1	0.00	Yes
AFF1.2	1.90	No
AFF1.3	0.54	No
AFF1.4	1.80	No
AFF1.5.1	0.00	Yes
AFF1.5.2	0.00	Yes
AFF1.5.3	0.12	No
AFF1.5.4	0.09	No
AFF1.6.1	0.26	No
AFF1.6.2	0.00	Yes
AFF1.6.3	0.00	Yes

Fixed Effects	Coefficient	Penalized (Yes/No)?
AV2.1.1	0.97	No
AV2.1.2	0.00	Yes
AV2.2.1	0.00	Yes
AV2.2.2	0.00	Yes
AV2.3.1	0.00	Yes
AV2.3.2	0.02	No
AV2.3.3	1.37	No
AV2.3.4	0.00	Yes
AV2.4	0.00	Yes
AV2.5.1	0.00	Yes
AV2.5.2	0.00	Yes
AV2.5.3	0.00	Yes
AV2.5.4	2.00	No
AV2.6	0.45	No
AV2.7.1	0.00	Yes
AV2.7.2	0.00	Yes
QS3.1	3.05	No
QS3.2.1	0.00	Yes
QS3.2.2	0.00	Yes
QS3.2.3	0.00	Yes
QS3.2.4	0.00	Yes
QS3.3.1	1.39	No
QS3.3.2	0.00	Yes
QS3.3.3	4.25	No
QS3.4	0.00	Yes
QS3.5.1	0.00	Yes
QS3.5.2	2.27	No
QS3.5.3	3.24	No
NRR4.1.1	2.34	No
NRR4.1.2	2.03	No
NRR4.1.3	5.36	No
NRR4.1.4	1.19	No
NRR4.1.5	-0.44	No
NRR4.2.1	2.81	No
NRR4.2.2	2.32	No
NRR4.3.1	0.99	No
NRR4.3.2	2.13	No
NRR4.4.1	-1.72	No
NRR4.4.2	0.57	No
NRR4.5.1	1.54	No
NRR4.5.2	-0.45	No
NRR4.6.1	0.38	No
NRR4.6.2	0.45	No
NRR4.6.3	0.00	Yes
NRR4.6.4	0.00	Yes
NRR4.7.1	1.51	No
NRR4.7.2	0.90	No

CHAPTER 7: CONCLUDING REMARKS AND FUTURE DIRECTIONS

Repeated measurements or longitudinal data are often correlated. This correlation needs to be accounted for in the analysis to produce unbiased parameter estimates. Marginal and conditional models are appropriate for addressing this temporal and spatial proximity. Conditional models include random effects for the within-subject dependency, while marginal methods require additional modeling steps to handle these dependencies. GEE is a population-average (marginal) approach that uses quasi-likelihood equations for parameter estimation. GEE takes into account the dependency of measurements by specifying a "working" correlation structure. Khajeh-Kazemi et al. (2011) stated that the "working" correlation matrix should be identified to sufficiently fit the data; otherwise, the parameter estimates will be inefficient but consistent. The quasi-likelihood estimators are estimates of the GEE where no likelihood function is explicitly specified. Thus, the response's joint distribution is not specified completely; GEE has limitations on the goodness-of-fit test and has complexity with comparing and choosing the best model.

It is impossible to check the regression model's adequacy without proper model checking, and the validity of inference cannot be assured. Therefore, GEE is not a modeling technique, but it is an estimating method. However, empirical parameter estimates and standard errors can be attained. Empirical estimators are more variable and smaller (in absolute value) than estimators from conditional methods. Moreover, empirical standard errors are underestimated unless there is a large sample size (Slavkovic, 2018). Another limitation of GEE is the sensitivity of the link function, which can affect the model fit.

On the other hand, LMM is a subject-specific (conditional) framework. Assume a distribution of intercepts, and every subject intercept is a random variable to account

for the variability between subjects. This describes how the covariate affects the response within-cluster, holding all other covariates and random effects constant (Muff et al., 2016). A possible modification can be applied to the GLMM to ensure the method's robustness when the assumptions are not satisfied. For example, empirical standard errors are introduced within the GLMM when the correlation structure is mis-specified (Koper & Manseau, 2009). Unlike GEE, LMM can be extended to allow multiple levels (i.e., clusters) in the longitudinal data. Also, GLMMs rely on likelihood methods and can thus undergo model selection procedures. Mixed-effects models provide valid inferences when the missing data mechanism is ignorable, while GEE requires a stronger missing data assumption, MCAR. The interpretation of the conditional and marginal models is equivalent when the response variable is normally distributed. However, Lee and Nelder (2004) stated that the conditional model is fundamental where the marginal error can be made. They supported the use of robust procedures but with likelihood-based methods. All the robust procedures used in GEEs are also applicable for LMMs. Twisk (2004) and Locascio and Atri (2011) provided a concise overview of the two advanced analytical approaches; LMM and GEE. While Ballinger (2004) and Parzen et al. (2011) highlighted the cautions and drawbacks of using GEE in longitudinal settings.

Classical regression or ANOVA models are not suited for repeated measurements because they ignore the analysis results when even a single measurement is missing. On the contrary, mixed models have the advantage of handling uneven spaces of repeated measurements as long as the missingness follows the MAR assumption. Another advantage of LMM is that it can be extended to the non-normal response variable (i.e., generalized linear mixed models). Hence, linear mixed models provide a flexible and general tool for correlated data analysis.

In longitudinal studies, it is far more common and challenging that study participant (i.e., units) do not always appear for a scheduled observation leading to missing observations. Therefore, the data is unbalanced over time as not all individuals have the same number of measurements in a given set of time points. The term "incomplete" is used to distinguish the missingness from another kind of unbalanced longitudinal design, meaning a particular intended measurement could not be obtained. It is always recommended to have balanced longitudinal designs since these designs can capture within-individual change. Missing data and panel attrition is one of the most frequently encountered issues in longitudinal panel designs, which also have more ways to address them. Results revealed that multiple imputations using the JM-SMC approach hold great promise for imputing longitudinal data.

Due to the simpler form and concave optimization property of the Adaptive LASSO penalization method, the adopted method of variable selection is more robust to outliers and useful for simultaneous variable selection and parameter estimation.

Variable selection algorithms can be easily implemented in any statistical software package. For handling missing data, the widely used multiple imputation, using the JM approach, are also flexible and easy to implement and available in multiple software packages (e.g., `jomo` in R). However, incorporating variable selection and missing data algorithms need to be compiled in a way in which to give substantive inferential conclusions and unbiased estimates. Du et al. (2022) stated that there is a lack of methods that address variable selection with the presence of missing data in the literature. Stacking (i.e., homogeneous) and grouping (i.e., heterogeneous) are two appealing combination algorithms because they can handle variable selection given imputed datasets obtained previously from imputation software. Simulation findings in Du et al. (2022) indicated that the imputation-stacking objective function approach

tends to be more efficient and has better estimation and variable selection properties.

To investigate the behavior or performance of the proposed combination approach, the thesis conducted comparative simulation studies. Wood et al. (2008) indicated that applying variable selection on multiply imputed data sets may be computationally infeasible when there is a large data set (N) and large imputed datasets (m). Also, Roberts et al. (2017) declared that the acceptable rate of missing data in longitudinal studies varies from 5% to 20%. However, the thesis proposed method, which tested a longitudinal data of $N = 600$ observations and hindered by 25% of missingness in each covariate and response variable performs quite efficaciously in selecting the correct true model (80%).

As a subject for future research, one could use the Bayesian variable selection for linear mixed models when longitudinal data is hindered by missingness. Limited work has been done in this direction, especially with mixed models (Zhao & Long, 2017). Bayesian variable selection provides a convenient approach when the covariates are large, as the standard selection methods are infeasible because they choose the preferred model by fitting all the possible models. Yang et al. (2005) proposed a fully Bayesian framework applied on multiply imputed data, and recent research by Beesley and Taylor (2021) proposed a shrinkage-based Bayesian variable selection technique. The former research study used linear regression, and the latter used a multistate modeling approach. However, another direction indicates an extension of the methodology proposed by Yang et al. (2020) to accommodate missing values handled by multiple imputation.

REFERENCES

Allison, P. D. (2001). *Missing data*. Sage publications.

Allison, P. D. (2009) Missing Data. In R. E. Millsap & A. Maydeu-Olivares. *The SAGE Handbook of Quantitative Methods in Psychology* (pp. 72-89). Thousand Oaks, CA: Sage Publications Inc.

Allison, P. D. (2012). Handling missing data by maximum likelihood. SAS global forum, 312, 1–21.

Arnau, J., Bono, R., Blanca, M. J., & Bendayan, R. (2012). Using the linear mixed model to analyze non-normal data distributions in longitudinal designs. Behavior research methods, 44(4), 1224-1238.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. Organizational research methods, 7(2), 127-150.

Beesley, L. J., & Taylor, J. M. (2021). Bayesian variable selection and shrinkage strategies in a complicated modelling setting with missing data: A case study using multistate models. *Statistical Modelling, 21*(1-2), 11-29.

Bhaskaran, K., & Smeeth, L. (2014). What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology, 43*(4), 1336-1339. doi:10.1093/ije/dyu080

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, *37*(4), 373-384.

Brooks, J. (2016). Food security and the sustainable development goals.

Buuren, S. V. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*(3), 219-242. doi:10.1177/0962280206074463

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by

chained equations in R. *Journal of statistical software*, 1-68.

Carlier, B. E., Schuring, M., Lötters, F. J., Bakker, B., Borgers, N., & Burdorf, A. (2013). The influence of re-employment on quality of life and self-rated health, a longitudinal study among unemployed persons in the Netherlands. BMC Public Health, 13(1), 1-7.

Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. Journal of Statistical Planning and Inference, 63(1), 39-54.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759-771.

Chen, C. S., & Shen, C. W. (2022). Distribution-free model selection for longitudinal zero-inflated count data with missing responses and covariates. *Statistics in Medicine*.

Chen, Q., & Wang, S. (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine, 32*(21), 3646-3659. doi:10.1002/sim.5783

Chen, X., & Yin, J. (2022). Simultaneous variable selection and estimation for longitudinal ordinal data with a diverging number of covariates. *AIMS Mathematics*, *7*(4), 7199-7211.

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family medicine and community health*, *8*(1).

Chung, W., & Cho, Y. (2022). Bayesian mixed models for longitudinal genetic data: theory, concepts, and simulation studies. *Genomics & informatics*, *20*(1).

Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to

analyse unbalanced repeated measures and longitudinal data. Statistics in medicine, 16(20), 2349-2380.

Demidenko, E. (2013). *Mixed models: Theory and applications with R*. Wiley.

Du, J., Boss, J., Han, P., Beesley, L. J., Kleinsasser, M., Goutman, S. A., ... & Mukherjee, B. (2022). Variable selection with multiply-imputed datasets: choosing between stacked and grouped methods. *Journal of Computational and Graphical Statistics*, (just-accepted), 1-35.

Enders, C. K. (2011). Analyzing longitudinal data with missing values. *Rehabilitation Psychology*, *56*(4), 267-288. doi:10.1037/a0025579

Epprecht, C., Guegan, D., Veiga, Á., & Correa da Rosa, J. (2021). Variable selection and forecasting via automated methods for linear models: LASSO/adaLASSO and Autometrics. *Communications in Statistics-Simulation and Computation*, *50*(1), 103-122.

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101.

FAO. (2021). *The State of Food Security and Nutrition in the World*. Organization of the United Nations, Rome.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.

Food Systems Summit Dialogues. (2021). *Towards a resilient, sustainable, equitable and healthy food system by 2030*. Food Systems Summit Dialogues. Retrieved from https://summitdialogues.org/wp-content/uploads/2021/09/20210921_The-Qatar-Pathway-to-a-Resilient-Sustainable-Equitable-Healthy-Food-System-by-2030-DRAFT.pdf

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of*

computational and graphical statistics, *7*(3), 397-416.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, *95*(452), 1304-1308.

Geronimi, J., & Saporta, G. (2017). Variable selection for multiply-imputed data with penalized generalized estimating equations. *Computational Statistics & Data Analysis*, *110*, 103-114. doi:10.1016/j.csda.2017.01.001

Gokalp Yavuz, F., & Arslan, O. (2019). Variable selection in elliptical linear mixed model. *Journal of Applied Statistics*, *47*(11), 2025-2043. https://doi.org/10.1080/02664763.2019.1702928

Goldstein, H., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 553-564.

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American journal of public health*, *79*(3), 340-349.

He, Y. (2010). Missing data analysis using multiple imputation: getting to the heart of the matter. *Circulation: Cardiovascular Quality and Outcomes*, *3*(1), 98-105.

Helms, M. (2004). Food sustainability, food security and the environment. *British Food Journal*.

Huque, M. H., Moreno-Betancur, M., Quartagno, M., Simpson, J. A., Carlin, J. B., & Lee, K. J. (2020). Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model. *Biometrical Journal*, *62*(2), 444-466.

Izraelov, M., & Silber, J. (2019). An assessment of the global food security index. *Food Security*, *11*(5), 1135-1152.

Jacobs, R., P. Smith andM. Goddard (2004) Measuring Performance: An examination of composite performance indicators. Technical paper series 29, Center for Health Economics, University of yok, York, UK.

Khajeh-Kazemi, R., Golestan, B., Mohammad, K., Mahmoudi, M., Nedjat, S., & Pakravan, M. (2011). Comparison of generalized estimating equations and quadratic inference functions in superior versus inferior Ahmed glaucoma valve implantation. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences, 16*(3), 235.

Koper, N., & Manseau, M. (2009). Generalized estimating equations and generalized linear mixed-effects models for modelling resource selection. *Journal of Applied Ecology*, 590-599.

Kowalski, J., Hao, S., Chen, T., Liang, Y., Liu, J., Ge, L., ... & Tu, X. M. (2018). Modern variable selection for longitudinal semi-parametric models with missing data. *Journal of Applied Statistics*, *45*(14), 2548-2562.

Lee, K., & Chen, R. (2019). Bayesian variable selection in a finite mixture of linear mixed-effects models. *Journal of Statistical Computation and Simulation*, *89*(13), 2434-2453. https://doi.org/10.1080/00949655.2019.1620746

Lee, Y., & Nelder, J. A. (2004). Conditional and marginal models: another view. *Statistical Science*, *19*(2), 219-238.

Li, H., Shu, D., He, W., & Yi, G. Y. (2019). Variable selection via the composite likelihood method for multilevel longitudinal data with missing responses and covariates. *Computational Statistics & Data Analysis*, *135*, 25-34. doi:10.1016/j.csda.2019.01.011

Li, E., Zhang, D., & Davidian, M. (2004). Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for

longitudinal measurements. Biometrics, 60(1), 1-7.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Liu, Y., Wang, Y., Feng, Y., & Wall, M. M. (2016). Variable selection and prediction with incomplete high-dimensional data. *The annals of applied statistics*, *10*(1), 418.

Locascio, J. J., & Atri, A. (2011). An overview of longitudinal data analysis methods for neurological research. Dementia and geriatric cognitive disorders extra, 1(1), 330-357.

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological methods*, *22*(1), 141.

Marino, M., Buxton, O. M., & Li, Y. (2017). Covariate selection for multilevel models with missing data. *Stat*, *6*(1), 31-46. doi:10.1002/sta4.133

Miller, M. E., Davis, C. S., & Landis, J. R. (1993). The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares. Biometrics, 1033-1044.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, *83*(404), 1023-1032.

Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. Chichester: John Wiley & Sons.

Muff, S., Held, L., & Keller, L. F. (2016). Marginal or conditional regression models for correlated non-normal data?. *Methods in Ecology and Evolution*, *7*(12), 1514-1524.

NCSS Statistical Software. (2020a) Mixed Models – Repeated Measures. NCSS Documentation (Chapter 222). Retrieved from

https://www.ncss.com/wpcontent/themes/ncss/pdf/Procedures/NCSS/Mixed_Models-Repeated_Measures.pdf

Ni, X., Zhang, D., & Zhang, H. H. (2010). Variable Selection for Semiparametric Mixed Models in Longitudinal Studies. *Biometrics, 66*(1), 79-88. doi:10.1111/j.1541-0420.2009.01240.x

Noda, K., Arakawa, H., Kimura-Ono, A., Yamazaki, S., Hara, E. S., Sonoyama, W., ... & Kuboki, T. (2015). A longitudinal retrospective study of the analysis of the risk factors of implant failure by the application of generalized estimating equations. Journal of prosthodontic research, 59(3), 178-184.

Pan, J. (2016). *Adaptive LASSO for mixed model selection via profile log-likelihood* (Doctoral dissertation, Bowling Green State University).

Pan, J., & Shang, J. (2018). Adaptive LASSO for linear mixed model selection via profile log-likelihood. *Communications in Statistics-Theory and Methods*, *47*(8), 1882-1900.

Panzarini, N. H., Matos, E., & Bittencourt, J. (2013). Food Security and Sustainability: efficient production and innovations considering the environment . *Global Advanced Research Journal of Food Science and Technology*, *2*(2), 019–022.

Parzen, M., Ghosh, S., Lipsitz, S., Sinha, D., Fitzmaurice, G. M., Mallick, B. K., & Ibrahim, J. G. (2011). A generalized linear mixed model for longitudinal binary data with a marginal logit link function. The annals of applied statistics, 5(1), 449.

Pitchiah, R., Rooba, S., & Kumar, U. (2021). Missing Data Imputation using Multiple Imputation with Adaptive LASSO for Parkinson's Disease Data. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(13), 5311-5327.

Premanandh, J. (2011). Factors affecting food security and contribution of modern technologies in food sustainability. *Journal of the Science of Food and*

Agriculture, *91*(15), 2707-2714.

Quartagno, M., & Carpenter, J. R. (2019). Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical journal. Biometrische Zeitschrift*, *61*(4), 1003–1019. https://doi.org/10.1002/bimj.201800222

Quartagno, M., Grund, S., & Carpenter, J. (2019). Jomo: a flexible package for two-level joint modelling multiple imputation. *R Journal*, *9*(1).

Roberts, M. B., Sullivan, M. C., & Winchester, S. B. (2017). Examining solutions to missing data in longitudinal nursing research. *Journal for Specialists in Pediatric Nursing*, *22*(2), e12179.

Shen, C. W., & Chen, Y. H. (2013). Model selection of generalized estimating equations with multiply imputed longitudinal data. Biometrical Journal, 55(6), 899-911.

Singh, P. (2021, February). *The Global Food Security index (GFSI)*. The Economist. Retrieved from https://impact.economist.com/sustainability/project/food-security-index/

Slavkovic, A. (2018). Lesson 12: Advanced Topics I - Generalized Estimating Equations (GEE) [Online statisticl program]. Retrieved from the Pennsylvania State University STAT 504 PennState Eberly College of Science site.

Stamatis, C. A., Broos, H. C., Hudiburgh, S. E., Dale, S. K., & Timpano, K. R. (2022). A longitudinal investigation of COVID-19 pandemic experiences and mental health among university students. *British Journal of Clinical Psychology*, *61*(2), 385-404.

Taavoni, M., & Arashi, M. (2022). Estimation in multivariate linear mixed models for longitudinal data with multiple outputs: Application to PBCseq data analysis. *Biometrical Journal*, *64*(3), 539-556.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the*

Royal Statistical Society: Series B (Methodological), *58*(1), 267-288.

Torero, M. (2014, October 15). *Food security brings economic growth — not the other way around*. International Food Policy Research Institute (IFPRI). Retrieved from https://www.ifpri.org/blog/food-security-brings-economic-growth-not-other-way-around.

Twisk, J. W. (2004). Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. *European journal of epidemiology*, *19*(8), 769-776.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. New York, NY: Springer.

Wang, S. X. (2001). *Maximum weighted likelihood estimation*(Doctoral dissertation, Ph. D. Thesis, University of British Columbia. 2001. Available from: https://open. library. ubc. ca/cIRcle/collections/ubctheses/831/items/1.0090880).

Wang, L., & Ma, W. (2021). Improved empirical likelihood inference and variable selection for generalized linear models with longitudinal nonignorable dropouts. Annals of the Institute of Statistical Mathematics, 73(3), 623-647.

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data?. *Statistics in medicine*, *27*(17), 3227-3246.

Wu, H., & Jones, M. P. (2021). Proportional likelihood ratio mixed model for discrete longitudinal data. Statistics in Medicine, 40(9), 2272-2285.

Xue, X., Lu, J., & Zhang, J. (2021). Item-Weighted Likelihood Method for Measuring Growth in Longitudinal Study With Tests Composed of Both Dichotomous and Polytomous Items. *Frontiers in Psychology*, *12*.

Yang, X., Belin, T. R., & Boscardin, W. J. (2005). Imputation and Variable Selection

in Linear Regression Models with Missing Covariates. *Biometrics, 61*(2), 498-506. doi:10.1111/j.1541-0420.2005.00317.x

Yang, M., Wang, M., & Dong, G. (2020). Bayesian variable selection for mixed effects model with shrinkage prior. *Computational Statistics, 35*(1), 227-243.

Yi, G. Y., Tan, X., & Li, R. (2015). Variable selection and inference procedures for marginal analysis of longitudinal data with missing observations and covariate measurement error. Canadian Journal of Statistics, 43(4), 498-518.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 1049-1060.

Zhao, Y., & Long, Q. (2017). Variable selection in the presence of missing data: Imputation-based methods. *Wiley Interdisciplinary Reviews: Computational Statistics, 9*(5). doi:10.1002/wics.1402

Zheng, X., Fu, B., Zhang, J., & Qin, G. (2018). Variable selection for longitudinal data with high-dimensional covariates and dropouts. *Journal of Statistical Computation and Simulation, 88*(4), 712-725.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association, 101*(476), 1418-1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology), 67*(2), 301-320.

APPENDIX A: LIST OF VARIABLES TAKEN INTO ACCOUNT BY THE DATASET OF GFSI

A complete description if the data can be found on The Economist Impact website: https://impact.economist.com/sustainability/project/food-security-index/. The data used here is a subset consisting of 990 observations from 110 countries from 2012 to 2020. Here is a list describing the indicators used as covariates in the model.

Affordability Category	
$x1$	1.1 Change in average food costs (**AFF1.1**)
$x2$	1.2 Proportion of population under global poverty line (**AFF1.2**)
$x3$	1.3 Inequality-adjusted income index (**AFF1.3**)
$x4$	1.4 Agricultural import tariffs (**AFF1.4**)
$x5$	1.5.1 Presence of food safety net programmes (**AFF1.5.1**)
$x6$	1.5.2 Funding for food safety net programmes (**AFF1.5.2**)
$x7$	1.5.3 Coverage of food safety net programmes (**AFF1.5.3**)
$x8$	1.5.4 Operation of food safety net program (**AFF1.5.4**)
$x9$	1.6.1 Access to finance and financial products for farmers (**AFF1.6.1**)
$x10$	1.6.2 Access to diversified financial products (**AFF1.6.2**)
$x11$	1.6.3 Access to market data and mobile banking (**AFF1.6.3**)
Availability Category	
$x12$	2.1.1 Food supply adequacy (**AV2.1.1**)
$x13$	2.1.2 Dependency on chronic food aid (**AV2.1.2**)
$x14$	2.2.1 Public expenditure on agricultural research and development (**AV2.2.1**)
$x15$	2.2.2 Access to agricultural technology, education and resources (**AV2.2.2**)
$x16$	2.3.1 Crop storage facilities (**AV2.3.1**)
$x17$	2.3.2 Road infrastructure (**AV2.3.2**)
$x18$	2.3.3 Air, port and rail infrastructure (**AV2.3.3**)
$x19$	2.3.4 Irrigation infrastructure (**AV2.3.4**)
$x20$	2.4 Volatility of agricultural production (**AV2.4**)
$x21$	2.5.1 Armed conflict (**AV2.5.1**)
$x22$	2.5.2 Political stability risk (**AV2.5.2**)
$x23$	2.5.3 Corruption (**AV2.5.3**)
$x24$	2.5.4 Gender inequality (**AV2.5.4**)
$x25$	2.6 Food loss (**AV2.6**)
$x26$	2.7.1 Food security strategy (**AV2.7.1**)
$x27$	2.7.2 Food security agency (**AV2.7.2**)
Quality and Safety Category	
$x28$	3.1 Dietary diversity (**QS3.1**)
$x29$	3.2.1 National dietary guidelines (**QS3.2.1**)
$x30$	3.2.2 National nutrition plan or strategy (**QS3.2.2**)
$x31$	3.2.3 Nutrition labeling (**QS3.2.3**)
$x32$	3.2.4 Nutrition monitoring and surveillance (**QS3.2.4**)
$x33$	3.3.1 Dietary availability of vitamin A (**QS3.3.1**)
$x34$	3.3.2 Dietary availability of iron (**QS3.3.2**)

$x35$	3.3.3 Dietary availability of zinc (**QS3.3.3**)
$x36$	3.4 Protein quality (**QS3.4**)
$x37$	3.5.1 Food safety mechanisms (**QS3.5.1**)
$x38$	3.5.2 Access to drinking water (**QS3.5.2**)
$x39$	3.5.3 Ability to store food safely (**QS3.5.3**)
Natural Resources and Resilience Category	
$x40$	4.1.1 Temperature rise (**NRR4.1.1**)
$x41$	4.1.2 Drought (**NRR4.1.2**)
$x42$	4.1.3 Flooding (**NRR4.1.3**)
$x43$	4.1.4 Storm severity (annual average loss) (**NRR4.1.4**)
$x44$	4.1.5 Sea level rise (**NRR4.1.5**)
$x45$	4.2.1 Agricultural water risk – quantity (**NRR4.2.1**)
$x46$	4.2.2 Agricultural water risk – quality (**NRR4.2.2**)
$x47$	4.3.1 Land degradation (**NRR4.3.1**)
$x48$	4.3.2 Grassland (**NRR4.3.2**)
$x49$	4.3.3 Forest change (**NRR4.3.3**)
$x50$	4.4.1 Eutrophication (**NRR4.4.1**)
$x51$	4.4.2 Marine biodiversity (**NRR4.4.2**)
$x52$	4.5.1 Food import dependency (**NRR4.5.1**)
$x53$	4.5.2 Dependence on natural capital (**NRR4.5.2**)
$x54$	4.6.1 Early-warning measures / climate-smart Agriculture (**NRR4.6.1**)
$x55$	4.6.2 Commitment to managing exposure (**NRR4.6.2**)
$x56$	4.6.3 National agricultural adaptation policy (**NRR4.6.3**)
$x57$	4.6.4 Disaster risk management (**NRR4.6.4**)
$x58$	4.7.1 Projected population growth (**NRR4.7.1**)
$x59$	4.7.2 Urban absorption capacity (**NRR4.7.2**)

APPENDIX B: R PROGRAM FOR THE PROPOSED METHODOLOGY

Generating data from uniform (-2, 2) distribution using the code in Pan (2016)

```
p = 9; q = 2
sig <- 1; ni<- 10; n <- 60
y <- NULL
x <- NULL
z <- NULL

subject <- kronecker(1:n, rep(1, ni))   # ID

true.beta <- c(1,1,0,0,0,0,0,0,0)
Dt <- matrix(c(9,4.8,
                 4.8,4), nrow = q, ncol = q, byrow = TRUE)

for(i in 1:n){
    x[[i]] <- matrix(runif(ni*p, -2, 2), nrow = ni, ncol = p,
byrow = T)
    z[[i]] <- matrix(c(1,1,1,1,1,1,1,1,1,1, runif(ni, -2, 2)),
nrow = ni, ncol = q)
    V <- z[[i]] %*% Dt %*% t(z[[i]]) + diag(ni)
    y.temp <- t(rmvnorm(1, mean = x[[i]] %*% true.beta, sigma
= sig*V))
    y[[i]] <- y.temp
}

n <- length(y)
y1 <- y[[1]]
x1 <- x[[1]]
z1 <- z[[1]]
for(i in 2:n){
    y1 <- rbind(y1, y[[i]])
    x1 <- rbind(x1, x[[i]])
    z1 <- rbind(z1, z[[i]])
}
```

Inducing missing values in x and y after reshaping the data in wide format

```
dataWM$x11.1[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.1 - 1.2 * dataW$y11.1)] = NA
  dataWM$x11.2[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.2 - 0.1 * dataW$y11.2)] = NA
  dataWM$x11.3[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.3 - 0.2 * dataW$y11.3)] = NA
  dataWM$x11.4[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.4 - 0.2 * dataW$y11.4)] = NA
  dataWM$x11.5[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.5 - 0.3 * dataW$y11.5)] = NA
```

```
  dataWM$x11.6[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.6 - 1.2 * dataW$y11.6)] = NA
  dataWM$x11.7[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.7 - 0.1 * dataW$y11.7)] = NA
  dataWM$x11.8[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.8 - 0.2 * dataW$y11.8)] = NA
  dataWM$x11.9[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.9 - 0.2 * dataW$y11.9)] = NA
  dataWM$x11.10[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.10 - 0.3 * dataW$y11.10)] = NA


  dataWM$x22.1[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.1 - 1.2 * dataW$y11.1)] = NA
  dataWM$x22.2[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.2 - 0.1 * dataW$y11.2)] = NA
  dataWM$x22.3[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.3 - 0.2 * dataW$y11.3)] = NA
  dataWM$x22.4[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.4 - 0.2 * dataW$y11.4)] = NA
  dataWM$x22.5[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.5 - 0.3 * dataW$y11.5)] = NA
  dataWM$x22.6[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.6 - 1.2 * dataW$y11.6)] = NA
  dataWM$x22.7[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.7 - 0.1 * dataW$y11.7)] = NA
  dataWM$x22.8[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.8 - 0.2 * dataW$y11.8)] = NA
  dataWM$x22.9[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.9 - 0.2 * dataW$y11.9)] = NA
  dataWM$x22.10[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.10 - 0.3 * dataW$y11.10)] = NA


  dataWM$x33.1[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.1 - 1.2 * dataW$y11.1)] = NA
  dataWM$x33.2[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.2 - 0.1 * dataW$y11.2)] = NA
  dataWM$x33.3[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.3 - 0.2 * dataW$y11.3)] = NA
  dataWM$x33.4[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.4 - 0.2 * dataW$y11.4)] = NA
  dataWM$x33.5[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.5 - 0.3 * dataW$y11.5)] = NA
  dataWM$x33.6[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.6 - 1.2 * dataW$y11.6)] = NA
  dataWM$x33.7[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.7 - 0.1 * dataW$y11.7)] = NA
  dataWM$x33.8[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.8 - 0.2 * dataW$y11.8)] = NA
```

```
  dataWM$x33.9[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.9 - 0.2 * dataW$y11.9)] = NA
  dataWM$x33.10[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.10 - 0.3 * dataW$y11.10)] = NA


  dataWM$x44.1[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.1 - 1.2 * dataW$y11.1)] = NA
  dataWM$x44.2[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.2 - 0.1 * dataW$y11.2)] = NA
  dataWM$x44.3[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.3 - 0.2 * dataW$y11.3)] = NA
  dataWM$x44.4[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.4 - 0.2 * dataW$y11.4)] = NA
  dataWM$x44.5[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.5 - 0.3 * dataW$y11.5)] = NA
  dataWM$x44.6[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.6 - 1.2 * dataW$y11.6)] = NA
  dataWM$x44.7[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.7 - 0.1 * dataW$y11.7)] = NA
  dataWM$x44.8[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.8 - 0.2 * dataW$y11.8)] = NA
  dataWM$x44.9[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.9 - 0.2 * dataW$y11.9)] = NA
  dataWM$x44.10[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.10 - 0.3 * dataW$y11.10)] = NA



  dataWM$x55.1[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.1 - 1.2 * dataW$y11.1)] = NA
  dataWM$x55.2[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.2 - 0.1 * dataW$y11.2)] = NA
  dataWM$x55.3[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.3 - 0.2 * dataW$y11.3)] = NA
  dataWM$x55.4[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.4 - 0.2 * dataW$y11.4)] = NA
  dataWM$x55.5[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.5 - 0.3 * dataW$y11.5)] = NA
  dataWM$x55.6[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.6 - 1.2 * dataW$y11.6)] = NA
  dataWM$x55.7[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.7 - 0.1 * dataW$y11.7)] = NA
  dataWM$x55.8[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.8 - 0.2 * dataW$y11.8)] = NA
  dataWM$x55.9[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.9 - 0.2 * dataW$y11.9)] = NA
  dataWM$x55.10[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.10 - 0.3 * dataW$y11.10)] = NA
```

```
  dataWM$x66.1[runif(nrow(dataW))  <  invlogit(-0.1  +  2.5  *
dataW$z22.1 - 1.2 * dataW$y11.1)] = NA
  dataWM$x66.2[runif(nrow(dataW))  <  invlogit(-2.3  +  1.5  *
dataW$z22.2 - 0.15 * dataW$y11.2)] = NA
  dataWM$x66.3[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.3 - 2.2 * dataW$y11.3)] = NA
  dataWM$x66.4[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.4 - 0.2 * dataW$y11.4)] = NA
  dataWM$x66.5[runif(nrow(dataW))   <   invlogit(-3   +   2.5   *
dataW$z22.5 - 0.3 * dataW$y.5)] = NA
  dataWM$x66.6[runif(nrow(dataW))  <  invlogit(-0.1  +  2.5  *
dataW$z22.6 - 1.2 * dataW$y11.6)] = NA
  dataWM$x66.7[runif(nrow(dataW))  <  invlogit(-2.3  +  1.5  *
dataW$z22.7 - 0.15 * dataW$y11.7)] = NA
  dataWM$x66.8[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.8 - 2.2 * dataW$y11.8)] = NA
  dataWM$x66.9[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.9 - 0.2 * dataW$y11.9)] = NA
  dataWM$x66.10[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.10 - 0.3 * dataW$y.10)] = NA


  dataWM$x77.1[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.1 - 1.2 * dataW$y11.1)] = NA
  dataWM$x77.2[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.2 - 0.1 * dataW$y11.2)] = NA
  dataWM$x77.3[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.3 - 0.2 * dataW$y11.3)] = NA
  dataWM$x77.4[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.4 - 0.2 * dataW$y11.4)] = NA
  dataWM$x77.5[runif(nrow(dataW))   <   invlogit(-3   +   2.5   *
dataW$z22.5 - 0.3 * dataW$y11.5)] = NA
  dataWM$x77.6[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.6 - 1.2 * dataW$y11.6)] = NA
  dataWM$x77.7[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.7 - 0.1 * dataW$y11.7)] = NA
  dataWM$x77.8[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.8 - 0.2 * dataW$y11.8)] = NA
  dataWM$x77.9[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.9 - 0.2 * dataW$y11.9)] = NA
  dataWM$x77.10[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.10 - 0.3 * dataW$y11.10)] = NA

  dataWM$x88.1[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.1 - 1.2 * dataW$y11.1)] = NA
  dataWM$x88.2[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.2 - 0.1 * dataW$y11.2)] = NA
  dataWM$x88.3[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.3  - 0.2 * dataW$y11.3)] = NA
  dataWM$x88.4[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.4 - 0.2 * dataW$y11.4)] = NA
```

```
  dataWM$x88.5[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.5 - 0.3 * dataW$y11.5)] = NA
  dataWM$x88.6[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.6 - 1.2 * dataW$y11.6)] = NA
  dataWM$x88.7[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.7 - 0.1 * dataW$y11.7)] = NA
  dataWM$x88.8[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.8  - 0.2 * dataW$y11.8)] = NA
  dataWM$x88.9[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.9 - 0.2 * dataW$y11.9)] = NA
  dataWM$x88.10[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.10 - 0.3 * dataW$y11.10)] = NA


  dataWM$x99.1[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.1 - 1.2 * dataW$y11.1)] = NA
  dataWM$x99.2[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.2 - 0.1 * dataW$y11.2)] = NA
  dataWM$x99.3[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.3 - 0.2 * dataW$y11.3)] = NA
  dataWM$x99.4[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.4 - 0.2 * dataW$y11.4)] = NA
  dataWM$x99.5[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.5 - 0.3 * dataW$y11.5)] = NA
  dataWM$x99.6[runif(nrow(dataW))  <  invlogit(-0.1  +  1.5  *
dataW$z22.6 - 1.2 * dataW$y11.6)] = NA
  dataWM$x99.7[runif(nrow(dataW))  <  invlogit(-2.3  +  0.5  *
dataW$z22.7 - 0.1 * dataW$y11.7)] = NA
  dataWM$x99.8[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.8 - 0.2 * dataW$y11.8)] = NA
  dataWM$x99.9[runif(nrow(dataW))  <  invlogit(-2.2  +  0.10  *
dataW$z22.9 - 0.2 * dataW$y11.9)] = NA
  dataWM$x99.10[runif(nrow(dataW))  <  invlogit(-3  +  2.5  *
dataW$z22.10 - 0.3 * dataW$y11.10)] = NA


  # Impose missing data in y
  dataWM$y11.1[runif(nrow(dataW))  <  invlogit(-0.1  +  4.5  *
dataW$z22.1 - 2.98 * dataW$z11.1)] = NA
  dataWM$y11.2[runif(nrow(dataW))  <  invlogit(-2.3  +  4.5  *
dataW$z22.2 - 3.61 * dataW$z11.2)] = NA
  dataWM$y11.3[runif(nrow(dataW))  <  invlogit(-2.2  +  3.5  *
dataW$z22.3 - 1.2 * dataW$z11.3)] = NA
  dataWM$y11.4[runif(nrow(dataW))  <  invlogit(-2.2  +  5.10  *
dataW$z22.4 - 0.8 * dataW$z11.4)] = NA
  dataWM$y11.5[runif(nrow(dataW))  <  invlogit(-2  +  2.5  *
dataW$z22.5 - 0.6 * dataW$z11.5)] = NA
  dataWM$y11.6[runif(nrow(dataW))  <  invlogit(-0.1  +  4.5  *
dataW$z22.6 - 2.98 * dataW$z11.6)] = NA
  dataWM$y11.7[runif(nrow(dataW))  <  invlogit(-2.3  +  4.5  *
dataW$z22.7 - 3.61 * dataW$z11.7)] = NA
```

```
  dataWM$y11.8[runif(nrow(dataW)) < invlogit(-2.2 + 3.5 *
dataW$z22.8 - 1.2 * dataW$z11.8)] = NA
  dataWM$y11.9[runif(nrow(dataW)) < invlogit(-2.2 + 5.10 *
dataW$z22.9 - 0.8 * dataW$z11.9)] = NA
  dataWM$y11.10[runif(nrow(dataW)) < invlogit(-2 + 2.5 *
dataW$z22.10 - 0.6 * dataW$z11.10)] = NA
```

JM-SMC multiple imputation following the approach of Huque et al. (2020) after reshaping the data in long format

```
formula <- as.formula(y11 ~ 1 + x11 + x22 + x33 + x44 + x55 +
x66 + x77 + x88 + x99 + z22 + (1 + z22 | ID))

MCMC.dry = jomo.lmer.MCMCchain(formula, data = dataLM, nburn =
2, output = 2)   # check the convergence of the MCMC sampler
MCMC.check = jomo.lmer.MCMCchain(formula, data = dataLM, nburn
= 5000, output = 2)
plot(MCMC.check$collectbeta[1,1,1:5000], type = "l", ylab =
expression(beta["e,0"]), xlab = "Iteration numb")


JM.imp <- jomo.lmer(formula, data = dataLM, nimp = 10, output
= 2, nburn = 5000)
```

Stack the imputed data sets

```
xx.list <- NULL
yy.list <- NULL
zz.list <- NULL
clus.list <- NULL

for(i in 1:10){
    xx.list[[i]]   <-   as.matrix(impList[[i]][,   paste0("x",
c(11,22,33,44,55,66,77,88,99))])
    yy.list[[i]] <- impList[[i]]$y11
    zz.list[[i]]   <-   as.matrix(impList[[i]][,   paste0("z",
c(11,22))])
    clus.list[[i]] <- impList[[i]]$clus
}

# stacking - converting to vectors
xx1 <- do.call("rbind", xx.list)
zz1 <- do.call("rbind", zz.list)
yy1 <-  do.call("c", yy.list)
clus1 <- do.call("c", clus.list)
```

Introduce the subject weights then use adaptive LASSO for variable selection (stacked adaptive LASSO)

```
e_weights <- 1/10

# Random effects
aa = rand.lam.sel(xxx, yyy, zzz, D.init, eps = 1e-5, lam,
e_weights)
bestBIC.R = aa$bic
lambdaBIC.R = lam[which.min(bestBIC.R)]
estr.bic = rand.sel(lambdaBIC.R, xxx, yyy, zzz, D.init, eps =
1e-5, e_weights)


# Fixed effects
bb = fix.lam.sel(xxx, yyy, zzz, estr.bic$beta, estr.bic$D, eps
= 1e-5, lam, e_weights)
bestBIC.F = bb$bic
lambdaBIC.F = lam[which.min(bestBIC.F)]
estf.bic = fix.sel(xxx, yyy, zzz, estr.bic$beta, estr.bic$D,
lambdaBIC.F, eps = 1e-5, e_weights)
```