8-2017

Power Analysis of Longitudinal Data with Time-Dependent Covariates Using Generalized Method of Moments

Niloofar Ramezani
University of Northern Colorado

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

POWER ANALYSIS OF LONGITUDINAL DATA WITH TIME-DEPENDENT COVARIATES USING GENERALIZED METHOD OF MOMENTS

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

Niloofar Ramezani

College of Education and Behavioral Sciences
Department of Applied Statistics and Research Methods

August 2017

This Dissertation by: Niloofar Ramezani

Entitled: *Power Analysis of Longitudinal Data with Time-Dependent Covariates using Generalized Method of Moments*

Has been approved as meeting the requirement for the degree of Doctor of Philosophy in College of Education and Behavioral Sciences, Program of Applied Statistics and Research Methods

Accepted by the Doctoral Committee

_____
Trent L. Lalonde, Ph.D., Research Advisor

_____
Jay R. Schaffer, Ph.D., Committee Member

_____
Susan R. Hutchinson, Ph.D., Committee Member

_____
Mehrgan Mostowfi, Ph.D., Faculty Representative

Date of Dissertation Defense _____

Accepted by the Graduate School

_____
Linda L. Black, Ed.D., LPC
Associate Provost and Dean
Graduate School and International Admissions

ABSTRACT

Ramezani, Niloofar. *Power Analysis of Longitudinal Data with Time-Dependent Covariates using Generalized Method of Moments.* Published Doctor of Philosophy dissertation, University of Northern Colorado, 2017.

Longitudinal data occur in different fields such as biomedical and health studies, education, engineering, and social studies. Planning advantageous research projects with both high power and minimum sample size is an important step in any study. The extensive use of longitudinal data in different fields and the importance of their power estimation, yet the limited resources about their respective power estimation tools, made it worthwhile to study their power estimation techniques.

The presence of time-dependent covariates triggers the need to use more efficient models such as generalized method of moments than the existing models which are based on generalized estimating equations. Not taking into consideration the correlation among observations and the covariates that change over time while calculating power and minimum sample size will cause expensive research being conducted without using data that are capable of answering the research questions (Williams, 1995). Two different power estimation and minimum sample size calculation techniques for longitudinal data in the presence of time-dependent covariate using generalized method of moments approaches are constructed in this study and their performances are evaluated.

DEDICATION

This dissertation is dedicated to:

My loving parents and my only brother.

For their unconditional love, endless support and encouragement, and for always believing in me.

ACKNOWLEDGEMENTS

I am thankful for the support I had along the way at the University of Northern Colorado. Firstly, I would like to thank my amazing advisor, Dr. Trent Lalonde, for always believing in me to explore a new area of research and finish up this dissertation. I am grateful for the endless encouragement, insight, guidance, and help he provided throughout my graduate studies and my dissertation work. I could have not asked for a more supportive and encouraging mentor, advisor, and professor and I am truly grateful and honored for the opportunity of being his student and advisee.

Additionally, I am appreciative of my committee members, Dr. Jay Schaffer, Dr. Susan Hutchinson, and Dr. Mehrgan Mostowfi for their careful reading of my dissertation, their time, and helpful suggestions.

I am immensely grateful for the unconditional support I have always received from my family: my mom, Shahnaz, my dad, Reza, and my one and only brother, Ali. I could have never accomplished any of this without their unlimited love, encouragement, and the invaluable life lessons I learned from them. They were always there for me no matter what. The countless conversations and their pieces of advice during difficult times helped me to be where I am today. So, I can never thank them enough.

I met the most wonderful people at the department over the years. I want to thank all of them especially Keyleigh Gurney, who is the most encouraging listener, Michael Floren and Robert Resch, who were so generous when sharing their programming knowledge, and all my friends especially Karen Traxler, Rita and James Buller.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Planning advantageous research projects with both high power and minimum sample size is an important step in any study which influences all future results of the research; therefore, comprehensive and innovative research studies are needed to address different issues associated with this step. If this step is not taken after systematic planning and with caution toward building valuable research design, the final results will not be valid and the outcome may not contribute to the body of the research. Therefore, studying power can greatly benefit almost any scientific study with statistical input where inferential statistical procedures are used.

The motivation for this research comes from some weaknesses of the current approaches which are being taken in designing longitudinal studies and estimating their statistical power. Longitudinal studies are extensively used across disciplines to model changes over time in the presence of multiple time point measurements while taking into consideration the dependence among repeated measurements per subject. For instance, when studying patients with knee osteoarthritis over years with the goal of modeling their body stiffness and pain, there are multiple time points when they check into a hospital regarding their pain and treatment; hence, there exist multiple measurements for each

patient. These multiple observations are correlated within each patient and the severity of each patient's pain may vary through a period of time making longitudinal models and their respective power estimation techniques more appropriate than the ones used for cross-sectional data which is a type of data collected at one time point. The correlation that exists among observations is why longitudinal data are sometimes referred to as correlated data. Correlated data are a more general case of longitudinal data which include any type of correlation that may exist among observations due to clustering or repeated measurements per subject over a period of time. Therefore, due to the fact that this dissertation is mainly focused on the correlation that exists among repeated observations over time, it is more appropriate to use the longitudinal term for this study.

When handling the correlation which exists among observations within longitudinal data, more advanced models are required to account for the dependence between multiple outcome values observed within each subject. Not taking into consideration the correlation among observations will result in unreliable conclusions when using flawed models for analyzing such data. Even worse, the erroneous theory being used for calculating power and the minimum sample size needed for analysis, may lead to expensive research being conducted that is incapable of answering the research questions (Williams, 1995). Sequentially observed over time, longitudinal data may be regarded to as a collection of numerous time series, which is a sequence of data points in successive order, one each per subject. This type of data allows conducting studies on the changes of the variability of the response in time with covariates possibly changing with time. One benefit in using longitudinal data is reducing the burden of recruiting a sizable number of subjects, compared to cross-sectional studies, by collecting repeated outcomes

from each subject. Longitudinal data analysis, which relies on the underlying probability mechanism of changes including growth, aging, time profiles, or effects of covariates over time, is technically more demanding and complex than cross-sectional data analysis. The presence of repeated measurements for each subject indicates that the observations from the same subjects are auto-correlated or serially correlated. This requires the development of statistical methodology with the ability to account for the serial correlation.

When studying responses over time, more advanced models such as conditional models, transition models, or marginal models are required (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2009). Conditional or subject-specific models are used when the regression coefficients are used to describe an individual's response to changing covariates while marginal or population averaged models are used when one does not attempt to control for unobserved subjects' random effects. The primary estimation method adopted for this study is generalized method of moments (GMM) which is a population averaged model. GMM is preferred in this study because it provides more efficient estimates compared to the other marginal models in the presence of time-dependent covariates, which is the primary interest in this dissertation (Lai & Small, 2007). Time-dependent covariates are the covariates that may vary within individuals throughout the study. Age, weight, and systolic blood pressure are some examples of such covariates. For instance, when modeling the effectiveness of a drug over time for patients with heart disease, variables such as patients' blood pressure or weight might change over time. This change may affect the treatment process of the patients potentially as the result of interactions between those time-dependent covariates and the status of the

heart disease or drug usage. However, patients' race or sex will remain constant so they will not fluctuate the way a drug affects patients' treatment over time. This type of covariate is called time-independent which remains constant through the whole study.

Borrowing the strength from the theory of generalized linear models is important in developing marginal components, which are suitable for incorporating correlation of the outcomes (Zeger & Liang, 1986). Examples of such models are generalized estimating equations (GEE) and GMM. GEE are based on quasi-likelihood inference that depends on the first two moments of the underlying distribution of the data and treats the correlation as a nuisance parameter (Liang & Zeger, 1986), whereas GMM may be regarded as a class of inference functions constructed through a limited set of moment conditions of the underlying statistical model with no need for complete specification of the probability model. Moments of a distribution mentioned above refer to mean, variance, skewness, and so forth.

GMM estimators are preferred to maximum likelihood (ML) estimators in this proposed study because according to Hall (2005), they are more robust due to not having any distributional assumptions, more tractable analytically, and more stable numerically. On the other hand, ML estimators are asymptotically more efficient only if the model is correctly specified. GMM estimators are more robust even in the presence of distributional misspecification (Hall, 2005). They also are more consistent with respect to the correct specification of only the limited set of the moment conditions in contrast with the ML estimators that require correct specification of every conceivable moment condition (Newey & McFadden, 1994).

The extensive use of longitudinal data and the importance of their power estimation, yet the limited resources about their respective power estimation tools, made it worthwhile to study the power estimation techniques for different types of longitudinal outcome variables. Although some valuable literature on the subject is currently available, there is less focus on instances in which there exist time dependent covariates. When trying to estimate the power of longitudinal studies in the presence of time-independent covariates, Rochon (1998), Liu and Liang (1997), and Lyles, Lin, and Williamson (2007) proposed some GEE-based techniques to estimate the minimum sample size. It is when the covariates vary through the study that there is a need for developing better methods to estimate the power and minimum sample size based on GMM which is so far the most appropriate marginal technique for modeling longitudinal data with time-dependent covariates. The estimation technique used in this paper for developing power estimation methods for longitudinal data with time-dependent covariates is the GMM approach.

Purpose of the Study

The purpose of this dissertation was to assess power estimation and minimum sample size calculation techniques for different hypothesis tests with the focus on longitudinal data. The objective of this study was to investigate various methods for power and minimum sample size calculation of longitudinal data that are gathered over time in the presence of time-dependent covariates using GMM. The primary methodology involved the use of GMM in estimating statistical power. GMM, which performs better in terms of efficiency than the previous methods that were based on GEE,

was used to extend the existing methods to a more efficient technique when dealing with longitudinal responses in the presence of time-dependent covariates.

Different approaches of power calculation for longitudinal data with time-dependent covariates were modeled and discussed using the GMM technique. In order to do that, the distribution of each statistic under null and alternative hypotheses needed to be estimated as knowing these distributions is a necessary element of power analysis. The performance of these approaches within the GMM technique was evaluated using a real data set and through a simulation study. Performance of the theoretically developed methodology at the end was compared to the empirical powers.

Rationale for the Study

When planning for any research project, it was important to consider different aspects of the data that need to be accounted for in the study. Assuring researchers collect enough data in the data collection process is crucial since without appropriate consideration of power and required minimum sample size, the entire study may fail and the research findings may be deceptive (Kraemer & Blasey, 2015). On the other hand, collecting more than enough data will result in wasted time and resources, often for minimal gain. The optimal sample size refers to a large enough sample size to get statistically significant results yet not too large to be only time consuming and expensive without a notable gain. This calculation tends to be a desired part of the protocol of nearly all scientific research, especially the studies involving human or animal subjects in which too small or too large sample sizes will have ethical, scientific, and budgetary implications.

In the process of planning research and finding out about the minimum sample size, for this dissertation I focused on longitudinal data due to their extensive use in different fields and especially by applied researchers and practitioners willing to answer research questions that address changes over time. Unfortunately, no extensive studies on power and sample size calculation had been completed within longitudinal designs in the presence of time-dependent covariates.

Within longitudinal studies, estimation techniques such as GEE and GMM, which address issues regarding longitudinal data, need to be applied to appropriately account for the correlation among repeated observations. Among these more advanced models, GMM was my main focus for this dissertation due to its higher efficiency compared to GEE in particular when dealing with time-dependent covariates (Lai & Small, 2007). GMM models provide consistent, efficient, and asymptotically normally distributed estimators with minimal use of information only from the moment conditions. According to Hansen (2007), GMM also takes care of both sampling and estimation error by its unique way of constructing tests. All of these characteristics make GMM a desirable method to be used when providing estimation of unknown parameters within various models; therefore, making it crucial to study the power estimation methods within this technique.

There is a gap in the literature regarding appropriate power analysis and sample size calculation techniques based on GMM, which is important to be studied as GMM is more appropriate than estimation methods such as GEE when working with longitudinal data in the presence of time-dependent covariates. In addition to the aforementioned advantages of GMM, it can also be seen as a generalization of many other estimation techniques such as least squares (LS), instrumental variables (IV), or maximum

likelihood (ML; Chaussé, 2010), which makes it even more important to come up with an efficient power analysis technique for GMM.

The advantages of GMM and the lack of available power analysis techniques for such models were the main rationales of this dissertation. In this study my aim was to provide an easier power and sample size calculation technique for applied researchers and practitioners with minimum knowledge about the distribution of the data. Such applied researchers are those who want to conduct cost effective research studies and at the same time be sure of selecting an appropriate model and optimal sample size for longitudinal data with time-varying covariates, which result in a high power of the performed tests. In this paper, different approaches of power estimation and sample size calculation for longitudinal data with time-dependent covariates using GMM are discussed to fill this gap. Previous methods for power estimation techniques are mainly based on GEE using Wald, likelihood ratio, and score tests. In the current study, the possibility of using the Wald test as well as the distant metric statistic, which is based on the difference of GMM-based quadratic forms, within GMM methods were investigated. These methods are no longer likelihood-based and rely on moment conditions. Moment conditions of a population are the assumed moments of the random variables and the analogous sample moment conditions can be computed using the data.

Research Questions

In order to develop power estimation and minimum sample size calculation methods for tests using GMM with a focus on longitudinal data with time-dependent covariates, this dissertation addressed the following questions:

Q1 How can power be calculated for hypothesis tests using longitudinal data with time-dependent covariates applying a Wald approach within a GMM estimation technique?

Q2 How can sample size be calculated for a desired level of power for hypothesis tests using longitudinal data with time-dependent covariates applying a Wald approach within a GMM estimation technique?

Q3 How can power be calculated for hypothesis tests using longitudinal data with time-dependent covariates applying a Distant Metric Statistic approach within a GMM estimation technique?

Q4 How can sample size be calculated for a desired level of power for hypothesis tests using longitudinal data with time-dependent covariates applying a Distant Metric Statistic approach within a GMM estimation technique?

Q5 How well do the proposed power calculation approaches within a GMM method perform compared to the empirical power?

This study had two phases. The first phase was where the first four research questions were addressed theoretically through a set of proofs in Chapter III. The second phase was where a practical power estimation procedure was developed and the fifth research question were answered empirically through the analysis of an exemplary data set and Monte Carlo simulation methods. It was necessary to develop the theoretical portion of this dissertation first before implementing the empirical component of the study. This is why the theoretical derivation of the power calculation procedure of this study and the proofs I constructed are presented in Chapter III along with answers to research questions 1 through 4 before answering question 5 in Chapter IV.

Methods

The performance of the two GMM-based power estimation techniques presented in this dissertation were evaluated using a pre-existing data set as well as a simulation study. The pre-existing data set consists of osteoarthritis initiative (OAI) data from a

multi-center study on osteoarthritis of the knee and contains follow-up information for about 4,000 subjects aged 45 and above over a period of up to 9 years. Studying these data helps understanding risk factors for progression of osteoarthritis of the knee. Osteoarthritis causes problems ranging from stiffness and mild pain to severe joint pain and even disability. The Western Ontario and McMaster Universities (WOMAC) disability score is typically treated as a continuous value indicating patients' pain, stiffness, and physical function with knee osteoarthritis. The average of the WOMAC scores for the left and right knee, which is a continuous variable, was used as the response which might be affected by different variables in the presence of time-dependent covariates. Considering this continuous response variable for the current study provides the opportunity of evaluating the effectiveness of the proposed power calculation techniques when modeling such response variables as the most common type of outcomes. This dataset was drawn from http:// www.oai.ucsf.edu. The OAI dataset is longitudinal, as desired for this study due to the repeated observations over time on each of the patients at multiple follow-up times. Using this dataset, a practical theoretical power estimation procedure for the pilot data sets was developed.

A simulation study was also used for evaluating the performance of different power calculation techniques in this dissertation. The data were simulated using Monte Carlo simulation in R version 3.2.2 (R Core Team, 2015). This simulation was based on the real dataset introduced above. Continuous responses were generated so they would be consistent with the outcome variable from the OAI data. Four sample sizes and two power estimation techniques were considered in this simulation study and the results of the estimated powers were compared to the empirical power and post-hoc powers at the

end to evaluate the performance of the new power techniques. More details about the simulated data are provided in Chapter III of this dissertation.

Chapter II includes an in-depth review of the most current literature pertaining to the power calculation techniques for longitudinal data, including reviews of previously studied methods of modeling longitudinal data as well as different power analysis techniques being performed for correlated data. Chapter III involves estimating power for tests using GMM and related theoretical proofs and procedures. Chapter IV includes the data analysis and results for this study using the OAI data set as well as a simulation study which were mainly used to evaluate the performance of the proposed methods. Finally, Chapter V consists of discussion, impact, limitations, and possible future work pertaining to the topics and methods discussed throughout this dissertation.

CHAPTER II

LITERATURE REVIEW

This chapter is dedicated to reviewing the literature on longitudinal data and different methods of analyzing this type of data with the purpose of providing the necessary background to discuss power analysis techniques of longitudinal data. The first section introduces the idea of power analysis and the important role it plays in any research study, specifically at the planning stage. The second section provides the background information regarding longitudinal data analysis. Within this section, a summary of different techniques of analyzing longitudinal data and estimating model parameters is provided. This subsection helps facilitate understanding of the differences between longitudinal modeling techniques, which are used when responses are measured at different time points and cross-sectional designs, which are used for modeling the outcomes that are measured at a single time point. After briefly introducing the generalized linear models (GLM), which are widely used for longitudinal data with continuous outcome variables, I mention different extensions to GLM in the subsequent sections of this chapter used for modeling different types of longitudinal responses. These extended models were developed to accommodate the correlation among observations inherent in longitudinal data with varying types of response variables, which are the main types of data considered in the current study. Two of the most important estimating

techniques that researchers use for longitudinal analyses are discussed, respectively, in the third and fourth sections of this chapter. These methods include generalized estimating equation (GEE) and generalized method of moments (GMM). The final section of this chapter is devoted to introducing power and discussing three of the most important available power analysis techniques and sample size calculation methods for correlated observations when model parameters are estimated using a GEE. These techniques were developed based on Wald statistics, likelihood ratio statistics, and score statistics. Finally, these techniques are discussed to support being adopted and extended in developing two new methods of power analysis of longitudinal data in the presence of time-dependent covariates based on GMM.

Introduction

Planning successful and cost-effective research projects is an important goal of every researcher in every field of study in order to answer research questions and help make policy. After designing a study and deciding the most appropriate type of statistical test to use, researchers need to perform the test. Studying the entire population is not practical or even possible for the majority of research studies. What can be done instead of looking into all the population observations is to take a sample of the population (Kraemer & Blasey, 2015). The sampling techniques and the details about taking a representative sample is an important topic but is not discussed here. What is important here is the number of subjects to sample, assuming that we already know how to sample them and what measures to use. For more information about different sampling techniques refer to Thompson (2012).

Making a mistake in the process of planning a study, such as when planning what data to collect and how many subjects to sample in order to study a population, is irrevocable and more of a challenge than when making a mistake in the data analysis. By not considering different aspects of the data that need to be collected or not collecting enough data during the timeline of data collection, the data that researchers have spent years and major resources to collect may not be useful in answering the research questions they posed. However, if the problem was in miscalculating the test-statistics or p-value or even the statistical technique used for analyzing the data, the analyses could easily be redone; however, there is almost nothing that statisticians can do with a researcher's data that are not appropriate or that are based on an insufficient sample size. This is where the importance of power calculations in planning of research projects comes to attention. Without appropriate consideration of power, hence the sample size, the entire enterprise of applying the scientific method is likely to fail and the research findings may be misleading (Kraemer & Blasey, 2015). Unfortunately, incorrect results, due to the use of inadequate sample size within a study, may be published and this problem has been highlighted in the highly cited article by Ioannidis (2005).

Sometimes there are multiple research designs and associated tests that can be used to answer one research question while each method requires a different sample size to get valid results. Performing the power analysis for all the possible tests and choosing the method which needs a smaller required sample size, compared to the other possible tests, can help researchers in getting the most feasible and cost effective design. Through power analysis, the optimal sample size can be picked for an appropriate test which will prevent researchers from ending up with a more expensive and time-consuming study for

a minimal gain (Kraemer & Blasey, 2015). These are a few reasons to emphasize the importance of the power analysis practice.

The focus of this dissertation was to study longitudinal data, due to their extensive use in different fields and especially by applied researchers who are interested in answering research questions that address changes over time. Due to the correlated nature of this type of data, regular power analysis techniques are not appropriate and more advanced methods are needed for sample size calculations. These power estimation and sample size calculation methods are tied to the estimation method used within longitudinal models of correlated data. The review of these estimation techniques starts with introducing models such as GLM and different estimation methods commonly used within the models which cannot address all the issues regarding longitudinal data such as the presence of different types of covariates and discrete responses. The review then suggests more advanced estimation techniques such as GEE and GMM, which can be performed within different models to appropriately account for the correlation among repeated observations for non-normal outcomes. Then, different methods of power analysis for such models by Rochon (1998), Liu and Liang (1997), and Lyles et al., (2007) are discussed in this review. Review of these three studies helps identify a gap in the literature regarding appropriate power analysis and sample size calculation techniques based on GMM, which is the main topic of this dissertation. The reason that GMM and its power analysis is important is because it can be seen as a generalization of many other estimation methods like least squares (LS), instrumental variables (IV), or maximum likelihood (ML; Chaussé, 2010). According to Lai and Small (2007), GMM is also more efficient when modeling longitudinal data in the presence of time-dependent covariates

which is the type of covariate that does not remain constant throughout the period of a study.

According to Hall (2005), GMM is proven to be a very flexible estimating technique since it does not require full distributional assumptions, which in practice may not be specified for different research studies. It only requires some assumptions about moment conditions. Moment conditions contain information about unknown parameters and are functions of the model parameters and the data, such that their expectation is zero at the true values of the parameters. By minimizing a quadratic form of the moment conditions, which is introduced in section four of this chapter, the GMM-based parameter estimates may be found. This estimation is obtained by finding the parameters that make the sample moment conditions as close to the population moment conditions as possible.

The flexibility of the GMM estimation technique can be observed in different real world examples. For instance in macroeconomics, GMM allows estimating a structural model equation. As another example, we can look at finance in which most data such as stock returns are characterized by skewed and heavy-tailed distributions. Because GMM does not impose any restriction on the distribution of the data, it is a good alternative in this area as well.

GMM is also a reliable estimation procedure for many models especially in economics. For example, GMM with the right moment conditions is more appropriate than ML in general equilibrium models, which suffer from endogeneity problems when attempting to explain the behavior of supply, demand, and prices in a whole economy with several interacting markets. In statistical models, endogeneity problems arise when there is a correlation between the explanatory variables and the error term as a result of

measurement error, autoregression with correlated errors, simultaneity, and omitted variables. Within finance studies, GMM is appealing in many cases due to the distribution-free feature, as there is no satisfying parametric distribution which reproduces the properties of stock returns. Some claim that the family of stable distributions is a good candidate but only the densities of the normal, Cauchy, and Levy distributions, which belong to this family, have a closed form expression. Therefore, GMM still is a better candidate for parameter estimation in finance (Hall, 2005).

GMM estimators are consistent, which is another important characteristic that one can look for in any estimation technique; however, efficiency and bias depend on the choice of moment conditions so cannot be justified without considering the chosen moment conditions for each design. Furthermore, GMM can be used to estimate the model parameters and perform inferences, in even non-linear dynamic models when only a set of population moment conditions, which are deduced from the assumptions of the models, are known (Hall, 2005).

The advantages of GMM and the lack of available power analysis techniques for such models provided the motivation to study this topic. This study is important in order to provide an easier power and sample size calculation technique for applied researchers with minimum knowledge about the distribution. Adopting the methods developed in this study, applied researchers and practitioners will end up with the most cost effective model selection and sample size with the highest possible power at the same time. GMM is specifically preferred to other models when dealing with longitudinal data and time-dependent covariates. More details regarding the GMM estimation technique are discussed in the GMM section.

Longitudinal Data

In the presence of multiple time points for subjects of a study and the interest of patterns of change over time, longitudinal data are formed with the main characteristics of dependence among repeated measurements per subject (Liang & Zeger, 1986). This correlation among measures for each subject introduces a complexity to the study due to violating the assumption of independence among the observations, which requires more complex models that enable researchers to take into consideration all aspects of such models.

For example, when modeling body pain and stiffness of patients with knee osteoarthritis over years, there are multiple measurements for each patient. The severity of each patient's pain may vary through a period of time, but these observations are correlated within each patient, making cross-sectional data models inappropriate. This example is explained in detail in Chapter III of this dissertation.

Many models have been developed for cross-sectional data where a single observation for each subject is available, but more studies regarding modeling of longitudinal data in the presence of varying types of responses and covariates need to be conducted regardless of their challenges. Conducting more studies in this area is important because of the opportunities repeated observations provide for researchers such as increased statistical power and robustness to model selection. The higher power of longitudinal studies is due to having the same number of subjects as a comparable cross-sectional study but more observations, due to multiple observations per subject, as well as generally smaller error terms resulting from these additional observations. Additionally, some model misspecification problems can be avoided within longitudinal studies as they

allow analyses that are insensitive to omitted covariates that do not change with time. This will result in robust model selections and inferences common in observational studies (Liang, Zeger, & Qaqish, 1992).

According to Zeger and Liang (1991), although using each subject as his own control will result in homogeneity among subjects over time and hence increased efficiency, there is an analytic cost researchers may pay by inconsistent estimates of precision by ignoring the existing correlation among subjects of longitudinal data. However, these challenges can be met and overcome by appropriate models that are specifically designed to capture the correlation among the observations and use them to have a greater power and make inferences. Some of these methods that can be used in modeling longitudinal data are discussed below.

Modeling Longitudinal Data

The early development of methods that can handle longitudinal data is traced back to the usefulness of the ANOVA paradigm for longitudinal studies and to the seminal paper by Harville (1977). The most common way of modifying ANOVA for longitudinal studies is repeated measures ANOVA, which simply models the change of measurements over time through partitioning of the total variation. The total variance may be partitioned for such models using time-dependent and time-independent variables. A time-dependent variable will take on values that may change for different observations on the same subject; however, a time-independent variable has the same value on the same subject for all observations.

After developing and adopting the repeated measures ANOVA technique for longitudinal studies, the idea of having random effects in a model in addition to fixed

effects and ending up with a mixed-effects model was developed. Mixed-effects models, which historically go back to 1930's when Wishart (1938) started the early contribution to the growth curve analysis, enable researchers to efficiently model longitudinal data. The early use of mixed-effects within ANOVA for longitudinal data analysis was mainly in life science, which Laird and Ware (1982) highlighted. This way of using ANOVA was among the first steps of developing mixed-effects model, which is probably the most widely used method of analyzing longitudinal data (Fitzmaurice et al., 2009).

The idea of randomly varying regression coefficients was also a common thread in the two-stage approach of longitudinal data analysis. A two-stage method is based on assuming that the repeated measurements on each subject follow a regression model with distinct regression parameters for each individual. While this method was used for years by different people in different ways, Rao (1965) was the one who formally used this two-staged model by specifying a parametric growth curve model based on the assumption of normality of the random growth curve parameters. Although relatively simple to use and providing the motivation for more advanced models from an historical perspective, the two-stage methods force some restrictions which are not necessary and are sometimes very inconvenient in terms of modeling. These restrictions include having only time-varying covariates in the first stage, the limitation of having the ability to introduce the between-subject covariates only in the second stage and finally putting unnecessary constraints on the choice of the design matrix for the fixed effects (Fitzmaurice et al., 2009).

Linear Mixed-Effects Models

It was in the early 1980s that Laird and Ware (1982) proposed their flexible class of linear mixed-effects models for longitudinal data based on the earlier work done on the general class of mixed models by Harville (1977). Repeated-measures ANOVA and growth curve models are considered as special cases of this model. Furthermore, the linear mixed-effects model for longitudinal data has fewer restrictions on the fixed and random effects design matrices as well as more efficient likelihood-based estimation of the model parameters (Fitzmaurice et al., 2009). The linear mixed-effects model is given by

$$Y_{it} = \boldsymbol{X'}_{it}\boldsymbol{\beta} + \boldsymbol{Z'}_{it}\boldsymbol{\gamma} + e_{it}, \tag{2.1}$$

where Y_{it} is the response variable of the ith subject observed repeatedly at different time points ($i = 1,\ldots,n, t = 1,\ldots,T$), n is the number of subjects, T is the number of time points, $\boldsymbol{X}_{it}$ is the design or covariance vector of tth measurement at time t for subject i for the fixed effects, $\boldsymbol{\beta}$ is the fixed effect parameter vector, $\boldsymbol{Z}_{it}$ is the design vector of the tth measurement measured for subject i for the random effects, $\boldsymbol{\gamma}$ is the random effect parameter vector following a normal distribution, $\boldsymbol{\gamma} \sim N(0, \boldsymbol{G})$ and $\boldsymbol{e}_{it}$ is the random error also following a normal distribution, $e_{it} \sim N(0, R)$. When using this model for longitudinal studies, subjects can be considered as clusters with different measurements across time per subject; therefore, there will be i subjects and t different time points.

Having the random effects within these mixed-effects models helps account for the correlation among the measurements per subject at different time points. Within the two normal distributions of the random vectors mentioned above, $\boldsymbol{G}$ is the covariance matrix of random effects $\boldsymbol{\gamma}$ and R_i is the covariance of error term e_{it}. Different

correlation structures among errors can be assumed but the most common one is the constant covariance, $\sigma^2 \boldsymbol{I}$. Additionally, the distributions of the random effects can vary from normal. According to Laird and Ware (1982), Different algorithms other than the Expectation-Maximization (EM) can be used to fit this general class of models to longitudinal data (e.g., Fitzmaurice et al., 2009; Jennrich & Schluchter, 1986). During the mid-1980s a very general class of linear models was proposed that could handle longitudinal unbalanced data in the presence of mistimed measurement or missing data as well as time-varying or time-invariant covariates and yet provide parsimonious and flexible covariance models (Fitzmaurice et al., 2009).

Nonlinear Models

Although the developments in methods of analyzing longitudinal continuous responses span about a century, many of the advances in methods for analyzing longitudinal discrete responses have been limited to the most 30 to 35 years. According to Fitzmaurice et al. (2009), when the response variables are discrete within a longitudinal study and no longer normally distributed, linear models are no longer appropriate. To solve this problem, statisticians have developed approximations of GLM (Wedderburn, 1974) for longitudinal data. A characteristic feature of GLMs is the addition of a non-linear transformation of the mean, which is assumed to be a linear function of the covariates that can introduce some issues in the regression coefficients of longitudinal data. This problem has been solved also by extending GLMs to handle longitudinal observations in a number of different ways that can be categorized into three main types of models. These categories are: (i) conditional models also known as random-effects or subject-specific models, (ii) or transition models and (iii) marginal or

population averaged models (Fitzmaurice et al., 2009). These models have some main differences from each other in terms of how they account for the correlation among the repeated observations within longitudinal data and the interpretations of the regression parameters resulting from these models (Fitzmaurice et al., 2009).

Conditional, Transition, and Marginal Models

Conditional models are appropriate when a researcher seeks to examine individual level data. For example, take the situation of modeling the academic success of students clustered into majors within a single university or the academic success of students over time. If the interpretation of the results seeks to explain what factors impact academic success of students and their individual trend, a conditional model would be appropriate (Zorn, 2001). Conditional or random effects models allow adding a random term to the model to capture the variation in the population of subjects and also the correlation among the observations. A linear version of conditional models can be specified as

$$Y_{it} = \boldsymbol{X'}_{it}\boldsymbol{\beta} + \nu_i + e_{it}, \tag{2.2}$$

where $\nu_i \sim N(0, G)$ is a random effect and $e_{it} \sim N(0, R)$ is the random error term. The conditional mean can be obtained as

$$E(Y_{it}|\nu_i) = \boldsymbol{X'}_{it}\boldsymbol{\beta} + \nu_i. \tag{2.3}$$

One example of conditional models is the Generalized Linear Mixed Model (GLMM), which is a GLM that includes a random effect and can be applied to longitudinal data. GLMM can be considered as an extension of the GLM in which the mean response model is conditioned on both measured covariates and an unobserved random effect. When averaging over the distribution of the random effects, the within-subject correlation among the repeated responses within longitudinal data is marginally

captured by allowing the regression coefficients to vary randomly from one individual to another through entering random effects in the model for the mean response. Within GLMMs, there usually is the assumption that the random effects are normally distributed (multivariate normal) and are independent of the covariates (Fitzmaurice et al., 2009).

The general form of GLMM can be written as Equation 2.1. What is different between the mixed-effects models and GLMMs is that the response variables can come from different distributions besides Gaussian (aka normal). Within GLMM, rather than modeling the responses directly, some link function is often applied, such as a log link. Let the linear predictor, $\boldsymbol{\eta}$, be the combination of the fixed and random effects excluding the residuals specified as below

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}, \tag{2.4}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects, $\boldsymbol{X}$ is the fixed effects design matrix, $\boldsymbol{\gamma}$ is the vector of random effects such that $cov(\boldsymbol{\gamma}) = \sigma^{\boldsymbol{2}}\boldsymbol{D}$ for at least positive definite matrix $\boldsymbol{D}$ and $\boldsymbol{Z}$ is the random effects design matrix. The link function, $g(.)$, relates the outcome, $\boldsymbol{Y}$, to the linear predictor , $\boldsymbol{\eta}$. One of the most common link functions is $g(.) = log_e(\frac{p}{1-p})$ and $g(E(\boldsymbol{Y})) = \boldsymbol{\eta}$. Both the estimations of fixed and random effects coefficients, respectively, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, can be found within these models. In the GLMM, the default optimization technique that is used is the Quasi-Newton method. Because a residual likelihood technique is used to compute the objective function, only the covariance parameters participate in the optimization. This model is not complicated and more details about it can be found in Agresti (2007).

Transition models also result from extending generalized linear models by modeling the mean and time dependence simultaneously via conditioning an outcome on

other outcomes to handle longitudinal data. Transition model, or Markov, is a specific kind of conditional model which accounts for the correlation between subjects of a longitudinal study by letting the past values influence the present observations, which are of interest by considering the sequential nature of longitudinal data. The fact that the conditional distribution of each response at any occasion is expressed given the past responses and covariates makes the transition models part of conditional ones (Fitzmaurice et al., 2009). In transition models,

$$E(Y_{it}|\boldsymbol{H}_{it}, \boldsymbol{X}_{it}) = \boldsymbol{X'}_{it}\boldsymbol{\beta} + \sum_{r=1}^{s} \alpha_r f_r(\boldsymbol{H}_{it}), \quad (2.5)$$

where $\boldsymbol{H}_{it} = \{Y_{i1}, \dots, Y_{it-1}\}$ denotes the history of the past responses at the tth occasion and $f_r(\boldsymbol{H}_{it})$ denotes some known functions of the history of the past responses with α_r as the coefficients of these past history functions of the responses.

Marginal models are also the extension of GLMs, which directly incorporate the within-subject association among the repeated measures of the longitudinal data into the marginal response distribution. The principal distinction between marginal and conditional models has often been asserted to depend on whether the regression coefficients describe an individual's response or the marginal response to changing covariates, that is, one that does not attempt to control for unobserved subjects' random effects (Lee & Nelder, 2004). Marginal models can be written as in Equation 2.6,

$$E(Y_{it}) = \boldsymbol{X'}_{it}\boldsymbol{\beta}, \quad (2.6)$$

where the parameters in $var(\boldsymbol{Y}) = \boldsymbol{\Sigma}$ are nuisance parameters with an arbitrarily chosen pattern.

These models include no random effect and are population averaged models such as GEE and GMM. According to Hansen (2007), marginal approaches are appropriate

when the researcher seeks to examine cluster level data, which is when inferences about the population average are of primary interest (Diggle, Liang, & Zeger, 1994) or when the expected values of the responses as a function of the current covariates are the applicable necessary results (Pepe & Anderson, 1994). For the example mentioned above about academic success of students, if the goal of a study is to compare the academic success between clusters or majors or to compare the academic success between males and females, a marginal model would be appropriate (Zorn, 2001). These models are called marginal because the mean response model at each occasion depends only on the covariates of interest, not like conditional models, which depend on previous responses and random effects.

Generalized Estimating Equations

For analyzing marginal models, Liang and Zeger (1986) developed the GEE approach as a multivariate extension of quasi-likelihood used to estimate the regression coefficients without completely specifying the response distribution. In this approach, a "working" correlation structure for the correlation between a subject's repeated measurements is proposed by Liang and Zeger (1986).

Assume n subjects are repeatedly measured over T times as before with J covariates $j = 1, \dots, J$. Let Y_{it} denote a response variable observed repeatedly at different time points. It is also possible that these repeated measures are observed within an unbalanced longitudinal data but for the sake of simplicity, balanced data are considered here. Suppose $\boldsymbol{X}_{it}$ is a covariates matrix including a $(r \times 1)$ vector of covariates associated with each response, Y_{it}. The marginal model is a regression model which separately models the mean response and the within-subject association among repeated

measures of the response variable. These three parts are the main features of marginal models that need to be specified:

1. $E(Y_{it}|\boldsymbol{X}_{it}) = \mu_{it}$ is the conditional expectation of each response which is assumed to be dependent on the covariates through a known link function $g(.)$. Therefore, the conditional expectation can be written as $E(Y_{it}|\boldsymbol{X}_{it}) = \mu_{it} = g(\boldsymbol{X}_{it}^T\boldsymbol{\beta})$ where $\boldsymbol{X}_{it}$ represents the covariates matrix and $\boldsymbol{\beta}$ represents the vector of parameters of interest.
2. $Var(Y_{it}) = \psi v(\mu_{it})$ is the conditional variance of each response given $\boldsymbol{X}_{it}$ which is assumed to be dependent on the mean and also on the covariates, $\boldsymbol{X}_{it}$. $v(\mu_{it})$ is a known variance which is a function of the mean and ψ is a possibly unknown scale or dispersion parameter. This scale parameter can be fixed and known or unknown in the estimation of such models.
3. Given the covariates, the conditional within-subject associations among the vector of repeated responses are assumed to be a function of an additional vector of association parameters. The conditional within-subject associations can be specified as $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{A}$ where $\boldsymbol{R}_i$ is the working correlation matrix that may depend on a vector of unknown parameters, $\boldsymbol{\alpha}$. In general, the assumed covariance among the responses can be written as below and referred to as working covariance within GEE emphasizing the fact that $\boldsymbol{V}_i$ is only an approximation to the true covariance which can be approximated as

$$\boldsymbol{V}_i(\boldsymbol{\alpha}) = \phi \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{\frac{1}{2}}, \tag{2.7}$$

where $\boldsymbol{A}_i = diag\{v(\mu_{it})\}$ is a diagonal matrix with diagonal elements $v(\mu_{it})$, which are specified entirely by the marginal means, by $\boldsymbol{\beta}$ and $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is a ($T \times$

T) correlation matrix, referred to as the working correlation within GEE and ϕ is an unknown scale or dispersion parameter.

Within the correlation matrix, $\boldsymbol{\alpha}$ represents a vector of parameters associated with a specified model for $Corr(\boldsymbol{Y}_{it})$, with typical element

$$\rho_{ist} = \rho_{ist}(\boldsymbol{\alpha}) = Corr(Y_{is}, Y_{it}; \boldsymbol{\alpha}), \quad s \neq t. \tag{2.8}$$

As both the $\boldsymbol{R}_i(\boldsymbol{\alpha})$ and $Var(Y_{it})$ can be incorrectly specified, the use of "working covariance" is preferred by many statisticians. When $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{I}$, the GEE is reduced to the quasi-likelihood estimating equation for a GLM that assumes the repeated observations are independent by the use of an identity matrix for the working correlation (Fitzmaurice et al., 2009). There are other correlation structures within GEE such as autoregressive, unstructured, and exchangeable that can be found in more detail in Liang and Zeger (1986) and Prentice (1988).

According to Fitzmaurice et al. (2009), the first two components of GEE correspond to the standard GLM with no distributional assumption. It is the third part that represents the main extension of the GLM to the longitudinal data. Therefore, the steps that marginal models take to make this extension first specify a GLM for longitudinal responses at each occasion and additionally include a model for the within-subject association among the repeated responses. Separately modeling the mean response and the association among responses is important in the interpretation of the regression parameters, $\boldsymbol{\beta}$, in the model for the mean response. The population-averaged interpretations of the $\boldsymbol{\beta}$ describe how the mean response in the population is related to the covariates.

The avoidance of the distributional assumptions is what makes these marginal models so unique and important as the specification of the joint multivariate distribution of $\boldsymbol{Y}_i$, the vector of responses, is not always possible. This avoidance of the full distributional assumptions is the reason for these models to be considered semi-parametric due to having $\boldsymbol{\beta}$ as a parametric component as well as a non-parametric component, which is determined by the nuisance parameters by the moments higher than just the first-order moments.

Assuming that there are n independent observations of a scalar response variable, $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ are the covariates associated with the response, the GEE estimator of $\boldsymbol{\beta}$ can be found as in Equation 2.9.

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} \boldsymbol{X}_i'[\boldsymbol{R}_i(\widehat{\boldsymbol{\alpha}})]^{-1}\boldsymbol{X}_i\right]^{-1}\left[\sum_{i=1}^{n} \boldsymbol{X}_i'[\boldsymbol{R}_i(\widehat{\boldsymbol{\alpha}})]^{-1}\boldsymbol{X}_i\right]. \tag{2.9}$$

A valuable feature of GEEs with time-independent covariates is that they produce efficient estimates if the working correlation structure is correctly specified (Lai & Small, 2007). GEE estimators remain consistent and provide correct standard errors even if the working correlation structure is incorrectly specified. However, when there are time-dependent covariates, Hu (1993) and Pepe and Anderson (1994) pointed out that the consistency of GEEs is not assured with arbitrary working correlation structures unless a key assumption is satisfied. When there are time-dependent covariates, Pepe and Anderson (1994) suggested that marginal models be estimated by generalized estimating equations with the independent working correlation in the presence of time-dependent covariates. Fitzmaurice et al. (2009) showed in detail the loss of efficiency when using a

GEE approach to estimate the unknown parameters of longitudinal models in the presence of time-dependent covariates.

According to Lai and Small (2007), GMM is a more efficient estimation approach for marginal regression models with time-dependent covariates. GEEs with the independent working correlation do not exploit all of the available estimating equations involving any time-dependent covariate. GMM, on the other hand, makes efficient use of all the estimating equations that are made available by time-dependent covariates providing more efficient estimates than GEEs with the independent working correlation under certain conditions. GMM also maintains the GEE approach with time-independent covariates' attractive feature of being consistent under all correlation structures for subjects' repeated measurements (Lai & Small, 2007).

Generalized Method of Moments

GMM was first introduced in the econometrics literature by Lars Hansen in 1982 and has had a large influence in econometrics (Hansen, 1982). From then, it has been developed and widely used by taking advantage of numerous statistical inference techniques. GMM has been used in agriculture, business cycles, commodity markets consumption, economics growth, education, environmental economics, equity pricing, health care, import demand, interest rates, inventories, investment, macroeconomic forecasts, microstructures in finance, technological innovation, and many other areas of economics (Hall, 2005).

Unlike ML estimation, GMM does not require complete knowledge and specification of the distribution of the data. Only specified moments derived from an underlying model are what GMM estimator needs. This method, under some

circumstances, is even superior to the ML estimator, which is one of the best available estimators for the classical statistics paradigm since the early 20th century (Hall, 2005). MLE performs well only if the distribution of the data are completely and correctly specified. However, this specification is not always possible. This problem happens under economic theory leaving researchers with the arbitrary choice of distribution. As a result of this limitation, an optimal estimator might not exist, which will possibly cause biased inferences under ML estimation. These circumstances also include the computational burden of MLE and its dependence on the joint probability distribution of the data, known as the likelihood function. So, even if the choice of the distribution coincides with the truth, with the currently available computer technology, numerically evaluating the likelihood function of the joint probability distribution would be burdensome. Another computational burden will be added to some models when more parameters need to be added to the model to complete the distributional specification of the data. Some models, specifically within economics, do not specify all aspects of the probability distribution of the data due to their parameters' nature. This is very burdensome within MLE as under these circumstances, the likelihood needs to be maximized according to some nonlinear constraints implied by such models while trying to estimate many additional parameters (Hall, 2005). Additionally, in models for which there are more moment conditions than model parameters, GMM estimation provides a straightforward way to test the specification of the proposed model, which is an important feature that is unique only to GMM estimation.

In contrast to the disadvantages of MLE mentioned above, GMM provides a computationally convenient framework for making inferences within such models

without the necessity of specifying the likelihood function (Hall, 2005). GMM, which has roots in the minimum χ^2 method, is an estimation procedure that enables researchers to avoid unwanted or unnecessary assumptions such as distributional assumptions regarding the model they try to fit. This type of model can be considered semi-parametric as the full shape of the distributional functions of data may not be known but the parameter of interest is finite-dimensional.

Within GMM, a certain number of moment conditions, which are functions of the model parameters and the data, need to be specified for the model. These moment conditions have the expectation of zero at the true values of the parameters. Through GMM models, consistent, efficient, and asymptotically normally distributed estimators are estimated that do not need to use any information other than the information that is contained in the moment conditions. This method also takes account of both sampling and estimation error by its unique way of constructing tests (Hansen, 2007).

According to Hansen (2007), GMM estimation begins with a vector of population moment conditions taking the form below for all t

$$E[f(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)] = 0, \tag{2.10}$$

where $\boldsymbol{\beta}_0$ is an unknown vector in a parameter, $\boldsymbol{x}_{it}$ is a vector of random variables, $i = 1, \dots, n;\ t = 1, \dots, T$ and $f(.)$ is a vector of functions.

The GMM estimator is the value of $\boldsymbol{\beta}$ which minimizes a quadratic form in weighting matrix, $\boldsymbol{W}$, and the sample moment $n^{-1}\sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})$. This quadratic form is shown in Equation 2.11.

$$Q(\boldsymbol{\beta}) = \{n^{-1}\textstyle\sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})\}' \boldsymbol{W} \{n^{-1}\textstyle\sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})\}, \tag{2.11}$$

where $\boldsymbol{W}$ is a positive semi-definite matrix which may depend on the data but converges in probability to a matrix of constants which is positive definite. By definition, the GMM estimator of $\boldsymbol{\beta}_0$ is

$$\widehat{\boldsymbol{\beta}} = arg\ \underset{\boldsymbol{\beta}\in\mathbb{P}}{min}\ Q(\boldsymbol{\beta}), \tag{2.12}$$

where $arg\ min$ stands for the value of the argument $\boldsymbol{\beta}$ which minimizes the function in front of it.

If some regularity conditions hold (Hall, 2005), then the first order conditions for this minimization imply

$$\frac{\partial Q(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = 0. \tag{2.13}$$

Solving Equation 2.13 provides the closed form solution for $\widehat{\boldsymbol{\beta}}$ as a function of data in linear models. Unfortunately, in non-linear models, this is typically impossible.
This calculation takes a lot of steps which can be done using a computer based routine. The process begins with some trial value of $\boldsymbol{\beta}$ which can be called $\boldsymbol{\beta}(0)$. If this is the value that minimizes $Q(\boldsymbol{\beta})$, then it should not be possible to find a value of $\boldsymbol{\beta}$ for which the minimand is smaller. Using some rules, the computer tries to find a possible value of $\boldsymbol{\beta}$, for example $\boldsymbol{\beta}(1)$, which satisfies $Q(\boldsymbol{\beta}[1]) < Q(\boldsymbol{\beta}[0])$. If this new value is found such that it meets the criterion mentioned above, $\boldsymbol{\beta}(1)$ becomes the new candidate value for $\widehat{\boldsymbol{\beta}}$ and the computer searches again for another possible value which is smaller than $\boldsymbol{\beta}(1)$, say $\boldsymbol{\beta}(2)$, such that $Q(\boldsymbol{\beta}(2)) < Q(\boldsymbol{\beta}(1))$. This updating process continues until it is judged that the value of $\boldsymbol{\beta}$ which minimizes $Q(\boldsymbol{\beta})$ has been found. Three aspects of this routine need to be considered before beginning the estimation procedure:

1. The starting value for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}(0)$ needs to be specified. Ideally $\boldsymbol{\beta}(0)$ needs to be as close as possible to the value which minimizes $Q(\boldsymbol{\beta})$ since meeting this condition reduces the number of iterations and hence the computational burden.
2. The iterative search method, by which the candidate value of $\widehat{\boldsymbol{\beta}}$ is updated on each step, needs to be conducted. In most of the problems, it is computationally infeasible to perform a search over the entire parameter space and some rules are used to limit the required calculations. For example in a class known as gradient methods, the value of $\boldsymbol{\beta}$ is updated on the ith step by

$$\boldsymbol{\beta}(i) = \boldsymbol{\beta}(i-1) + \xi_i D\big(\boldsymbol{\beta}(i-1)\big), \tag{2.14}$$

where ξ_i is a scalar known as the step size and $D(.)$ is a vector known as step direction which is a function of the gradient $\frac{\partial Q(\boldsymbol{\beta}(i-1))}{\partial \boldsymbol{\beta}}$ and hence reflects the curvature of the function at $\boldsymbol{\beta}(i-1)$. $D(\boldsymbol{\beta}(i-1))$ determines the direction in which to update $\boldsymbol{\beta}(i-1)$ and ξ_i determines how far to go in that direction.
3. The convergence criterion used to judge when the minimum has been reached needs to be specified next. This convergence can be assessed in a number of different ways. For example, if $\boldsymbol{\beta}(i)$ is the value which minimizes $Q(\boldsymbol{\beta})$, then the updating routine should not move away from this point, suggesting that the minimum has been found if

$$\|\boldsymbol{\beta}(i+1) - \boldsymbol{\beta}(i)\| < \varepsilon, \tag{2.15}$$

where ε is an arbitrarily small positive constant. A typical value of ε is 10^{-6} or less. Convergence can be assessed by a number of other ways which can be found in Hall (2005).

Seven elements of the GMM framework according to Hall (2005) are as below. The first element is identification, which refers to the importance of the population moment conditions in having a successful estimation and how they must not only be valid but also provide sufficient information to identify the parameter vector. Decomposition of moment conditions into identifying restrictions which contain the information that goes into the estimation and over-identifying restrictions, which are a reminder that manifests itself in the estimated sample moment is the second element of this framework. The third element of this framework describes the asymptotic properties saying that when the consistent GMM estimator is appropriately scaled, it has a normal limiting distribution, which is important for hypothesis testing and performing other inferences. The fourth element pertains to the estimated sample moment, which is shown to have a limiting normal distribution with the characteristics that directly depend on the function of data in the over-identifying restrictions. Long run covariance estimation, which emphasizes the necessity of consistently estimating the long run variance of the sample moment while trying to use the asymptotic normality in practical inference procedures, forms the fifth element. The sixth element is the optimum choice of weighting matrix, which depends on the long run variance of the sample moment. The last element of a GMM framework is about model diagnostics, which considers the bias provided for testing the validity of the GMM model specification via the estimated sample moments.

Moment Selection for Longitudinal Data

Moment selection is an important and yet challenging part of GMM models, which is one of the best and most efficient models when modeling longitudinal data especially in the presence of time-dependent covariates (Lai & small, 2007). The

desirable properties of the selected moments depend upon the question that needs to be answered in a study. Hall (2005) assumed that the objective of the study is mostly in regard to making inferences about an unknown parameter vector, $\boldsymbol{\beta}_0$, based on some asymptotic theories. Under this context, he argued it is desirable for the selected vector to satisfy three conditions. The first condition is the orthogonality condition, which refers to the fact that the estimation should be based on valid information. The efficiency condition is the second condition that emphasizes the importance of making the inference based on estimates that are asymptotically the most precise ones. The third condition is the non-redundancy condition so that the selected moment condition does not suffer from redundancy of elements resulting in declining the asymptotic approximation quality to finite sample behavior.

There are two existing approaches to resolve this issue of moment selection in general. The first option is finding the optimal moment condition theoretically, which is the one that satisfies both the orthogonality and efficiency conditions. The score vector will always be the optimal moment condition as it will result in the GMM estimator, which also is the ML estimator. Unfortunately, within many models, this option is infeasible. Therefore, more restrictions forcing more practical settings are necessary. The second approach, which seems more realistic, is to develop data-based methods for moment selection. This is a more practical approach as in most circumstances a researcher needs to decide about the moments without any knowledge of the underlying data generation process and only based on the data. The only point that needs to be considered within this approach is that the use of the data does not contaminate the limiting distribution theory as the moment selection must perforce be based upon the

data. This introduces some other criteria, which are not discussed here as this is not the main topic of the current dissertation. For more details regarding moment selection, see Hall (2005).

When analyzing longitudinal data, there are two types of correlations that need to be taken into consideration; the correlation inherent from the repeated measures of the responses and the correlation due to the feedback created between the responses at a particular time and the predictors at other times. These added complexities will make the process of finding the moment conditions more complicated. When using a generalized method of moments for estimating the coefficients in such data, the necessity of taking approaches that make use of all the valid moment conditions with each time-dependent and time-independent covariate is what is highlighted in some references (Lalonde, Wilson, & Yin, 2014).

Lai and Small (2007) suggested using GMM for longitudinal models in a way to use optimal information provided by time-dependent covariates, when obtaining estimates. The choice of moment conditions within their approach depends on the type of time-dependent covariates, which they classified into three types. Type I and type II time-dependent covariates are covariates for which there is no "feed-back" from the response process to the covariate process. Type I time-dependent covariates have the additional feature which is based on the situation of past values of the covariate being uncorrelated with current residuals.

For the repeated observations taken over T times on n subjects with J covariates, assume that observations y_{is} and y_{kt} are independent whenever $i \neq k$. Making the decision about the type of time-dependent covariates is based on the equation

$$E\left[\frac{\partial \mu_{it}(\boldsymbol{\beta})}{\partial \beta_j}\{y_{it} - \mu_{it}(\boldsymbol{\beta})\}\right] = 0, \tag{2.16}$$

where $\mu_{it}(\boldsymbol{\beta})$ represents the expectation of $\boldsymbol{y}_{it}$ based on the vector of covariate values, $\boldsymbol{x}_{it}$ and $\boldsymbol{\beta}$ denotes the vector of parameters that describes the marginal distribution of y_{it}. If Equation 2.16 holds for all s and t, then the jth covariate is classified as type I. Type I covariates plausibly satisfy a condition that their outcomes are independent of past and future outcomes of the response. For this type of covariate, there will be T^2 moment conditions. Variables like age, time variables, and treatment assignment for each subject at a certain time point in a randomized crossover trial can be classified into type I covariates (Lai & Small, 2007).

If Equation 2.16 holds for $s \geq t$ but fails to hold for some $s < t$, the jth covariate is said to be type II**.** This type of covariate is used in many time-series models. This type of covariate is common in a linear model with autoregressive responses (Lalonde et al., 2014). For each of the type II covariates, there will be $\frac{T(T+1)}{2}$ moment conditions.

If Equation 2.16 fails to hold for any $s > t$, the jth covariate is said to be type III**.** This will occur if there is some feedback loop or common response to an omitted variable; therefore, this type of covariate occurs when it changes randomly and its distribution may depend on past values of the response. There will be T moment conditions valid for each type III covariate. To clarify the distinction between types II and III of time-dependent covariates, the study of infectious diseases and vitamin A deficiency in Indonesian children, which was first presented by Zeger and Liang (1991), is considered here. Considering diarrheal disease as an outcome variable and xerophthalmia, which is an ocular condition due to vitamin A deficiency, as the time-

dependent covariate, xerophthalmia can be specified as a type III covariate. It is because there is a feedback cycle in which xerophthalmia increases the risk of diarrheal disease, which further increases the risk of future xerophthalmia. In contrast, considering respiratory disease as an outcome and the same covariate of xerophthalmia, this time xerophthalmia is classified as a type II covariate because there is no evidence of a feedback cycle (Diggle et al., 1994).

Lalonde et al. (2014) argued that there can be theoretically more than three types of time-dependent covariates. Concentrating on using valid moment conditions, they provided a method to choose valid equations to determine the impact of time-dependent covariates on the response over time. In their recommended models, there is no need to classify the covariates into different types but in order to identify the appropriate moment conditions which result in consistent and efficient estimators, they revisited Lai and Small's (2007) procedures and defined the forth type of covariates before presenting their different yet related approach. Type IV covariate is in direct contrast to type II in which the future responses are not affected by the previous process so there is no feedback from the covariate process to the response process. For this type of covariate, there will be $\frac{T(T+1)}{2}$ moment conditions. Lalonde et al. (2014) used the example of a weight loss study for clarification. The weight loss will impact the blood pressure as the future covariate but the blood pressure has no impact on future weight loss. So, this covariate can be classified as a type IV covariate because the future responses are not affected by the previous covariate process and there is no feedback from the covariate process to the response process.

Lalonde et al. (2014) showed that incorrectly specifying the type of covariate may result in significant changes in the standard errors, hence inaccurate conclusions. This is why after embracing the approach by Lai and Small (2007) regarding classifying the variables into different types as well as adding a new type of covariate, Lalonde et al. (2014) moved to a more general method which depicts each of the valid moment conditions rather than designating them in a group of a certain type. Their approach is based on using a correlation technique to decide about the moment conditions that should be included. Then, using the continuously updating GMM or two-step GMM is recommended in obtaining estimates and selecting moment conditions without assuming that feedback is always present over time, or if present, occurs at the same degree. Continuously updated GMM results from continuing the multi-step procedure to obtain the iterated GMM estimator. This approach was first suggested by Hansen, Heaton, and Yaron (1996) in which the dependence of the weighting matrix on the unknown parameters is acknowledged and taken care of during the optimization procedure. There is fairly compelling evidence to suggest there are gains to iteration in terms of finite sample performance of the estimator but in most cases the two-step estimator is applied. Two-step estimators on the other hand benefit from not having the numbers of equations and parameters in the nonlinear GMM step grow with the number of perfectly measured regressors, conferring a computational simplicity (Erickson & Whited, 2002). For more details about these two approaches in moment selection see Lai and Small (2007) and Lalonde et al. (2014).

Power

The power of a statistical test can be taken to be the probability of obtaining statistically significant results when testing a false null hypothesis, H_0, against a specific alternative hypothesis, H_a. Statistical power depends on the sample size (n), significance criterion (α), type of test, and the population effect size among other things (Cohen, 1992).

According to Cohen (1992), power analysis is a very important aspect of most of the studies especially in social and behavioral sciences as in every single study, researchers are trying to formulate and test different null hypotheses with the hope of rejecting them to proceed to establish facts about the phenomena under study.

The power function, $\pi(\theta)$, is the probability of rejecting the null hypothesis, H_0, when the true value of the parameter is θ_1 for a simple hypotheses $H_0: \theta = \theta_0$ versus $H_a: \theta = \theta_1$. This probability can be specified as

$$Power = P(reject\ H_0 | H_1 is\ true). \tag{2.17}$$

The computation of the power of a hypothesis test can be summarized in three steps. These steps include defining the region of acceptance, specifying the critical parameter value, which is an alternative to the value specified in the null hypothesis and finally calculating the power. The effect size can be found by using the difference between the critical parameter value and the value from the null hypothesis. When the null hypothesis is false and the researcher's hypothesis is true, the effect size will be greater than zero. The power of the test for such positive effect size is the probability that the test will lead to rejecting the null hypothesis, which provides support for the theory. This will form the last step, which is computing the power after assuming that the true

population parameter is equal to the critical parameter value, rather than the value specified in the null hypothesis. Based on that assumption, the probability of the sample estimate of the population parameter falling outside the region of acceptance is the power of the test (Kraemer & Blasey, 2015).

Derivation of the minimum sample size in applied research is an important component that needs to be considered at the design stage of any study that is designed by researchers to address some scientific hypotheses. It is important for researchers to come up with the correct sample size they need to perform a hypothesis test and make inferences. The ideal sample size is the one that is not too small to rob a study of power to detect the significant effects when they actually exist and not too large to be very time-consuming and costly to perform or lead to over-powered tests (Rochon, 1998). Usually there is no formula for the power of different tests, but power is estimated for different values of sample size and based on the preferred value of the power, the minimum sample size can be chosen.

When trying to calculate the power of the tests within longitudinal studies to figure out the required sample size, the process is more complicated than it is for cross-sectional data. In general, in any study including longitudinal data, in order to perform power analyses and sample size calculations, one needs to examine the asymptotic mean and variance of the test-statistic under both the alternative and null hypotheses. After specifying the significance level and possibly the parameter values, the sample size needed to test the hypothesis can be computed in different ways, three of which are explained below in detail. These three methods, which may be used within longitudinal studies, are based on using the Wald test, the likelihood ratio test, and the score test.

Power Calculation Using the Wald Test

Rochon (1998) adopted GEE as the underlying statistical approach to sample size calculations on longitudinal continuous and discrete responses. This approach allows practitioners to design the study with the same analytic procedure that will be applied in the analysis. If the underlying assumptions are correctly specified, adequate power will be calculated to detect significant differences in the study using a GEE analysis. This approach uses the damped exponential family of correlation structures. Under this approach, the correlation between two observations separated by T time points is $\varphi^{T^{\theta}}$, where φ is the correlation between observations separated by one unit of time and θ is a damping parameter.

Assume the repeated measures are recorded at the same set of time points $t = \{1, 2, \dots, T\}$ for all the subjects of the study. For this hypothetical study, each subject is considered as a cluster. Assuming there are $i = 1, \dots, n$ clusters or subjects, $\boldsymbol{\mu}'_i = [\mu_{i1} \dots \mu_{iT}]$ is the vector of expected values across the repeated measures of the ith subject and $\boldsymbol{X}_i$ is a $(T \times r)$ design matrix in the ith cluster. Let $\boldsymbol{X} = \boldsymbol{I}_i \otimes \boldsymbol{X}_i$ be the overall design matrix across all the n subjects where $\otimes$ represents the outer product of two vectors which forms a matrix. The repeated measure response matrix can be specified as

$$\boldsymbol{Y}_i = [\boldsymbol{Y}_{i1} \dots \boldsymbol{Y}_{iT}] = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1T} \\ y_{21} & y_{22} & \cdots & y_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nT} \end{bmatrix},$$

where $i = 1, \dots, n$. The regression model can be written as

$$g[E(Y_{it}|\boldsymbol{X}_{it} = \boldsymbol{x}_{it})] = \boldsymbol{x}_{it}\boldsymbol{\beta}, \tag{2.18}$$

where $g(.)$ is a known link function, $\boldsymbol{x}_{it}$ is a $(1 \times r)$ vector and finally $\boldsymbol{\beta}$ is a $(r \times 1)$ vector of regression coefficients that needs to be estimated.

In order to find the power and then the minimum sample size of a statistical test, first the hypothesis needs to be specified. Suppose the specific desired hypothesis can be expressed as

$$\begin{cases} H_0: \boldsymbol{H\beta} = \boldsymbol{h}_0 \\ H_1: \boldsymbol{H\beta} \neq \boldsymbol{h}_0' \end{cases} \tag{2.19}$$

where $\boldsymbol{H}$ is an $(h \times r)$ full rank matrix and $\boldsymbol{h}_0$ is an $(h \times 1)$ conformable vector of constant elements.

Within this test, the vector of the parameters, $\boldsymbol{\beta}$, can be estimated using different estimating techniques. Adopting the GEE method, after assuming the same design matrix, mean vector and covariance matrix within each of the clusters, we may take sums across individuals and use them to find the estimators using GEE. According to McCullagh and Nelder (1989), the estimator of $\boldsymbol{\beta}$ using GEE can be found using this equation

$$\widehat{\boldsymbol{\beta}} = \left[\sum_i \boldsymbol{X'}_i \boldsymbol{W}_i \boldsymbol{X}_i\right]^{-1} \left[\sum_i \boldsymbol{X'}_i \boldsymbol{W}_i h(\boldsymbol{\mu}_i)\right], \tag{2.20}$$

where $\boldsymbol{W}_i = \boldsymbol{\Delta}_i' \boldsymbol{V}_i^{-1} \boldsymbol{\Delta}_i$. According to Rochon (1998), this estimated $\boldsymbol{\beta}$ has the model-based covariance matrix, $cov_{MB}(\widehat{\boldsymbol{\beta}})$, that can be calculated using the equation below

$$cov_{MB}(\widehat{\boldsymbol{\beta}}) = \left[n \sum_i \boldsymbol{D'}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right]^{-1}. \tag{2.21}$$

According to Liang and Zeger (1986), the robust covariance matrix for $\widehat{\boldsymbol{\beta}}$, $cov_R(\widehat{\boldsymbol{\beta}})$, is obtained using the sandwich estimator as

$$cov_R(\widehat{\boldsymbol{\beta}}) = n^{-1}\left[\sum_i \boldsymbol{D'}_i \boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right]^{-1}\left[\sum_i \boldsymbol{D'}_i \boldsymbol{V}_i^{-1}\boldsymbol{\Gamma}_i\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right]\left[\sum_i \boldsymbol{D'}_i \boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right]^{-1}, \quad (2.22)$$

where $\boldsymbol{\Gamma}_i$ is the true covariance matrix among the set of repeated measures in the ith cluster defined as

$$\boldsymbol{\Gamma}_i = Var(\boldsymbol{Y}_i).$$

This robust covariance is used to protect the inferences from deviations in the working covariance structure $\boldsymbol{V}_i$ from the true covariance pattern $\boldsymbol{\Gamma}_i$. $\boldsymbol{V}_i$ which is used in Equation 2.22 and can be defined as

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

This is difficult to do at the design stage as little is known about the true covariance structure and one needs to wait until starting the analysis stage to calculate the residuals for estimating $\boldsymbol{\Gamma}_i$.

The parameter estimation from above can be applied in calculating the Wald test statistic, which is utilized for testing the aforementioned null hypothesis in Equation 2.19. The Wald test statistic has an approximate chi-square distribution

$$T_W = n\left(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h_0}\right)'\left[\boldsymbol{H}\widehat{var}\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\psi}\right)\boldsymbol{H'}\right]^{-1}\left(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h_0}\right) \sim \chi^2_{(h),\lambda_W}, \quad (2.23)$$

where $\widehat{var}\left(\widehat{\boldsymbol{\beta}}\right)$ can be either the estimate of the model-based covariance matrix or the estimate of the robust covariance matrix for $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\psi}$ is a vector of scale or dispersion parameters. This chi-square distribution has the approximate non-centrality parameter of λ_W which can be approximated as

$$\hat{\lambda}_W \approx n\left(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h_0}\right)'\left[\boldsymbol{H}\widehat{var}\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\psi}\right)\boldsymbol{H'}\right]^{-1}\left(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h_0}\right). \quad (2.24)$$

So, sample size can be estimated as

$$n \approx \frac{\hat{\lambda}_W}{\left(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h_0}\right)'\left[\boldsymbol{H}\widehat{var}\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\psi}\right)\boldsymbol{H}'\right]^{-1}\left(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h_0}\right)}. \tag{2.25}$$

In order to estimate the power, a specific vector needs to be specified for the alternative hypothesis, $\boldsymbol{h}_1$, so the original hypothesis should be stated as

$$\begin{cases} H_0: \boldsymbol{H\beta} = \boldsymbol{h}_0 \\ H_1: \boldsymbol{HB} = \boldsymbol{h}_1 \end{cases}. \tag{2.26}$$

Assuming that α represents the type I error, $\chi^2_{(h);1-\alpha}$ is the critical value from the central $\chi^2_{(h)}$ distribution. Using this critical value, power can be calculated by finding the probability

$$\Pr\left(\chi^2_{h,(\lambda_W)} \geq \chi^2_{h,1-\alpha}\right), \tag{2.27}$$

for the Wald test, with $\chi^2_{h,1-\alpha}$ denoting the $100(1-\alpha)$th percentile of the central chi-square with h degrees of freedom. So, the power associated with the Wald test statistic is

$$1 - \gamma = \int_{\chi^2_{(h);1-\alpha}}^{\infty} f(x; h, \lambda_W)dx, \tag{2.28}$$

where γ represents the type II error and $f(x; h, \lambda_W)$ is the probability density function of $\chi^2_{(h),\lambda_W}$.

A strict application of the theory requires a true value of $\boldsymbol{\beta}$ and the exact covariance of its estimator. A consistent estimator of this parameter can be applied which will result in two circumstances. One is that T_W is only asymptotically distributed as a chi-square distribution which some believe might affect the efficiency. However, believing that efficiency is negatively affected is in disagreement with what Lipsitz, Fitzmaurice, Orav, and Laird (1994) suggested regarding the high efficiency of GEE procedures, even for small sample sizes. The other circumstance is that the non-centrality parameter for this asymptotic distribution is an approximation and so its influence is

unclear. However, neither of these two influences seem to be a problem with a large sample size (Rochon, 1998). After finding the power and solving the non-centrality equation for the minimum sample size, the required sample size for the particular hypothesis a researcher is considering can be estimated.

The Wald test statistic explained above can also be used to estimate the conditional power calculated for an appropriate expanded data set. Lyles et al. (2007) came up with this method of estimating the power with no dependence on the assumed distribution of the response variable to an expanded data set composed of one record for each possible value of the outcome per combination of covariate value. The procedure for creating this expanded dataset is briefly explained at the end of this chapter.

Power Calculation Using the Likelihood Ratio Test

Having the same regression model of this general form as Equation 2.18,

$$g[E(\boldsymbol{Y}_{it}|\boldsymbol{X}_{it} = \boldsymbol{x}_{it})] = \boldsymbol{x}_{it}\boldsymbol{\beta},$$

the vector of regression coefficients, $\boldsymbol{\beta}$, needs to be estimated.

Trying to test the hypothesis from Equation 2.19, the likelihood ratio (LR) test statistic is given by

$$T_{LR} = -2[l(\widehat{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\psi}}^*) - l(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}})], \tag{2.29}$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\psi}}$ are unrestricted ML estimators of $\boldsymbol{\beta}$ and a vector of $\boldsymbol{\psi}$ of scale or dispersion parameters, $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\boldsymbol{\psi}}^*$ are the corresponding ML estimators under the null hypothesis and $l(.)$ denotes the log-likelihood function.

T_{LR} follows an asymptotic central chi-square distribution with h degrees of freedom under the null hypothesis. The distribution of this test-statistic under the

alternative hypothesis is a non-central chi-square distribution, $\chi^2_{h,(\lambda_{LR})}$, with the non-centrality parameter specified by λ_{LR},

$$\lambda_{LR} = -2[l^*(\boldsymbol{\beta}, \boldsymbol{\psi}) - l(\boldsymbol{\beta}, \boldsymbol{\psi})], \tag{2.30}$$

where $l(\boldsymbol{\beta}, \boldsymbol{\psi})$ is the log-likelihood evaluated at the true parameters and $l^*(\boldsymbol{\beta}, \boldsymbol{\psi})$ is the log-likelihood evaluated at the true parameters after imposing the restrictions designated under the null hypothesis. Testing the hypothesis (2.26) with a specified vector of $\boldsymbol{h}_1$ for the alternative hypothesis, power can be calculated as

$$\Pr\left(\chi^2_{h,(\lambda_{LR})} \geq \chi^2_{h,1-\alpha}\right), \tag{2.31}$$

with $\chi^2_{h,1-\alpha}$ denoting the $100(1-\alpha)th$ percentile of the central chi-square with h degrees of freedom where α represents the type I error also known as the critical value from the central $\chi^2_{(h)}$ distribution. Using this critical value, the power associated with T_{LR} likelihood ratio test statistic is

$$1 - \gamma = \int_{\chi^2_{(h);1-\alpha}}^{\infty} f(x; h, \lambda_{LR}) dx, \tag{2.32}$$

where γ represents the type II error and $f(x; h, \lambda_{LR})$ is the probability density function of $\chi^2_{(h),\lambda_{LR}}$.

Power Calculation Using the Score Test

Liu and Liang (1997) developed the use of score tests in the process of sample size and power calculation for correlated observations by extending what Self and Mauritsen (1988) did before for cross sectional studies. Within this multivariate extension, Liu and Liang (1997) used a quasi-score test statistic based on GEE models to derive the minimum sample size needed. The likelihood ratio-based model would not be

feasible for GEE models in general because the complete distribution function is often unspecified.

Liu and Liang (1997) used the term "sample size" in their paper as the number of clusters in which the cluster is formed by subjects in longitudinal studies. They first came up with a test statistic for correlated data.

Considering the general regression Equation 2.18,

$$g[E(\boldsymbol{Y}_{it}|\boldsymbol{X}_{it} = \boldsymbol{x}_{it})] = \boldsymbol{x}_{it}\boldsymbol{\beta},$$

This time two sets of vectors of covariates $\boldsymbol{X}_{ij}$ and $\boldsymbol{X}_{n_{ij}}$ are considered

$$g[E(\boldsymbol{Y}_{it}|\boldsymbol{X}_{it} = \boldsymbol{x}_{it})] = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{x}_{n_{it}}\boldsymbol{\beta}_n, \tag{2.33}$$

where $\boldsymbol{\beta}$ is $(p \times 1)$ vector of the parameters of interest and $\boldsymbol{\beta}_n$ is a $(q \times 1)$ vector of nuisance parameters. Testing the hypothesis (2.26), the quasi-score statistic based on GEE is as below

$$T = S_\beta\big(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{n_0}, \alpha\big)' \Sigma_0^{-1} S_\beta\big(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{n_0}, \alpha\big), \tag{2.34}$$

where α is the parameter used to specify the exchangeable or autoregressive correlations, $\boldsymbol{\beta}_0$ is a vector of parameters of interest under the null hypothesis and $\widehat{\boldsymbol{\beta}}_{n_0}$ is the estimator of $\boldsymbol{\beta}_n$ under H_0. The covariance matrix under the null hypothesis, Σ_0, as well as the score function, $S_\beta\big(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{n_0}, \alpha\big)$, are defined as below where $\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_n\end{bmatrix}$,

$$S_\beta\big(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{n_0}, \alpha\big) = \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)' V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i), \tag{2.35}$$

$$\Sigma_0 = cov_{H0}\big[S_\beta\big(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{n_0}, \alpha\big)\big], \tag{2.36}$$

and $\widehat{\boldsymbol{\beta}}_{n_0}$, which is the estimator of $\boldsymbol{\beta}_n$ under H_0, can be obtained from solving

$$S_{\beta_n}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_n, \alpha) = \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}_n}\right)' V_i^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) = 0, \tag{2.37}$$

where $\boldsymbol{V}_i$ is the covariance matrix of $\boldsymbol{y}_i$, characterized by parameters α and $\boldsymbol{\mu}_i = E(\boldsymbol{y}_i)$. Under the null hypothesis, as $n \to \infty$, T converges to a χ_p^2 distribution; however, under the alternative hypothesis, T converges to an asymptotic non-central chi-square distribution with the non-centrality parameter as

$$\lambda = \xi' \Sigma_1^{-1} \xi, \tag{2.38}$$

where ξ is the expectation of $S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}_n)$ under H_1 and is approximated by

$$\xi = E_{H_1}[S_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{n_0})] \approx \sum_{i=1}^{n} P_i^* \boldsymbol{V}_i^{-1}(\boldsymbol{\mu}_i^1 - \boldsymbol{\mu}_i^*), \tag{2.39}$$

where $\boldsymbol{\mu}_{it}^1 = g^{-1}(\boldsymbol{X}_{it}' \boldsymbol{\beta}_1 + \boldsymbol{X}_{n_{it}}' \boldsymbol{\beta}_{n_1})$ and $\boldsymbol{\mu}_{it}^* = g^{-1}\left(\boldsymbol{X}_{it}' \boldsymbol{\beta}_0 + \boldsymbol{X}_{n_{it}}' \boldsymbol{\beta}_{n_0}^*\right)$.

The above are evaluated at $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_{n_0}^*$ in which $\boldsymbol{\beta}_{n_0}^*$ is the limiting value of $\widehat{\boldsymbol{\beta}}_{n_0}$ under given $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{n_1}$ as $n \to \infty$. This limiting value can be found by solving

$$\lim_{n \to \infty} n^{-1} E_{H_1}\left[S_{\boldsymbol{\beta}_n}\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}_{n_0}^*\right); \boldsymbol{\beta}_1, \boldsymbol{\beta}_{n_1}\right] = 0. \tag{2.40}$$

Σ_1 is the covariance of $S_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{n_0}, \alpha\right)$ under H_1 which is approximated by

$$\Sigma_1 = cov_{H_1}[S_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{n_0}\right) \approx \sum_i P_i^* V_i^{-1} cov_{H_1}(y_i) V_i^{-1} P_i^{*\prime}, \tag{2.41}$$

where

$$P_i^* = \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)' - I_{\boldsymbol{\beta}\boldsymbol{\beta}_n}^* I_{\boldsymbol{\beta}_n\boldsymbol{\beta}_n}^{*-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}_n}\right)',$$

$$I_{\boldsymbol{\beta}\boldsymbol{\beta}_n}^* = \sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)' V_i^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}_n}\right),$$

$$I_{\boldsymbol{\beta}_n\boldsymbol{\beta}_n}^* = \sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}_n}\right)' V_i^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}_n}\right).$$

After specifying all the essential elements for performing the hypothesis test, the statistical power for testing a null versus alternative hypothesis can be approximated from the non-central chi-square distribution mentioned above. Conversely, the non-central parameter can be derived by specifying the type I and type II errors. Within power analysis and sample size calculations, both values of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{n_1}$ need to be specified under the alternative hypothesis since the distribution of T, which is a function of $\boldsymbol{y}_{it}$, depends on both the parameter of interest, $\boldsymbol{\beta}$ and the nuisance parameter, $\boldsymbol{\beta}_n$.

To calculate the sample size, first assume that the cluster sizes are identical across the clusters, for all i for convenience. Assume the covariates $\{(\boldsymbol{x}_t, \boldsymbol{x}_{n_t}), t = 1, \dots, T\}$ have the joint distribution

$$P[\boldsymbol{x}_t = \boldsymbol{u}_{tl}, \boldsymbol{x}_{n_t} = \boldsymbol{v}_{tl}; t = 1, \dots, T] = \boldsymbol{\pi}_l, \quad l = 1, \dots, L, \tag{2.42}$$

where $\{(\boldsymbol{u}_{tl}, \boldsymbol{v}_{tl}; t = 1, \dots, T), l = 1, \dots, L\}$ are the L possible distinct values for $\{(\boldsymbol{x}_t, \boldsymbol{x}_{n_t}), t = 1, \dots, T\}$.

Taking the expectation with respect to the joint distribution specified above, it can be used to find ξ as

$$\xi = nE[\boldsymbol{P}^*\boldsymbol{V}^{-1}(\boldsymbol{\mu}^1 - \boldsymbol{\mu}^*)] = n\sum_{l=1}^{L} \boldsymbol{\pi}_l \boldsymbol{P}_l^* \boldsymbol{V}_l^{-1}(\boldsymbol{\mu}_l^1 - \boldsymbol{\mu}_l^*). \tag{2.43}$$

Then Σ_1 is reduced to

$$\Sigma_1 = nE\left(\boldsymbol{P}^*\boldsymbol{V}^{-1} cov_{H_1}(\boldsymbol{y}) \boldsymbol{V}^{-1}\boldsymbol{P}^{*\prime}\right) = n\sum_{l=1}^{L} \boldsymbol{\pi}_l \boldsymbol{P}_l^* \boldsymbol{V}_l^{-1} cov_{H_1}(\boldsymbol{y}_l) \boldsymbol{V}_l^{-1} \boldsymbol{P}_l^{*\prime}. \tag{2.44}$$

Defining $\tilde{\xi} = E[\boldsymbol{P}^*\boldsymbol{V}^{-1}(\boldsymbol{\mu}^1 - \boldsymbol{\mu}^*)]$ and $\tilde{\Sigma}_1 = E\left(\boldsymbol{P}^*\boldsymbol{V}^{-1} cov_{H_1}(\boldsymbol{y}) \boldsymbol{V}^{-1}\boldsymbol{P}^{*\prime}\right)$, the non-centrality parameter derived from a non-central chi-square distribution and the given valued of the nominal power and significance level of the test can now be expressed as

$$\nu = n\tilde{\xi}'\tilde{\Sigma}_1^{-1}\tilde{\xi}. \tag{2.45}$$

The sample size required to achieve the nominal power is approximately

$$n = \nu/(\tilde{\xi}'\tilde{\Sigma}_1^{-1}\tilde{\xi}). \tag{2.46}$$

Finally, Equation 2.40 can be expressed as Equation 2.47 below which is used to find the solution of λ_0^* :

$$\sum_{l=1}^{L} \pi_l \left(\frac{\partial \tilde{\boldsymbol{\mu}}_l^*}{\partial \boldsymbol{\beta}_n}\right)' \boldsymbol{V}_l^{-1}(\tilde{\boldsymbol{\mu}}_l^1 - \tilde{\boldsymbol{\mu}}_l^*) = 0. \tag{2.47}$$

The expected value of $\boldsymbol{y}$, known as $\tilde{\boldsymbol{\mu}}^1$, can be calculated given $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{\boldsymbol{n}_1}$. Equation 2.47 can be solved using the GEE method with the weights $\{\boldsymbol{\pi}_l, l = 1, \dots, L\}$.

All parameters in the models under the null and alternative hypotheses have to be specified in the process of sample size calculation so once the focus is on correlated observations, parameters α which represent within-cluster associations need to be included too. These parameters appear in the working covariance

$$\boldsymbol{V}_i = \boldsymbol{\Delta}_i^{\frac{1}{2}} \boldsymbol{R}(\alpha) \boldsymbol{\Delta}_i^{\frac{1}{2}}, \tag{2.48}$$

where $\boldsymbol{\Delta}_i = diag[var(\boldsymbol{y}_{i1}), var(\boldsymbol{y}_{i2}), \dots, var(\boldsymbol{y}_{iT})]$ and $\boldsymbol{R}(\alpha) = corr(\boldsymbol{y}_i)$ is an $(T \times T)$ working correlation matrix.

Common choices for the working correlation matrix are mentioned in Diggle et al. (1994) and Fitzmaurice, Laird, and Rotnitzky (1993) of which some are listed here. For an uncorrelated structure, $\boldsymbol{R}(\alpha) = \boldsymbol{I}$ can be used which is an $(T \times T)$ identity matrix. If there exists an exchangeable correlation structure, $corr(\boldsymbol{y}_{it}, \boldsymbol{y}_{ik}) = \alpha, t \neq k$ can be used. For an autoregressive correlation, $corr(\boldsymbol{y}_{it}, \boldsymbol{y}_{ik}) = \alpha^{|t-k|}$ is appropriate. These three correlation structures may be used for sample size calculations in practice. If there

exists a correlation that is unstructured, $corr(\boldsymbol{y}_{it}, \boldsymbol{y}_{ik}) = \alpha_{jk}$ may be used in which α contains $n(n-1)/2$ pairwise correlations (Liu & Liang, 1997).

To summarize the sample size calculation process for correlated observations based on a quasi-score test statistic, four main steps need to be taken. First, the regression model for the marginal mean and parameter values for both H_0 and H_1 need to be specified. Second, a working correlation structure along with its corresponding parameter values should be specified. Third, a distribution for the configuration on discrete covariates needs to be assumed. At the end, the weighted GEE at Equation 2.47 needs to be solved for $\boldsymbol{\beta}^*_{n_0}$. In addition, the non-central parameter needs to be evaluated and the sample size needs to be estimated using Equation 2.46.

The only disadvantage of this sample size formula in the univariate case is its sensitivity to the distribution of the covariates. This is one of the reasons that has led some researchers to use the likelihood-based sample size formula which outperforms the score test-based formula for univariate observations (Liu & Liang, 1997). One alternative, which uses the approximate likelihood ratio, can be found in the work done by Hanfelt and Liang (1995).

Due to the absence of power estimation and minimum sample size calculation techniques for GMM, in conjunction with the higher efficiency of the GMM estimation technique for longitudinal data in the presence of time-dependent covariates, it is important to develop such methods. To the best of my knowledge, no studies exist on power estimation and minimum sample size calculation of longitudinal data using a GMM estimation technique. Two GMM-based approaches were developed in the current study to help applied researchers and practitioners in calculating the required sample size

and the optimal power for longitudinal studies in the presence of time-dependent covariates. In the next chapter, different options using GMM for estimating power and minimum sample size for testing hypotheses in longitudinal studies with time-dependent covariates are assessed.

CHAPTER III

METHODOLOGY

This chapter is dedicated to examining different methods of estimating statistical power and required sample size when working with longitudinal data in the presence of time-dependent covariates. These methods are based on using the generalized method of moments (GMM) as it is a more efficient estimation technique for longitudinal studies with time-dependent covariates compared to other estimation techniques such as generalized estimating equations (GEE; Lai & Small, 2007).

This chapter includes four sections that reveal the methodology that was used for the current study. First, a summary of the research methods used in this study is provided. Second, the process of the GMM technique for obtaining estimates of parameters within longitudinal studies is presented. Third, the power estimation tools based on GMM are explained. Fourth, the data set and description of data simulation schemes and conditions for Monte Carlo simulation are described.

Introduction

The research questions given in Chapter I are addressed in this chapter to develop power estimation and minimum sample size calculation methods for tests using GMM with a focus on longitudinal data with time-dependent covariates. This dissertation addressed the following questions:

Q1 How can power be calculated for hypothesis tests using longitudinal data with time-dependent covariates applying a Wald approach within a GMM estimation technique?

Q2 How can sample size be calculated for a desired level of power for hypothesis tests using longitudinal data with time-dependent covariates applying a Wald approach within a GMM estimation technique?

Q3 How can power be calculated for hypothesis tests using longitudinal data with time-dependent covariates applying a Distant Metric Statistic approach within a GMM estimation technique?

Q4 How can sample size be calculated for a desired level of power for hypothesis tests using longitudinal data with time-dependent covariates applying a Distant Metric Statistic approach within a GMM estimation technique?

Q5 How well do the proposed power calculation approaches within a GMM method perform compared to the empirical power?

The first four questions are being addressed in this chapter through some proofs I constructed due to the importance of developing the theoretical derivation of the power calculation procedures before implementing the empirical component of this study.

Various methods to properly model longitudinal data have been studied by Fitzmaurice et al. (1993), Gueorguieva (2001), and others and a discussion of these methods was given in the previous chapter. When trying to estimate the statistical power for such data, current research is mainly based on GEE techniques. GEE is appropriate for time-independent covariates but not for time-dependent covariates. The primary methodology of the existing approaches involves the use of the Wald test, the likelihood ratio test, and the score test as proposed by Rochon (1998), Lyles et al. (2007), and Liu and Liang (1997), respectively. However, this dissertation focused on time-dependent covariates. In the presence of such covariates, the models explored before based on GEE are not as efficient as the ones that can be developed based on GMM. Time-dependent

covariates are modeled more efficiently when using GMM according to Lai and Small (2007).

Chapters I and II introduced and expanded on the need to find power estimation for longitudinal data in the presence of time-dependent covariates. The purpose of this chapter is to describe a method that uses GMM instead of GEE to estimate the power and the minimum sample size when testing different hypotheses in longitudinal data. This is necessary to study because longitudinal data that contain time-dependent covariates arise in many research situations, such as health data research, in which covariates do not necessarily remain constant throughout the whole study.

Using Generalized Method of Moments in Longitudinal Studies

When testing hypotheses about parameter vectors, different techniques can be used. For instance, in economics, there are many cases in which a particular theory implies some restrictions on the parameter vectors of the econometric model. Consequently, the accuracy of the theory can be assessed by testing whether such restrictions are met using the data (Hall, 2005). Such tests can be performed in every discipline in which some theory needs to be evaluated using real data. The power can be estimated and the required sample size may be calculated for these hypothesis tests. To do this, in general first the test needs to be defined. Suppose the specific desired hypothesis test can be expressed as

$$\begin{cases} H_0: \boldsymbol{H\beta} = \boldsymbol{h}_0 \\ H_1: \boldsymbol{H\beta} \neq \boldsymbol{h}_0' \end{cases} \tag{3.1}$$

where $\boldsymbol{H}$ is a full rank matrix and $\boldsymbol{h}_0$ is a conformable vector of constant elements. Then, the estimation technique needs to be picked. Within the chosen test, the vector of parameters, $\boldsymbol{\beta}$, can be estimated using different estimation techniques.

One of these techniques is GMM, which provides a computationally convenient framework for making inferences within such models without the need to specify the likelihood function (Hall, 2005). Instead, within GMM, a certain number of moment conditions need to be specified for the model. This will result in a partially specified model, which uses the moment conditions to obtain estimates. As discussed in Chapter II, according to Hansen (2007), GMM estimation begins with a vector of population moment conditions taking the form

$$E[f(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)] = 0, \tag{3.2}$$

where $\boldsymbol{\beta}_0$ is an unknown vector which is to be estimated, $\boldsymbol{x}_{it}$ is a vector of random variables, where $i = 1, \dots, n;\ t = 1, \dots, T$ and $f(.)$ is a vector of functions. The GMM estimator is the value of $\boldsymbol{\beta}$ which minimizes the quadratic form

$$Q(\boldsymbol{\beta}) = \{n^{-1} \sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})\}' \boldsymbol{W} \{n^{-1} \sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})\}, \tag{3.3}$$

where $\boldsymbol{W}$ is a positive semi-definite weighting matrix which may depend on the data but converges in probability to a matrix of constants which is positive definite and $n^{-1} \sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})$ is the sample moment. By definition, the GMM estimator of $\boldsymbol{\beta}_0$ is

$$\widehat{\boldsymbol{\beta}} = arg \min_{\boldsymbol{\beta} \in \mathbb{P}} Q(\boldsymbol{\beta}), \tag{3.4}$$

where $arg\ min$ stands for the value of the argument $\boldsymbol{\beta}$ which minimizes the function in front of it. Hansen (1982), Hansen (2007), and Lai and Small (2007) discussed the GMM theory and their results are used in the current study in detail.

Power Estimation Using Generalized Method of Moments

Considering Equation 3.1, when trying to test a hypothesis about the vector of parameters, the hypothesis can also be written as

$$\begin{cases} H_0: r(\boldsymbol{\beta}) = \mathbf{0} \\ H_1: r(\boldsymbol{\beta}) \neq \mathbf{0}' \end{cases} \tag{3.6}$$

where $r(\boldsymbol{\beta}) = \boldsymbol{H\beta} - \boldsymbol{h}_0$.

In order to test the hypothesis (3.6) using GMM estimators, there exist some statistics, which can be viewed as extensions to the GMM framework of the Wald and distance metric statistic (DM). Unfortunately, some references such as Hall (2005) refer to the DM statistic as a likelihood ratio test; however, this is not accurate as GMM is not a likelihood-based method and the DM statistic is built based on the distance between two quadratic forms within the GMM framework. Thus, "DM statistic" is the preferred name for this statistic in the current dissertation.

To facilitate the presentation of these test statistics, unrestricted and restricted estimators of $\boldsymbol{\beta}$ within GMM need to be defined. The unrestricted estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}$ which is defined above. The restricted estimator of $\boldsymbol{\beta}$, denoted as $\widetilde{\boldsymbol{\beta}}$, is the value of $\boldsymbol{\beta}$ which minimizes $Q(\boldsymbol{\beta})$ subject to $r(\boldsymbol{\beta}) = \mathbf{0}$. It is assumed that both of these minimizations use the same weight matrix $\boldsymbol{W} = \boldsymbol{S}^{-1}$.

The first statistic considered in this dissertation is used within the Wald test that examines whether the unrestricted estimator, $\widehat{\boldsymbol{\beta}}$, satisfies the restrictions with due allowance for sampling error. This statistic can be written as

$$T_W^* = n\left(r(\widehat{\boldsymbol{\beta}})\right)^T \left[R(\widehat{\boldsymbol{\beta}})\left(\boldsymbol{G}_n(\widehat{\boldsymbol{\beta}})^T \boldsymbol{S}^{-1} \boldsymbol{G}_n(\widehat{\boldsymbol{\beta}})\right)^{-1} R(\widehat{\boldsymbol{\beta}})^T\right]^{-1} \left(r(\widehat{\boldsymbol{\beta}})\right), \tag{3.7}$$

where n is the number of subjects, $\widehat{\boldsymbol{\beta}}$ is the unrestricted GMM estimator of the unknown parameters, $\boldsymbol{S}^{-1}$ is the weight matrix, $r(\boldsymbol{\beta}) = \boldsymbol{H\beta} - \boldsymbol{h}_0$,

$$R(\boldsymbol{\beta}) = \frac{\partial r(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'},$$

and

$$\boldsymbol{G}_n(\boldsymbol{\beta}) = E\left[\frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] = n^{-1}\sum_{i=1}^{n}\frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

The second statistic is the DM statistic that examines the impact on the GMM minimand of the imposition of the restrictions. This statistic is

$$T_{DM}^{*} = n\left[Q(\widetilde{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}})\right], \tag{3.8}$$

where within the DM statistic, $Q(.)$ is the quadratic form from Equation 3.3 which needs to be found based on the restricted and unrestricted parameter estimators, respectively, and then to be used in finding the difference between the respective quadratic forms. These statistics, in the context of maximum likelihood (ML) theory, are asymptotically equivalent under the null hypothesis, which can also be extended to the GMM setting (Hall, 2005).

In order to estimate the power of the hypothesis (3.6) using each of the Wald or DM tests, one would need to find the distribution of these test statistics under the null and alternative hypotheses. According to Hall (2005), the limiting distribution of the T_W^* and T_{DM}^* under the null hypothesis is $T_W^* \xrightarrow{d} \chi_s^2$ and $T_{DM}^* \xrightarrow{d} \chi_s^2$ as $n \to \infty$ where s is the rank of $R(\boldsymbol{\beta})$.

Under the alternative hypothesis, both the Wald and DM statistics follow a non-central chi-square distribution, $\chi_s^2(\lambda)$, with the non-centrality parameter λ,

$$\lambda = \boldsymbol{\mu}_R^T[R(\boldsymbol{\beta_0})(\boldsymbol{G}_0^T\boldsymbol{S}^{-1}\boldsymbol{G}_0)^{-1}R(\boldsymbol{\beta_0})^T]^{-1}\boldsymbol{\mu}_R > 0, \quad (3.9)$$

where $\boldsymbol{S}^{-1}$ is the weight matrix, $R(\boldsymbol{\beta_0})$ is the $R(\boldsymbol{\beta})$, defined above, under the null hypothesis, $\boldsymbol{G}_0$ is the $\boldsymbol{G}_n(\boldsymbol{\beta})$ under the null hypothesis and $\boldsymbol{\mu}_R$ is $\sqrt{n}\boldsymbol{\beta_0}$ when $\boldsymbol{h}_0 = \boldsymbol{0}$. $\boldsymbol{\mu}_R$ is equal to $\sqrt{n}(\boldsymbol{H\beta_0} - \boldsymbol{h}_0)$ when $\boldsymbol{h}_0 \neq \boldsymbol{0}$.

The proofs of these distributional assumptions are provided below. The proof regarding the distribution of the Wald statistic is based on one of the linear model theories about the quadratic form's distributions which is mentioned here as Theorem 3.1 (Ravishanker & Dey, 2002). The distributional properties of these statistics have been mentioned in (Hall, 2005), but I constructed the proofs regarding the actual distribution of these statistics.

Theorem 3.1. According to this theorem (Ravishanker & Dey, 2002), if $\boldsymbol{Y}$, a random vector, follows a normal distribution of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a full rank positive definite matrix, $\boldsymbol{A}$ is a symmetric matrix with $rank(\boldsymbol{A}) = m$; then, $\boldsymbol{Y}^T\boldsymbol{AY} \sim \chi^2\left(m, \frac{\boldsymbol{\mu}^T\boldsymbol{A\mu}}{2}\right)$ if any one of the following three conditions are met:

1. $\boldsymbol{A\Sigma}$ is an idempotent matrix of rank m.
2. $\boldsymbol{\Sigma A}$ is an idempotent matrix of rank m.
3. $\boldsymbol{\Sigma}$ is a g-inverse of $\boldsymbol{A}$ with $rank(\boldsymbol{A}) = m$.

This can be applied in finding the distribution of the Wald statistic in Proof 3.1 as the specific case and Proof 3.2 as the general case.

Proof 3.1. Consider the Wald statistic specified in Equation 3.7. It can be written as below

$$T_W^* = n\left(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h}_0\right)^T\left[R\left(\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{G}_n\left(\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{S}^{-1}\boldsymbol{G}_n\left(\widehat{\boldsymbol{\beta}}\right)\right)^{-1}R\left(\widehat{\boldsymbol{\beta}}\right)^T\right]^{-1}\left(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h}_0\right),$$

in which $R(\boldsymbol{\beta}) = \frac{\partial r(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$ will be **H** after taking the derivative of $r(\boldsymbol{\beta})$. For the sake of simplicity, let's write $\boldsymbol{G}_n(\widehat{\boldsymbol{\beta}}) = \boldsymbol{G}$. Now the Equation 3.7 can be written as

$$T_W^* = n\big(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h}_0\big)^T [\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\big(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h}_0\big).$$

Defining $[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1} = \boldsymbol{B}$, T_W^* can be simplified as

$$\begin{aligned} T_W^* &= n\big(\widehat{\boldsymbol{\beta}}^T\boldsymbol{H}^T - \boldsymbol{h}_0^T\big)\boldsymbol{B}\big(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h}_0\big) \\ &= n\big[\widehat{\boldsymbol{\beta}}^T\boldsymbol{H}^T\boldsymbol{B}\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{H}^T\boldsymbol{B}\boldsymbol{h}_0 - \boldsymbol{h}_0^T\boldsymbol{B}\boldsymbol{H}\widehat{\boldsymbol{\beta}} + \boldsymbol{h}_0^T\boldsymbol{B}\boldsymbol{h}_0\big]. \end{aligned} \tag{3.10}$$

Under the common special case that $\boldsymbol{h}_0 = \boldsymbol{0}$ and by substituting $\boldsymbol{B}$, Equation 3.10 can simply be written as

$$\begin{aligned} T_W^* &= n\big[\widehat{\boldsymbol{\beta}}^T\boldsymbol{H}^T\boldsymbol{B}\boldsymbol{H}\widehat{\boldsymbol{\beta}}\big] = n\big[\widehat{\boldsymbol{\beta}}^T\boldsymbol{H}^T[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{H}\widehat{\boldsymbol{\beta}}\big] \\ &= \big(\sqrt{n}\widehat{\boldsymbol{\beta}}^T\big)\boldsymbol{H}^T[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{H}\big(\sqrt{n}\widehat{\boldsymbol{\beta}}\big). \end{aligned} \tag{3.11}$$

Using Theorem 3.1, assume

$$\boldsymbol{A} = \boldsymbol{H}^T[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{H},$$

and

$$\boldsymbol{Y} = \sqrt{n}\widehat{\boldsymbol{\beta}}.$$

Because $\widehat{\boldsymbol{\beta}}$ is asymptotically normal, $\sqrt{n}\widehat{\boldsymbol{\beta}} \dot{\sim} N(\sqrt{n}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}$. It is shown below that $\boldsymbol{A\Sigma}$ is an idempotent matrix meaning that $(\boldsymbol{A\Sigma})(\boldsymbol{A\Sigma}) = \boldsymbol{A\Sigma}$.

Substituting $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$,

$$\begin{aligned} &(\boldsymbol{A\Sigma})(\boldsymbol{A\Sigma}) \\ &= \{\boldsymbol{H}^T[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\}\{\boldsymbol{H}^T[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\}. \end{aligned}$$

Due to the fact that $\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\,\boldsymbol{H}^T[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1} = \boldsymbol{I}$, then

$$\begin{aligned} (\boldsymbol{A\Sigma})(\boldsymbol{A\Sigma}) &= \boldsymbol{H}^T[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{I}\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1} \\ &= \boldsymbol{H}^T[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1} = \boldsymbol{A\Sigma}. \end{aligned}$$

This proves that $\boldsymbol{A\Sigma}$ is idempotent, which is the one condition that needs to be met in order to conclude that the quadratic form Equation 3.11 is distributed as a chi-square. So, substituting $\boldsymbol{Y} = \sqrt{n}\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\mu} = \sqrt{n}\boldsymbol{\beta}$ in $\boldsymbol{Y}^T\boldsymbol{AY} \sim \chi^2\left(m, \frac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}}{2}\right)$, we can say

$$\boldsymbol{Y}^T\boldsymbol{AY} = \left(\sqrt{n}\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{H}^T[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{H}\sqrt{n}\widehat{\boldsymbol{\beta}} = T_W^* \sim \chi^2(s, \lambda),$$

where the non-centrality parameter under the null hypothesis is defined as

$$\lambda = \frac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}}{2} = \frac{1}{2}\left(\sqrt{n}\boldsymbol{\beta_0}\right)^T\boldsymbol{H}^T[\boldsymbol{H}(\boldsymbol{G}_0^T\boldsymbol{S}^{-1}\boldsymbol{G_0})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{H}\left(\sqrt{n}\boldsymbol{\beta_0}\right)$$

$$= \frac{1}{2}\boldsymbol{\mu}_R^T[R(\boldsymbol{\beta}_0)(\boldsymbol{G}_0^T\boldsymbol{S}^{-1}\boldsymbol{G_0})^{-1}R(\boldsymbol{\beta}_0)^T]^{-1}\boldsymbol{\mu}_R,$$

defining $\boldsymbol{\mu}_R = \sqrt{n}\boldsymbol{\beta_0}$. □

This proof was for the common case where $\boldsymbol{h}_0 = \boldsymbol{0}$ which would result in a special case where $r(\boldsymbol{\beta})$ is reduced to $\boldsymbol{H\beta}$; however, it was of interest to also find the distribution of the Wald statistic for the general case where $\boldsymbol{h}_0 \neq \boldsymbol{0}$ to be able to generalize these results. In that case $r(\boldsymbol{\beta}) = \boldsymbol{H\beta} - \boldsymbol{h}_0$ and the distribution of the Wald statistic will be chi-square as proven below.

Proof 3.2. Knowing that $r\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h}_0 \sim N(\boldsymbol{H\beta} - \boldsymbol{h}_0, \boldsymbol{H\Sigma H}^T)$, consider the Wald statistic

$$T_W^* = n\left(r\left(\widehat{\boldsymbol{\beta}}\right)\right)^T\left[R\left(\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{G}_n\left(\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{S}^{-1}\boldsymbol{G}_n\left(\widehat{\boldsymbol{\beta}}\right)\right)^{-1}R\left(\widehat{\boldsymbol{\beta}}\right)^T\right]^{-1}\left(r\left(\widehat{\boldsymbol{\beta}}\right)\right)$$

$$= n\left(r\left(\widehat{\boldsymbol{\beta}}\right)\right)^T[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\left(r\left(\widehat{\boldsymbol{\beta}}\right)\right)$$

$$= \left(\sqrt{n}r\left(\widehat{\boldsymbol{\beta}}\right)^T\right)[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\left(\sqrt{n}r\left(\widehat{\boldsymbol{\beta}}\right)\right).$$

This time $\boldsymbol{A}$ and $\boldsymbol{Y}$ are different from before; calling them $\boldsymbol{A}^*$ and $\boldsymbol{Y}^*$, which are

$$\boldsymbol{A}^* = [\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1},$$

and

$$\boldsymbol{Y}^* = \sqrt{n}r(\widehat{\boldsymbol{\beta}}),$$

which still follows a normal distribution but with a different mean and variance. To find the mean and variance of $\boldsymbol{Y}^*$, consider

$$\sqrt{n}\widehat{\boldsymbol{\beta}} \dot{\sim} N(\sqrt{n}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \rightarrow \sqrt{n}\boldsymbol{H}\widehat{\boldsymbol{\beta}} \dot{\sim} N(\sqrt{n}\boldsymbol{H}\boldsymbol{\beta}, \boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^{\boldsymbol{T}})$$

$$\rightarrow \sqrt{n}(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h}_0) \dot{\sim} N(\sqrt{n}(\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{h}_0), \boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^{\boldsymbol{T}}),$$

so, $\boldsymbol{Y}^* \sim N(\sqrt{n}(\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{h}_0), \boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^{\boldsymbol{T}})$.

Knowing that the new variance covariance matrix of $\boldsymbol{Y}^*$ is

$$\boldsymbol{\Sigma}^* = \boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^{\boldsymbol{T}} = \boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^{\boldsymbol{T}},$$

it can be shown that $\boldsymbol{A}^*\boldsymbol{\Sigma}^*$ is an identity matrix, hence an idempotent one,

$$(\boldsymbol{A}^*\boldsymbol{\Sigma}^*)(\boldsymbol{A}^*\boldsymbol{\Sigma}^*) =$$

$$= \{[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T\}\{[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T\}$$

$$= I^2 = I = \boldsymbol{A}^*\boldsymbol{\Sigma}^*.$$

Therefore,

$$\boldsymbol{Y}^{*T}\boldsymbol{A}^*\boldsymbol{Y}^* = \left(\sqrt{n}r(\widehat{\boldsymbol{\beta}})^T\right)[\boldsymbol{H}(\boldsymbol{G}^T\boldsymbol{S}^{-1}\boldsymbol{G})^{-1}\boldsymbol{H}^T]^{-1}\left(\sqrt{n}r(\widehat{\boldsymbol{\beta}})\right) \sim \chi^2(s, \lambda),$$

where the non-centrality parameter is as below under the null hypothesis

$$\lambda^* = \frac{\boldsymbol{\mu}^{*T}\boldsymbol{A}^*\boldsymbol{\mu}^*}{2} = \frac{1}{2}\left(\sqrt{n}(\boldsymbol{H}\boldsymbol{\beta}_{\boldsymbol{0}} - \boldsymbol{h}_0)\right)^T[\boldsymbol{H}(\boldsymbol{G}_0^T\boldsymbol{S}^{-1}\boldsymbol{G}_{\boldsymbol{0}})^{-1}\boldsymbol{H}^T]^{-1}\left(\sqrt{n}(\boldsymbol{H}\boldsymbol{\beta}_{\boldsymbol{0}} - \boldsymbol{h}_0)\right)$$

$$= \frac{1}{2}\boldsymbol{\mu}^{*T}_R[R(\boldsymbol{\beta}_0)(\boldsymbol{G}_0^T\boldsymbol{S}^{-1}\boldsymbol{G}_{\boldsymbol{0}})^{-1}R(\boldsymbol{\beta}_0)^T]^{-1}\boldsymbol{\mu}^*_R,$$

defining $\boldsymbol{\mu}^*_R = \sqrt{n}(\boldsymbol{H}\boldsymbol{\beta}_{\boldsymbol{0}} - \boldsymbol{h}_0)$. □

The DM statistic follows the same distribution as the Wald statistic with the same non-centrality parameter because these statistics are identical. The proof for their identity can be found in Hall (2005) and Newey and McFadden (1994).

Power Estimation Steps Using Generalized Method of Moments

In every power estimation procedure, the distribution of the statistic used for testing the hypothesis needs to be known under the null and alternative hypotheses. Both Wald and DM statistics are distributed as a central chi-square distribution under the null hypothesis and a non-central chi-square distribution with the non-centrality parameter given in Equation 3.9 under the alternative hypothesis. Knowing all this information, the statistical power can be estimated using the following steps.

Considering the repeated measures used before at T time points for n subjects, in order to find the power and then the required sample size of a statistical test, first the hypothesis needs to be specified and tested as shown in (3.6)

$$\begin{cases} H_0: r(\boldsymbol{\beta}) = \mathbf{0} \\ H_1: r(\boldsymbol{\beta}) \neq \mathbf{0} \end{cases}.$$

Then the statistic, which is used to test this hypothesis, needs to be specified. Because the GMM approach is being adopted for this study, the Wald Equation 3.7 and the DM statistic Equation 3.8, where

$$\text{Wald: } T_W^* = n\left(r(\widehat{\boldsymbol{\beta}})\right)^T \left[R(\widehat{\boldsymbol{\beta}})\left(\boldsymbol{G}_n(\widehat{\boldsymbol{\beta}})^T \boldsymbol{S}^{-1} \boldsymbol{G}_n(\widehat{\boldsymbol{\beta}})\right)^{-1} R(\widehat{\boldsymbol{\beta}})^T\right]^{-1} \left(r(\widehat{\boldsymbol{\beta}})\right),$$

and

$$\text{DM: } T_{DM}^* = n\left[Q(\widetilde{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}})\right].$$

Then distribution of these statistics under the null and alternative hypothesis needs to be specified. Their asymptotic distribution under the null hypothesis is equivalent as below

$$H_0: T_W^* \xrightarrow{d} \chi^2_{(s)}, \qquad T_{DM}^* \xrightarrow{d} \chi^2_{(s)}.$$

Newey and West (1987) showed that the asymptotic equivalence of the statistics extends to the alternative hypothesis. As discussed above, under the alternative hypothesis, the Wald and DM statistics have an approximate non-central chi-square distribution of $\chi^2_{(s),\lambda}$ with the Equation 3.9 non-centrality parameter

$$\lambda = \frac{1}{2}\boldsymbol{\mu}_R^T[R(\boldsymbol{\beta}_0)(\boldsymbol{G}_0^T\boldsymbol{S}^{-1}\boldsymbol{G_0})^{-1}R(\boldsymbol{\beta}_0)^T]^{-1}\boldsymbol{\mu}_R.$$

In order to estimate the power, assuming that α represents the type I error, $\chi^2_{(s);1-\alpha}$ is the critical value from the central $\chi^2_{(s)}$ distribution. Using this critical value, power can be calculated by finding the probability of

$$\Pr\left(\chi^2_{s,(\lambda)} \geq \chi^2_{s,1-\alpha}\right), \tag{3.12}$$

with $\chi^2_{s,1-\alpha}$ denoting the $100(1-\alpha)$th percentile of the central chi-square with s degrees of freedom. So, the power associated with the Wald and DM test statistics is

$$1-\gamma = \int_{\chi^2_{(s);1-\alpha}}^{\infty} f(x_t; s, \lambda)dx, \tag{3.28}$$

where γ represents the type II error and $f(x_t; s, \lambda)$ is the probability density function of $\chi^2_{(s),\lambda}$.

Different steps to estimate the statistical power of longitudinal data using two aforementioned Wald and Distant Metric statistics can be summarized in Table 3.1.

Table 3.1

Statistical Power Estimation Steps

	Test Statistics	
	Wald	Distance Metric
Test Statistic	$n\left(r(\widehat{\boldsymbol{\beta}})\right)^T\left[R(\widehat{\boldsymbol{\beta}})\left(\boldsymbol{G}_n(\widehat{\boldsymbol{\beta}})^T S^{-1}\boldsymbol{G}_n(\widehat{\boldsymbol{\beta}})\right)^{-1}R(\widehat{\boldsymbol{\beta}})^T\right]^{-1}\left(r(\widehat{\boldsymbol{\beta}})\right)$	$n[Q(\widetilde{\boldsymbol{\beta}})-Q(\widehat{\boldsymbol{\beta}})]$
Step 1	Calculate the non-centrality parameter	Calculate the non-centrality parameter
Step 2	Find the critical value	Find the critical value
Step 3	$\Pr(\chi^2_{s,(\lambda)} \geq \chi^2_{s,1-\alpha})$ $\int_{\chi^2_{(s);1-\alpha}}^{\infty} f(x_t; s, \lambda)dx$	$\Pr(\chi^2_{s,(\lambda)} \geq \chi^2_{s,1-\alpha})$ $\int_{\chi^2_{(s);1-\alpha}}^{\infty} f(x_t; s, \lambda)dx$

Model Evaluation

The first four research questions were answered theoretically in this chapter by providing the proofs I constructed. To check the performance of the proposed theoretical GMM-based methods for estimating power and calculating the required sample sizes, a real data analysis and a simulation study were conducted. The real data set was used as an exemplar data set. The fifth question regarding the comparison of the exact power using the proposed GMM-based power calculation approaches to the empirical power was addressed using the simulated data. I constructed the R functions to accomplish these power and sample size estimates in Chapter IV of this dissertation. I developed a practical technique for estimating the theoretical powers using GMM in Chapter IV, which can be adopted by applied researchers and practitioners.

Evaluation of the performance of the proposed methods using GMM was carried out primarily via comparisons between the proposed methods and the empirical power

from the simulation study. The proposed methods were used to estimate the exact power for the pilot data, the post-hoc power of the simulated data sets, and build their bootstrap confidence intervals. Then, the hypothesis tests were performed on the simulated data sets over and over to compute the empirical power. The comparison of the empirical power and the estimated power was proposed as an appropriate method to evaluate the performance of the proposed GMM-based methods. Two sets of comparisons were made in Chapter IV; first, the comparison of the exact estimated theoretical powers of the pilot data and the post-hoc powers of the simulated data to see how well the estimated theoretical powers lined up with the post-hoc powers of different sizes of simulated data. Second, the comparison of the exact theoretical powers of the pilot data and the empirical powers, which come from the rejection rates while performing the hypothesis tests on the simulated data sets.

This simulation was not intended to compare the proposed GMM-based power estimation methods and the previously studied methods based on GEE. It rather was for comparing the exact power calculation to the empirical results of performing the hypothesis test on the simulated data sets multiple times to check the adequacy of the estimated power.

None of these data sets had been analyzed previously under the current methodological frame.

Example Data Set: Osteoarthritis Initiative

This study involved the use of the proposed power estimation techniques on one real data set to evaluate the performance of the proposed models. Using this pilot data, practical power estimation methods were developed, which can be adopted by researchers

in different fields. This data set was also used as an exemplary data set for future calculations. This data set contains characteristics of interest such as longitudinal data and different types of covariates including time-dependent covariates. These covariates were expected to provide a valid application of methodology to assess the efficiency of the proposed power estimation techniques and a comparison of them to the previously studied models.

The dataset used consists of data from the osteoarthritis initiative (OAI) which can be found at www.oai.ucsf.edu. The OAI data consist of a multi-center study on knee osteoarthritis in more than 4,000 subjects over a period of nine or more years. For the sake of simplicity, data from up to the 5^{th} follow-up year were considered. If the number of complete cases was large enough, it dropped to three follow-ups. Where there exist problems with convergence in the process of using the proposed models on the OAI dataset, the covariates were adjusted to overcome the potential issue.

Many variables were gathered; however, this research focused on modeling Western Ontario and McMaster Universities' (WOMAC) disability score, which is typically treated as a continuous variable. This dataset contains longitudinal data by the fact that observations were gathered on the same subjects over time and are thus more related to each other than observations from other subjects. The subjects' age and BMI at each time point as well as the subjects' gender were utilized as fixed effect regression predictors in the model. Age and BMI can be two of the time-dependent covariates in this study, which do not remain constant over time.

The proposed power estimation techniques were applied to this dataset to check how the power and required sample sizes can be estimated using Wald and DM statistics.

It was of interest to test the BMI, which is a time-dependent covariate within the model mentioned below

$$WOMAC_{it} = \beta_0 + \beta_1 Age_{it} + \beta_2 BMI_{it} + \beta_3 Sex_{it} + t_2 + t_3 + \varepsilon_{it},$$

where age is a type I time-dependent covariate, t_2 is a type I time-dependent covariate and a time indicator of the second follow-up time, t_3 is a type I time-dependent covariate and a time indicator of the third follow-up time and sex is a time-independent covariate

Within this model, the following hypothesis was tested

$$\begin{cases} H_0: \beta_2 = \mathbf{0} \\ H_1: \beta_2 \neq \mathbf{0} \end{cases}.$$

For this hypothesis, the power for different sample sizes was estimated using the Wald and DM statistics within the GMM-based power estimation method. The R functions I developed were used to perform each estimation.

Simulation Study

Simulated data were also used for evaluating the performance of the two power calculation techniques proposed in this dissertation. The data were simulated using Monte Carlo simulation in R version 3.2.2 (R Core Team, 2015).

This simulation was based on the real dataset introduced above to ensure that the simulated data are representative of the values seen in reality. Using the real dataset, predictor values and effect coefficients directly came from the OAI dataset and continuous response values were simulated based on them. Having predictors and effect coefficients coming from the real data helped get the time-dependent covariates in the simulated data to behave as they would in a real situation. Including these time-dependent covariates in the simulated data also helped to check the performance of the two GMM-based power estimation methods proposed in this study in the presence of such

covariates. Within this simulation, the same model used for the real data analysis was used and the hypothesis test for the BMI was performed. Within the simulation study, rejection rates and post-hoc powers of each simulated data were recorded for comparison with the estimated theoretical powers of the pilot data.

The nature of this simulation study is different from other simulation studies as this study was used for evaluating the effectiveness of the statistical power calculation methods proposed rather than checking the appropriateness or efficiency of different estimation techniques or statistical models, which are common in simulation studies. This study focused on developing two power estimation techniques for longitudinal data in the presence of time-dependent covariates; not on developing a new coefficient estimation technique. So, in the current simulation study, values such as standard errors were not used to compare different techniques. Instead, at the end of this study, the estimated power for different sample sizes using GMM-based power estimation methods was calculated for 3,600 data sets within different sample sizes. Then the 95% bootstrap confidence interval for each set of the estimated post-hoc powers was calculated. Finally, the actual hypothesis tests within the simulated data sets were performed using a Wald test and a DM test and the empirical power based on the rejection rates was calculated. After calculating the empirical rejection rates for the simulated data sets using Wald and DM tests, whether or not the empirical power for each method fell into the respective calculated 95% confidence intervals of the estimated power was reported as well as how close those values are to the theoretical powers. Having the empirical powers close to the estimated theoretical powers and the theoretical powers falling into the calculated confidence intervals of the estimated post-hoc powers are justifications that the proposed

power calculation methods are performing well. Tables are provided in Chapter IV of this dissertation to summarize the results of this simulation study for different sample sizes and statistical power estimation techniques.

Within this simulation study, I tried to exemplify the proofs which I showed to work for large sample sizes and see if they work for smaller sample sizes as well. Different sample sizes, which were used for this simulation study, include 25, 50, 100, and 200 subjects with three observations per subject. Sample sizes of 100 and 200 were chosen according to the simulation study by Lyles et al. (2007) which focused on a GEE-based technique for power estimation of longitudinal data using the Wald test. Two smaller sample sizes of 25 and 50 were also added to this study to compare the accuracy of the estimated statistical power for the smaller sample sizes to the higher sample size of 100 and 200. This comparison was of interest to see whether the methods that were shown to work for large sample sizes according to the proofs Newey and West (1987) and I constructed work as well in terms of the accuracy of the estimated power for the small sample sizes or not.

Three thousand and six hundred replicated samples were generated for each sample size within this simulation study. This number was calculated based on the theory from Robert and Casella (2013) explained below. Trying to use the existing literature to decide the number of replications resulted in two values based on two power estimation simulation studies of longitudinal data. Lyles et al. (2007) used 2,000 randomly generated data sets and Liu and Liang (1997) used 5,000 replications. Neither of these simulation sizes was selected for the current study because of the differences that existed between the nature of their simulation studies and the simulation study used for this dissertation.

Therefore, another method needed to be considered for determining the number of replications. The method adopted for this study was based on finding the required simulation size to achieve a desired level of accuracy of the recorded results (Robert & Casella, 2013). Each rejection of the hypothesis tests, which is a binary variable resulting in having a binomial distribution for the recorded results, was recorded. What the binomial random variable provides is an upper bound for the variance needed to calculate the number of replications as shown in Equation 3.29

$$M = \frac{[SD]^2}{d^2}, \tag{3.29}$$

where M is the number of replications, d is the level of accuracy, and $[SD]^2$ is the variance of the simulation outcome which comes from the sampling distribution of the recorded statistics. Given that I wished to report the p-values from the hypothesis tests with two digits of accuracy in order to decide the rejection of the null hypothesis, I needed the standard error to be half of the distance between two consecutive reported p-values with two digits of accuracy. Therefore, $d = {\cdot 01}/{2} = .005$. Using the variance of the binomial distribution and the desired power of .9, Equation 3.29 resulted in the minimum simulation size of 3,600, which is the required sample size for the empirical power. When reporting the estimated power, the beta distribution can be used to find the variance used in Equation 3.29. This is because power follows a beta distribution, $beta\ (a, b)$, with the shape parameters a and b where $b = 1$ according to Gupta and Nadarajah (2004). When $a = 1$, the resulting distribution will be the power function distribution which is a special case of the beta distribution (Gupta & Nadarajah, 2004). Considering the variance of this beta distribution, the number of replications were 3,300 which is the required sample size for the exact power calculation process. Three thousand

and three hundred (3,300) is smaller than 3,600; therefore, the number of replicates I decided to use for this study was 3,600 to ensure having a good-enough replication and an acceptable precision. I developed the R codes for performing the above power estimation procedures and shared them with the public at the end of this study.

The final tables, which are provided in the results of this study, helped in making conclusions regarding the performance of the proposed power estimation techniques for different sample sizes. These tables (similar to Table 3.2) summarize the power calculation results for sample sizes of 25, 50, 100, and 200 subjects.

Information in tables similar to Table 3.2 will be used to make the final conclusion about the performance of the proposed power estimation technique using the Wald statistic by comparing the empirical power applying the Wald test to the 95% confidence interval of the estimated power using the Wald method. I made the final conclusion about the performance of the proposed power estimation technique using the DM statistic by comparing the empirical power which applies the DM test to the 95% confidence interval of the estimated power adopting the DM method. Finally, the theoretical powers of each sample size were compared to the 95% bootstrap confidence interval of the estimated power for each simulated data (post-hoc power).

Table 3.2

Simulation Results for Each Sample Size

	Estimated Power	Hypothesis Test Wald Test (Reject or Not)	Hypothesis Test DM Test (Reject or Not)
Estimated Values	Summary of the post-hoc powers	Empirical power using Wald test for each sample size	Empirical power using DM test for each sample size
Confidence Intervals	95% bootstrap confidence interval	95% confidence interval of the rejection rate	95% confidence interval of the rejection rate

At the end, these results were compared across four sample sizes and the empirical power and the estimated power are closer to each other and higher in value when sample sizes are larger.

CHAPTER IV

RESULTS

This chapter describes the simulation procedure used in the process of estimating statistical power and required sample size when working with longitudinal data in the presence of time-dependent covariates using generalized method of moments (GMM). The main purpose of this simulation was to compare the exact power calculation, based on the methods developed in Chapter III of this dissertation, to the empirical results of performing the hypothesis test on the simulated data sets. In addition, providing the comparison of the post-hoc powers of the simulated data sets was of interest to evaluate the performance of the developed theory on smaller sample sizes.

Introduction

In this study, I aimed to develop power estimation and sample size calculation techniques for longitudinal data with time-dependent covariates using GMM. The reason for using GMM within the power estimation techniques instead of previously developed methods which were based on generalized estimating equations is the higher efficiency of GMM compared to GEE when dealing with time-dependent covariates (Lai & Small, 2007). However, when GMM is adopted as an estimation technique within a longitudinal model, prior to the current study, there was no existing methodology to estimate the power of such models.

The simulation study was carried out in R version (3.3.2); I wrote all the programs, including the estimation algorithm, for this study. To evaluate and compare the aforementioned approaches accurately, this study followed the simulation scheme based on real data to assure the simulated data show the same behavior as the real longitudinal type of data with time-dependent covariates. This method of simulation provides a comparable replication of the data analyses in real life scenarios. The simulation results are reported in text and presented in tables and figures relative to each of the research questions mentioned in Chapter I.

The remainder of this chapter is divided into the following sections. The first section briefly discusses the research questions and the answers to those questions. The second section describes the steps in the simulation study and the steps that needed to be taken in the process of transforming and generating the outcome variable, controlling the effect size, and simulating the final data sets for different conditions. In the third section, I discuss the algorithm for GMM estimation, which was used in the process of power estimation in the next sections. Section four presents problems with the convergence of the GMM algorithm and the solution to resolve this issue. Section five contains the issues I faced in the process of completing this study in terms of the run time and the steps which were taken to make the large simulation possible in a reasonable amount of time. Section six includes some issues associated with the distant metric (DM) statistic and the reasons causing such problems. The seventh section describes the GMM power estimation procedure and how tied it is with the number and magnitude of the responses, effect size, parameter estimates, and the sample sizes used within the theoretical power calculation. Section eight contains the simulation study results comparing the behavior of

the techniques proposed in Chapter III to the empirical results and post-hoc powers under different conditions to address the concern with smaller sample sizes. Lastly, the ninth section is dedicated to the summary and implications for the power estimation of longitudinal data with time-dependent covariates using GMM

Research Questions and Their Answers

Two statistics including Wald and DM statistics, which are used for testing statistical hypotheses when GMM is used, were discussed in Chapter III. The Wald statistic is,

$$T_W^* = n\left(r(\widehat{\boldsymbol{\beta}})\right)^T \left[R(\widehat{\boldsymbol{\beta}})\left(\boldsymbol{G}_n(\widehat{\boldsymbol{\beta}})^T \boldsymbol{S}^{-1} \boldsymbol{G}_n(\widehat{\boldsymbol{\beta}})\right)^{-1} R(\widehat{\boldsymbol{\beta}})^T\right]^{-1} \left(r(\widehat{\boldsymbol{\beta}})\right), \tag{4.1}$$

where n is the number of subjects, $\widehat{\boldsymbol{\beta}}$ is the unrestricted GMM estimator of the unknown parameters, $\boldsymbol{S}^{-1}$ is the weight matrix, $r(\boldsymbol{\beta}) = \boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{h}_0$,

$$R(\boldsymbol{\beta}) = \frac{\partial r(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}, \tag{4.2}$$

and

$$\boldsymbol{G}_n(\boldsymbol{\beta}) = E\left[\frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] = n^{-1}\sum_{i=1}^{n} \frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \tag{4.3}$$

where $f(\boldsymbol{x}_{it}, \boldsymbol{\beta})$ specifies the moment conditions.

The DM statistic is,

$$T_{DM}^* = n\left[Q(\widetilde{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}})\right], \tag{4.4}$$

where $Q(.)$ is the quadratic form from the GMM algorithm which needs to be found based on the restricted and unrestricted parameter estimators $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$, respectively, and then to be used in finding the difference between the two quadratic forms.

In order to use these statistics within the power estimation and sample size calculation process, their distributions under the null and alternative hypotheses needed to be identified. It has been proven that their asymptotic distributions are central and non-central Chi Square under null and alternative hypotheses, respectively. I provided a new proof for the Wald statistic's distributions in Chapter III of this dissertation. According to Hall (2005), Wald and DM statistics' distributions under the null and alternative hypotheses are identical; this theory was used in answering the research questions regarding using the DM statistic in the process of power estimation. Finally, different steps that needed to be taken to estimate statistical power and calculate optimal sample size are discussed. The first four questions were addressed in the previous chapter through some methodology and proofs I constructed due to the importance of developing the theoretical derivation of the power calculation procedures before implementing the empirical component of this study. To summarize, the first four research questions given in Chapter I were addressed in Chapter III; they are also briefly discussed below in multiple steps to be adopted by researchers and applied practitioners. More details about each of the answers to the research questions are discussed later in this chapter.

Suppose the repeated measures for a study are recorded at T time points for n subjects and researchers will test the hypothesis

$$\begin{cases} H_0: r(\boldsymbol{\beta}) = \mathbf{0} \\ H_1: r(\boldsymbol{\beta}) \neq \mathbf{0}' \end{cases}$$

where $r(\boldsymbol{\beta}) = \boldsymbol{H\beta} - \boldsymbol{h}_0$.

Q1 How can power be calculated for hypothesis tests using longitudinal data with time-dependent covariates applying a Wald approach within a GMM estimation technique?

Answer to Question 1. In order to estimate power for longitudinal data in the presence of time-dependent covariates using the Wald approach, adopting a GMM estimation technique, three steps need to be taken which are summarized below:

Step 1. It was proven in Chapter III that the Wald statistic has a central chi-square distribution under the null hypothesis. As Hall (2005) discussed, $T_W^* \xrightarrow{d} \chi_s^2$ as $n \to \infty$ where s is the rank of $R(\boldsymbol{\beta})$. Therefore, one needs to find the degrees of freedom of this chi-square distribution and find the chi-square critical value for the degrees of freedom which depends on the number of parameters that are tested in the null hypothesis.

Step 2. It also was proven in Chapter III that the Wald statistic under the alternative hypothesis has a non-centrality parameter, which needs to be calculated before moving to the next step. Under the alternative hypothesis, the Wald statistic follows a non-central chi-square distribution, $\chi_s^2(\lambda)$, with the non-centrality parameter λ, which can be calculated as below

$$\lambda = \boldsymbol{\mu}_R^T[R(\boldsymbol{\beta_0})(\boldsymbol{G}_0^T\boldsymbol{S}^{-1}\boldsymbol{G}_0)^{-1}R(\boldsymbol{\beta_0})^T]^{-1}\boldsymbol{\mu}_R > 0, \tag{4.5}$$

where $\boldsymbol{S}^{-1}$ is the weight matrix or $\boldsymbol{W}$, $R(\boldsymbol{\beta_0})$ is the $R(\boldsymbol{\beta})$, defined above, under the null hypothesis, $\boldsymbol{G}_0$ is the $\boldsymbol{G}_n(\boldsymbol{\beta})$ under the null hypothesis and $\boldsymbol{\mu}_R$ is $\sqrt{n}\boldsymbol{\beta_0}$ when $\boldsymbol{h}_0 = \boldsymbol{0}$. $\boldsymbol{\mu}_R$ is equal to $\sqrt{n}(\boldsymbol{H\beta_0} - \boldsymbol{h}_0)$ when $\boldsymbol{h}_0 \neq \boldsymbol{0}$.

Step 3. Then, the power can be calculated by integrating the probability distribution function of the non-central chi-square with the non-centrality parameter found in step 2. This integration starts from the central chi-square critical value found in step 1 and goes to infinity. This gives the power for a data set with a known sample size.

$$\Pr\left(\chi_{s,(\lambda)}^2 \geq \chi_{s,1-\alpha}^2\right) = \int_{\chi_{(s);1-\alpha}^2}^{\infty} f(x_t; s, \lambda)dx. \tag{4.6}$$

Q2 How can sample size be calculated for a desired level of power for hypothesis tests using longitudinal data with time-dependent covariates applying a Wald approach within a GMM estimation technique?

Answer to Question 2. In order to answer this question, power should be calculated for different sample sizes using multiple steps depending on the size of the pilot data.

Scenario 1 is when the size of the pilot data is larger than the sizes of the data sets considered for the future studies. In that case, multiple subsamples of the pilot data set must be taken for each of the sample sizes considered as possible options for future studies. Then, within each set of sample sizes, the non-centrality parameters need to be calculated for each sub-sample of each size. These non-centrality parameters of each sample size need to be averaged at the end and the power needs to be calculated for the averaged non-centrality parameter.

The reason for averaging the non-centrality parameters first and then finding the theoretical power for them rather than finding the power multiple times for each sub-sample and then averaging them, which is what I originally implemented, is the sensitivity of the power to the non-centrality parameter of each sub-sample and higher variance of power than the real power value for each sample size, which results in skewing the final averaged power. For example, when the non-centrality parameter of one of the sub-samples gets small, the resulting power of that sub-sample gets extremely small; using this extremely small power and averaging it along with the other powers will skew the mean of the powers at the end. But once all the non-centrality parameters of the representative sub-samples are averaged and then one theoretical power for the mean of

the non-centrality parameters is calculated, the power is representative of the actual power (close to the post-hoc power of 3,600 simulated datasets). The mathematical theory behind the relationship between the magnitude of the non-centrality parameters and how they affect the integration process within the power calculation procedure is discussed later.

Scenario 2 is when the size of the pilot data is smaller than the sizes of data sets considered for future studies. In that case, multiple data sets of the desired sizes need to be simulated using the characteristics of the data. This simulation process to expand the pilot dataset can be performed following the steps from Lyles et al. (2007). After this step is completed, the same steps as described above should be repeated for each of the simulated data sets within each sample size to calculate the power for desired sizes of sample. To simplify these steps, the procedure mentioned above is summarized in six steps as below:

Step 1. Determine the appropriate model and the hypothesis to be tested.

Step 2. Determine the "true" effects of the alternative.

Step 3. Determine the sample sizes of interest.

Step 4a. If the pilot data are larger than the sample sizes of interest, sub-samples of covariates and their respective responses should be selected. If effect sizes for the study were chosen to differ from the original model, the new outcomes must be generated.

Step 4b. If the pilot data are smaller than the sample sizes of interest, data sets of the sizes of interest should be randomly generated. The article by Lyles et al. (2007) can clarify the steps of generating data that are representative of the original pilot data.

Step 5. Use software to obtain the non-centrality parameters for all sub-samples / simulated samples. The programs I wrote in R 3.2.2 can be used to find the non-centrality parameters.

Step 6. Use the average of all non-centrality parameters to calculate the final power for each sample size.

Q3 How can power be calculated for hypothesis tests using longitudinal data with time-dependent covariates applying a Distant Metric Statistic approach within a GMM estimation technique?

Answer to Question 3. In order to estimate power for longitudinal data in the presence of time-dependent covariates using the DM approach within a GMM estimation technique three steps need to be taken which are summarized below:

Step 1. According to Hall (2005), the limiting distribution of the T_{DM}^{*} under the null hypothesis is $T_{DM}^{*} \xrightarrow{d} \chi_s^2$ as $n \to \infty$ where s is the rank of $R(\boldsymbol{\beta})$. So, one needs to find the degrees of freedom of this chi-square distribution and find the chi-square critical value for the degrees of freedom which depends on the null hypothesis being tested.

Step 2. According to Hall (2005), the DM statistic under the alternative hypothesis has a non-centrality parameter that needs to be calculated before moving to the next step. Under the alternative hypothesis, the DM statistic follows a non-central chi-square distribution, $\chi_s^2(\lambda)$, with the non-centrality parameter λ, which can be calculated as using Equation 4.5.

Step 3. Then, the power can be calculated by integrating the probability distribution function of the non-central chi-square with the non-centrality parameter found in step 2. This integration starts from the central chi-square value found in step 1 and goes to infinity. This gives the power for a data set with a known sample size.

$$\Pr\left(\chi^2_{s,(\lambda)} \geq \chi^2_{s,1-\alpha}\right) = \int_{\chi^2_{(s);1-\alpha}}^{\infty} f(x_t; s, \lambda)dx. \tag{4.6}$$

Q4 How can sample size be calculated for a desired level of power for hypothesis tests using longitudinal data with time-dependent covariates applying a Distant Metric Statistic approach within a GMM estimation technique?

Answer to Question 4. The answer to this question is identical to the second question due to the fact that both Wald and DM statistics have the same asymptotic distribution according to Hall (2005).

Once the first four research questions were successfully answered through some theoretical proofs in Chapter III, it was time to complete the empirical aspect of this study to evaluate how well the theoretically developed power estimation methods work for smaller sample sizes. The fifth question, which was not answered in Chapter III, is answered in this chapter using real and simulated data. This question is as below:

Q5 How well do the proposed power calculation approaches within a GMM method perform compared to the empirical power?

This comparison was made multiple times using simulated data to check the adequacy of the estimated power using the GMM-based Wald test as well as the DM statistic. As emphasized in Chapter III, this simulation was not intended to be a comparison of the proposed GMM-based power estimation methods and the previously studied methods based on GEE. Instead, it was designed to compare the exact power calculation to the empirical results of performing the hypothesis test on the simulated data sets multiple times to check the adequacy of the estimated power for smaller sample sizes.

To check the performance of the proposed theoretical GMM-based methods for estimating power and calculating the required sample sizes, a real data analysis and a

simulation study was performed and is reported in this chapter. The real data set was used as an exemplar data set as well as a pilot data set in the process of simulating data for the simulation study. This simulated data set, which is explained in detail below, was used to answer the fifth question regarding the comparison of the exact power using the proposed GMM-based power calculation approaches to the empirical rejection rates and post-hoc powers. I constructed the R functions to accomplish these power and sample size estimates which are shown in the Appendix C.

Evaluation of the performance of the proposed methods using GMM was carried out primarily via comparisons between the proposed methods and the empirical power. The proposed methods were used to estimate the exact power of the pilot data and the post-hoc powers of the simulated data and build their bootstrap confidence intervals. Then, the hypothesis test was performed on the simulated data sets over and over to compute the empirical power. The comparison of the empirical power, post-hoc powers, and the estimated theoretical powers was proposed as an appropriate method to evaluate the performance of the proposed GMM-based methods.

Simulation Study

The algorithm for the simulation study consisted of randomly extracting unique time-dependent and time-independent covariates from the real data set consisting of osteoarthritis initiative (OAI) data, discussed in Chapter III, based on different sample size conditions and effect sizes, then, generating the longitudinal response variables. The simulated data at the end were used to evaluate the performance of the proposed methods using GMM and the empirical power. The steps are discussed in detail in this section.

Exemplar Data Set

Real data were used as an exemplar data set for simulating multiple data sets of sizes 25, 50, 100, and 200 each 3,600 times. The use of real data in the process of data simulation is to ensure the consistency of the simulated data with the outcome variable from the OAI data in the presence of multiple types of time-dependent covariates. This data set contains characteristics of interest such as longitudinal data and different types of covariates including time-dependent covariates. These covariates were expected to provide a valid application of methodology to assess the efficiency of the proposed power estimation techniques.

The dataset, which was used in this study and explained in detail in Chapter III, consisted of OAI data. The number of complete cases used for this study was 2,456. Each subject had three follow-up measurements, which resulted in 7,368 records in the pilot data set.

Many variables were gathered; however, this research focused on modeling the Western Ontario and McMaster Universities' (WOMAC) disability score, which is typically treated as a continuous variable. The subjects' age and BMI at each time point as well as the subjects' sex were utilized as fixed effect regression predictors in the model. Age and BMI are two of the time-dependent covariates in this study, which do not remain constant over time. It was of interest to test the effect of BMI, which was treated as a type II time-dependent covariate within the model mentioned in Chapter III. This means there may be feedback between BMI and WOMAC disability score. The model is,

$$WOMAC_{it} = \beta_0 + \beta_1 Age_{it} + \beta_2 BMI_{it} + \beta_3 Sex_{it} + t_2 + t_3 + \varepsilon_{it}, \quad (4.7)$$

where age is a type I time-dependent covariate, t_2 is a type I time-dependent covariate and a time indicator of the second follow-up time for subjects, t_3 is a type I time-dependent covariate and a time indicator of the third follow-up time for subjects, and sex is a time-independent covariate. Type I time-dependent covariates are not stochastic and change predictably.

Within this model, the following hypothesis were tested

$$\begin{cases} H_0: \beta_2 = \mathbf{0} \\ H_1: \beta_2 \neq \mathbf{0} \end{cases}.$$

For this hypothesis, the power for different sample sizes was estimated using the Wald and DM statistics within the GMM-based power estimation method. The alternative hypothesis when calculating power is β_2 being equal to the population value for this parameter.

Data Generation

There is no data generating process for GMM due to the fact that GMM is a distribution free technique. Thus, in order to randomly generate data for this study and decide about the distribution of the random terms that needed to be used in the process of generating random responses, I needed to fit a mixed effect model. This required normalizing the response variable, which did not follow a normal distribution. Although the proposed power estimation process can also be applied to non-normal responses, the responses were normalized for this study. The reason for transforming the original WOMAC scores to a normal response was to keep the focus on estimating power, which was the main purpose of this study, not on non-linear modeling.

Figure 4.1 shows the histogram of the WOMAC scores which I believed to follow a Gamma distribution.

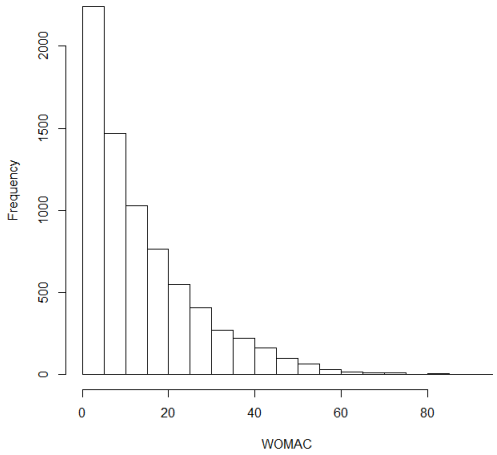

Figure 4.1. WOMAC Scores Histogram

In order to validate what the histogram of the WOMAC score was implying about the distribution of the response, I drew a Cullen and Frey graph in R using the "descdist" function from the "fitdistrplus" package. Cullen and Frey (1999) introduced their skewness-kurtosis graph, known as a Cullen and Frey graph, for the choice of distributions. Figure 4.2 shows the result, which implies the same type of distribution for the response variable. After seeing the Gamma distribution is a reasonable distribution for the WOMAC scores, the parameters of the Gamma distribution needed to be specified. Function "fitdist" from the "fitdistplus" package was used to fit a given distribution by maximum likelihood or matching moments. They suggested a shape of 0.95 and a rate of 0.08 for the Gamma distribution, which was fitted to the WOMAC scores. These estimated parameters of Gamma distribution were used in specifying the original distribution of the WOMAC score when applying transformations to it to normalize it for the future simulation steps.

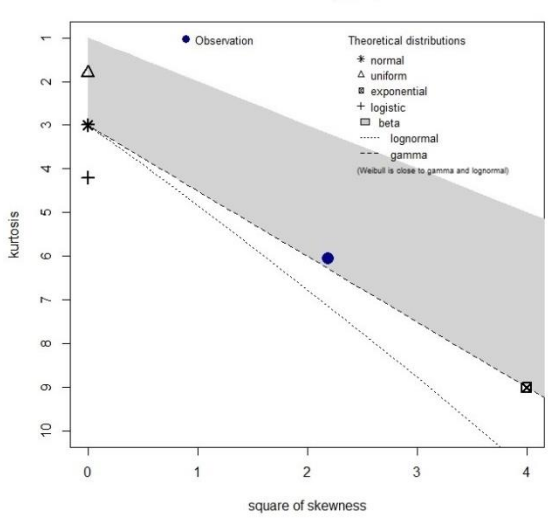

Figure 4.2. Cullen and Frey Graph of WOMAC Scores

To normalize the outcome WOMAC score variable, which had a Gamma distribution, a transformation needed to be applied to create the normalized WOMAC scores in the population data set. The idea of the final transformation that was applied to the response variable comes from the combination of two theories mentioned in Bain and Engelhardt (2009). The theories imply that no matter what the distribution of a variable is, if the cumulative distribution function of it is taken, then the cumulative distribution is passed into an inverse normal distribution function; the resulting values follow a normal distribution. This transformation was done using “pgamma” with the shape of .95 and rate of .08 on the WOMAC scores. The resulting values of the cumulative distribution were passed into the “qnorm” function in R to get the normalized WOMAC scores.

Figure 4.3 shows how the new transformed response looks. After fitting the normal distribution to it, the parameters for the normal distribution they followed were estimated to be 0.2 for the mean and 1.01 for the standard deviation using maximum likelihood estimation technique.

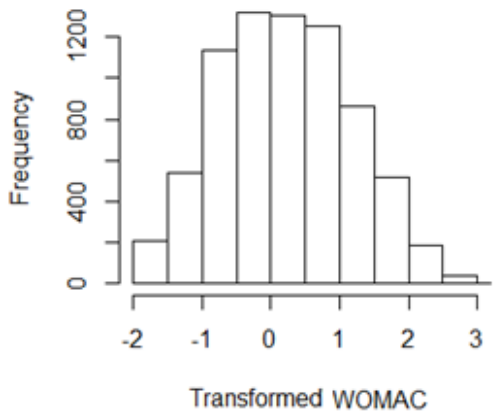

Figure 4.3. Normalized WOMAC Scores Histogram

Once the new normalized outcome was created, the goal changed to generating outcome variables that follow the same normal distributions as the new normalized outcome in the process of simulating data. As explained above, a mixed effect model was fitted to the transformed WOMAC score to figure out the coefficients of each of the covariates to be used in the simulation process later as well as to find out about the distribution of the random terms, which needed to be used in the response variable generating process. The model can be written as below

$$Y_{it} = \boldsymbol{X'}_{it}\boldsymbol{\beta}_{it} + u_{0i} + \varepsilon_{it},$$

where $\boldsymbol{X}_{it}$ is the matrix of the covariates, $\boldsymbol{\beta}_{it}$ is the vector of parameters, u_{0i} is the random intercept for each person, and finally ε_{it} is the random error term. Both the random intercept and random error follow a normal distribution with the mean of zero but different constant variances that needed to be estimated for the population used in the simulation process by fitting this model.

The “lmer” function from the “lme4” package in R was used to fit this random intercept model and the results are shown in Tables 4.1and 4.2 These results show the

two important pieces of information needed for the data simulation procedure: first, the coefficients for each variable and second, the variances to be used for generating random normal intercepts and error terms.

Table 4.1

Linear Mixed Model – Fixed Effects Estimates

Parameter	Estimate	Standard Error	t Value
Intercept	-1.695228	0.158606	-10.688
Sex	0.209491	0.033393	6.274
Age	0.003564	0.001819	1.959
BMI	0.048821	0.003109	15.705
t2	-0.092723	0.015575	-5.953
t3	-0.103803	0.015902	-6.528

Note. REML criterion at convergence: 16610.16

Table 4.2

Linear Mixed Model - Random Effects Estimates

Parameter	Variance	Standard Deviation
ID (Intercept)	0.5632	0.7505
Residual	0.2936	0.5418

The random error terms were randomly simulated from a normal distribution with the mean and variance of the estimated random effects from the aforementioned linear mixed model fitted to the population data,

$$\varepsilon_{it} \sim N(\mu = 0, \sigma = 0.54),$$

$$u_{0i} \sim N(\mu = 0, \sigma = 0.75).$$

The generated random intercept for each person stays the same for the three time points and the random error terms vary within each person as well as across patients. The reason for having the same randomly generated intercept for each person is to capture the autocorrelation that exists among the repeated measurements of each subject.

The steps for generating the response variable are described below. To make sure the generated response values followed the same distribution as the transformed WOMAC scores, first, the response values were simulated for all subjects in the pilot dataset. Then, the generated response values were plotted and they had the same distribution as the transformed WOMAC scores. Figure 4.4 shows the generated response values for the entire pilot population. Maximum likelihood estimation of the parameter of the normal distribution the generated responses follow was almost the same as the original transformed WOMAC scores. Some tests were applied to compare their distributions and there was no significant difference between the distribution of actual transformed response and the generated responses. The Kolmogorov–Smirnov test, which is a nonparametric test of the equality of continuous, one-dimensional probability distributions, was also applied to subsamples of the data to compare the distribution of the transformed WOMAC scores and the generated response variable and there was no significant difference in the distributions of the two variables (D = 0.11, p = .07). The reason for using sub-samples of the population to perform the Kolmogorov–Smirnov test is the large size of the population, which would result in the significance of any tests applied on them. As a result, random samples of 100 patients with unique IDs were

sampled with replacement multiple times and each time the Kolmogorov–Smirnov test resulted in no significant difference in the distributions of the transformed WOMAC scores and the generated response variables. This reassured me that the same procedure of generating response variables could be applied for the simulation procedure and the generated data would behave the same as the real data; hence, simulated data would be representative of the values seen in reality. Data generation code can be found in the Appendix A.

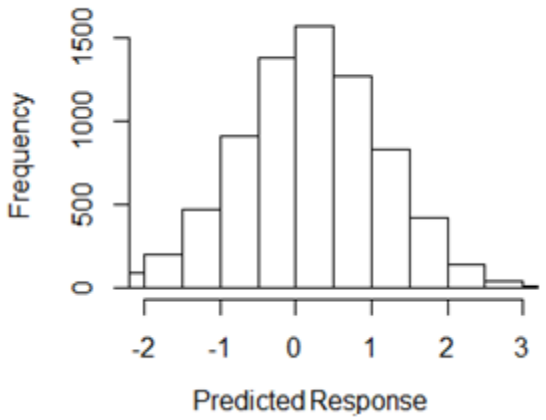

Figure 4.4. Histogram of the Generated Response Variable

Simulation Conditions and Procedure

Different sample sizes used for this simulation study include 25, 50, 100, and 200 subjects with three observations per subject due to having three follow-up times. Sample sizes of 100 and 200 were chosen according to the simulation study by Lyles et al. (2007), which focused on a GEE-based technique for power estimation of longitudinal data using the Wald test. Two smaller sample sizes of 25 and 50 were also added to this study to compare the accuracy of the estimated statistical power for the smaller sample sizes to the higher sample sizes of 100 and 200. This comparison was of interest to see

whether the methods that were shown to work for large sample sizes according to the proofs I constructed work as well in terms of the accuracy of the estimated power for the small sample sizes.

There were 3,600 data sets generated for each sample size within this simulation study. This number was calculated based on the theory and explained in Chapter III. Below, four steps for simulating the data sets for this study are summarized, but before looking at each step in detail, the entire simulation process is explained in one paragraph.

In summary, to simulate the data for this study, for each data set of size 25, 25 unique IDs from the population data set were randomly selected and then all three cases of predictors for each ID were selected. The "true" parameter values from the linear model, fitted to the entire population, were used to randomly generate responses for each case. This process was replicated 3,600 times to complete 3,600 data sets of size 25. Then, this process was repeated for sample sizes of 50, 100, and 200. This data simulation procedure is explained in detail below:

Step 1: Extracting the X values from the real dataset. At this step, for each sample size, unique IDs from the population data set were randomly chosen 3,600 times, which was the number of replications. Therefore, 3,600 datasets were randomly simulated within each sample size. The number of IDs chosen at this step depended on the sample size condition. There were four sets of sample sizes for this study: 25, 50, 100, and 200. So, for example, for sample size of 25, 25 unique ID's were selected from the population 3,600 times. This selection for each ID included their three time points resulting in 75 records of covariates.

Step 2: Creating the fixed effects. Using the extracted vector of covariates measured at each follow-up time for each subject formed the design matrix of the fixed effects (i.e., $\boldsymbol{X}_{it}$). The estimated coefficients of the fixed parameters from the original mixed effect model, called the "true" parameter values. formed the fixed effect parameter vector (i.e., $\boldsymbol{\beta}$).

Step 3: Generating the random effects. For each dataset, two random terms were generated at this step to be used as two random error terms in the process of generating the response variable. As explained before, one random intercept was generated for each subject following a normal distribution with the mean of 0 and standard deviation of 0.75 as

$$u_{0i} \sim N(\mu = 0, \sigma = 0.75),$$

where $i = 1, \ldots, n$ and n = 25, 50, 100, 200.

Then, this random number was used at all three follow-up times per each subject. Using the same random intercept term per subject was imposed to ensure that the similarities and autocorrelation that existed among the repeated measurements of each patient are being captured using this random intercept. Finally, the three random terms were generated for each subject following a normal distribution with the mean of 0 and standard deviation of 0.54 as

$$\varepsilon_{it} \sim N(\mu = 0, \sigma = 0.54),$$

where $i = 1, \ldots, n$ specifies the number of subjects, which for this study were four sets of sample sizes n = 25, 50, 100, 200 and $t = 1, \ldots, T$ specifies the number of repeated measures for each subject, which for this balanced study are the same per subject ($T = 3$).

Step 4: Generating the response values. At the final step, for each subject, the multiplication of the design matrix and "true" parameters from the model using the pilot data were added to the random intercept generated for that subject. Finally, the random error term was added to the previous addition to form the final transformed WOMAC score. Notice, at this step, to be consistent with the generated responses from the pilot data, the coefficient of BMI was being multiplied by 15 to increase the effect size as explained above. This step and why the effect size was increased for this study are explained below.

Controlling the Effect Size

When estimating the power of this study, the estimated powers ended up being very small, ranging from .05 to .1, for different data sets with sample sizes of 25 to 200 subjects. The small magnitude of the power would make it difficult, in the next steps of the simulation study, to see the changes in the magnitude of the estimated power values with the changes in the sample sizes. So, the effect size for the estimated parameter coefficient for BMI needed to be increased. BMI is the covariate whose effect size was controlled since BMI is the time-dependent covariate, which was tested in this study, hence; the magnitude of the power, which was calculated for the hypothesis test related to this variable, was directly affected by the changes in the effect size of this covariate.

Different constants ranging from 2 to 30 were multiplied by the coefficient of covariate of interest to find out which one would have the desired effect on the final estimated powers while using the same GMM estimates for all of the other parameters. Fifteen was the multiplier used for this study as it resulted in higher values of power, but not too high such as .999, for different sample sizes with the ability of capturing higher

ranges of power values. Consequently, when final response values were generated for this study, the BMI coefficient was multiplied by 15 but everything else stayed the same. Changing the effect size necessitated generating new responses. After generating the new responses, a random intercept mixed effect model was fitted to the population data set using the newly generated response. The estimated coefficients at this step were used as the "true" values of the parameters for the rest of the study. These "true" values are listed later in Table 4.4.

Algorithm for Generalized Method of Moments Estimation

The GMM estimation technique was explained in detail in Chapter III. The GMM method was used to estimate the parameters of the model within this study, so they could be used in the process of power estimation, which is explained in the next section. As explained in Chapter II and III of this dissertation and according to Hansen (2007), GMM estimation begins with a vector of population moment conditions taking the form in Equation 4.8 for all t

$$E[f(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)] = 0, \tag{4.8}$$

where $\boldsymbol{\beta}_0$ is an unknown vector in a parameter, $\boldsymbol{x}_{it}$ is a vector of random variables, $i = 1, \dots, n;\ t = 1, \dots, T$ and $f(.)$ is a vector of functions.

The GMM estimator is the value of $\boldsymbol{\beta}$ which minimizes the quadratic form shown in Equation 4.9

$$Q(\boldsymbol{\beta}) = \{n^{-1}\sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})\}' \boldsymbol{W} \{n^{-1}\sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})\}, \tag{4.9}$$

where $\boldsymbol{W}$ is a positive semi-definite weighting matrix, which may depend on the data but converges in probability to a matrix of constants which is positive definite and

$n^{-1}\sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})$ is the average of the sample moments. Therefore, by definition, the GMM estimator of $\boldsymbol{\beta}_0$ is

$$\widehat{\boldsymbol{\beta}} = arg\min_{\boldsymbol{\beta}\in\mathbb{P}} Q(\boldsymbol{\beta}), \tag{4.10}$$

where $arg\ min$ stands for the value of the argument $\boldsymbol{\beta}$ which minimizes the function.

Due to the lack of software availability to perform a GMM estimation and no known software package to perform the continuously updating GMM procedure, I wrote the R code for obtaining the parameter estimates of this GMM model due to its more efficient estimators in the second-order sense than the 2-step or k-step GMM estimators to improve the finite sample properties (Hall, 2005).

Four steps, which are summarized below, were taken in writing this R code:

Step 1: Defining the moment conditions. The moment conditions for this study needed to be defined depending on the types of time-dependent covariates in the model used within this study.

The moment conditions for this study were defined using Equation 4.10, which is defined for the repeated observations taken over T times on n subjects with J covariates, assuming that observations y_{is} and y_{kt} are independent whenever $i \neq k$.

$$E\left[\frac{\partial \mu_{is}(\boldsymbol{\beta})}{\partial \beta_j}\{y_{it} - \mu_{it}(\boldsymbol{\beta})\}\right] = 0, \tag{4.11}$$

where $\mu_{is}(\boldsymbol{\beta})$ represents the expectation of response measured for the ith subject at sth time, $\boldsymbol{y}_{it}$, based on the vector of covariate values, $\boldsymbol{x}_{it}$ and vector of parameters, $\boldsymbol{\beta}$.

To define the type of the time-dependent covariates, as explained in Chapter II, if Equation 4.11 holds for all s and t, then the jth covariate is classified as type I with T^2 moment conditions (Lai & Small, 2007). Variables age and time indicators were

classified as type I time-dependent covariates in this study as they could plausibly satisfy the condition that their outcomes are independent of past and future outcomes of the response. Therefore, nine moments were defined for each of them. If Equation 4.11 holds for $s \geq t$ but fails to hold for some $s < t$, the jth covariate is said to be type II**.** This type of covariate is common in a linear model with autoregressive responses (Lalonde et al., 2014) and BMI satisfied the conditions to be classified as a type II covariate with $\frac{T(T+1)}{2}$ moment conditions. Thus, six moment conditions were defined for this covariate.

In total, there were 39 moments that needed to be defined and entered into the GMM function. Three moment conditions for the intercept, three moment condition for the time-independent sex, nine moment conditions for the type I time-dependent covariate age, nine moment conditions for the time indicator t_2, nine moment conditions for the time indicator t_3, and finally six moment conditions for BMI, which is a type II time-dependent covariate. For each subject and considering the three time points, these moments were defined and saved to be used at the next step.

Step 2: Forming the vectors of the moment conditions. The vectors of sample moment conditions were defined at this step using the created moment conditions mentioned above. They were summed up for all subjects within each data set and finally averaged.

Step 3: Forming the weighting matrix. The weighting matrix, shown in Equation 4.12, was created using the suggestion of Lai and Small (2007),

$$W = S^{-1} = \left[\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})\, f(\boldsymbol{x}_{it}, \boldsymbol{\beta})^T\right]^{-1}. \tag{4.12}$$

Step 4: Defining the quadratic form. The quadratic form that was minimized is formed using Equation 4.9

$$Q(\boldsymbol{\beta}) = \left\{ n^{-1} \sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta}) \right\}' \boldsymbol{W} \left\{ n^{-1} \sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta}) \right\}.$$

As is obvious from the equation, this quadratic form is directly affected by the number of subjects and therefore the number of subjects involved in building the moment conditions, which are later summed up. This causes some issues with the power calculation for models using the GMM estimation technique. This issue is explained in detail under the GMM power section.

Step 5: Minimizing the quadratic form. The aforementioned quadratic form from Equation 4.9 was then minimized to find the GMM estimate of the parameters in the model. To do this, the "optim" function from the "stats" package was used. This general-purpose optimization works based on Nelder–Mead, quasi-Newton, and conjugate-gradient algorithms. Nelder–Mead method, that uses only function values and is robust but relatively slow, was used here. The code written for estimating the parameters using the continuously updating GMM estimation can be found in the Appendix B.

Convergence Problem

The GMM procedure and finding the GMM estimates within each simulated data set faced some issues regarding convergence. Not only did it affecting the final results of the estimates, it was biasing the final values of the quadratic term at the estimated parameters.

I had to re-write parts of the GMM function as well as monitor the convergence or non-convergence of the GMM process within each simulated data set to determine the

optimum number of iterations that needed to be used within each GMM estimation process. At the end to solve the issue for every simulated data set, the number of iterations had to be maxed to 10,000 iterations. Doing so, not even one non-convergence issue happened anymore for any of the 3,600 data sets for each of the four desired sample sizes, which resulted in close to unbiased estimates of parameters. Solving this issue insured the accurate and asymptotically unbiased estimation of the parameters of the model using the GMM estimation technique. Unfortunately, increasing the number of iterations resulted in the GMM process taking even longer to run, which was another issue that needed to be resolved within this study. This procedure is explained in the next section.

Issues Regarding the Run Time

There were many issues with the run time of the simulation, which made it impossible to finish this study in a reasonable amount of time. Overall, the expected run time was estimated to be over 515 days. The details regarding how long each part of the simulation process and data analyses originally took are described below.

Table 4.3 shows the original run time of this study for just GMM estimation, resulting in 114 days of run time excluding all the power calculation process. Including the power calculations would approximately triple the simulation run time. This run time estimate was made by assuming the time will increase linearly as the number of runs increased; however, the growth was not linear which resulted in an even longer run time.

The full process of running the GMM estimation procedure, hypothesis testing, and power calculation for one run of each sample size was 206 minutes. Assuming there was a linear growth in time by increasing the number of runs, 3,600 runs would take

741,600 minutes or 515 days. This run time was not feasible so I considered reducing the number of replications at this point but did not want to sacrifice the accuracy of this study by reducing the number of replications.

Table 4.3

Run Time for GMM Procedure

	Run Time for 1 Run in Minutes	Run time for 3,600 runs in minutes	Run time for 3,600 runs in Days
n=25	4.3163	15,538.68	10.79075
n=50	8.0935	29,136.6	20.23375
n=100	10.2232	36,803.52	25.558
n=200	22.842	82,231.2	57.105

Multiple options including renting space on Amazon web services, getting access to the university's super computer, and using multiple cores to run the simulation, parallelizing the simulation, and re-writing parts of the code were considered. Almost all of these options had to be taken advantage of in order to finish the originally proposed simulation without having to change the number of replications. These steps included first, removing any "filter" function from the study and replacing filters with other selection options which would take a shorter time than "filter."

A second option was to produce only the required statistics and results and removing extra information which was being stored. Any additional piece of information, that was originally being extracted, was removed to speed up the simulation process.

Re-writing parts of the GMM function, which was the most time-consuming part of this process, was a third option. The structure of all the data frames was changed to

vectors and matrices and the algorithm was written according to the new structure of the data. This step was the most effective solution I could come up with to reduce the run time.

Fourth, I had to parallel-program the entire code. This parallelization included two parts: (1) running the code on multiple computers while making sure the same seed was used for all of them and (2) parallelizing the workload into multiple cores of each computer to take advantage of all the cores of each machine used for running part of the analysis. The regular machines used for parts of this analysis had eight cores and the super computer, which was used for other parts of the analysis, had 12 cores.

Fifth, the machines with eight cores were used to run the analysis on the smaller sample sizes and the super computer, which to which I had to request access, was used to run the analysis on the larger sample sizes.

I was able to decrease the run time of the GMM estimation, hypothesis testing using Wald statistic, and post-hoc power estimation to about 90 hours or less than four days, which was a great improvement from the original 515 days to run the same thing. Of course, multiple machines and parallel programming, resulting in using multiple cores on each machine, were involved in achieving the goal of decreasing the run time. These 90 hours do not include the run time for the DM tests and calculating their power and generating the data. Data generation procedure was taken out of the original code and run separately to save time. Completing those tasks took another 90 hours or so, which needed to be done separately.

Distance Metric Statistic and Issues Regarding This Statistic

There was some uncertainty regarding calculating the DM statistic and getting different results from this test compared to what the Wald statistic, which was the main focus of this study. According to Hall (2005), "the DM test examines the impact on the GMM minimand of the imposition of the restrictions" which needs to be calculated using Equation 4.4. Within this equation, in order to find the DM statistic, a function of two quadratic forms needs to be found as below

$$T_{DM}^{*} = n\left[Q\left(\widetilde{\boldsymbol{\beta}}\right) - Q\left(\widehat{\boldsymbol{\beta}}\right)\right],$$

where within the DM statistic, $Q(.)$, is the quadratic form from the GMM algorithm which needs to be found based on the restricted, $\widetilde{\boldsymbol{\beta}}$ and unrestricted, $\widehat{\boldsymbol{\beta}}$, parameter estimators, respectively and then to be used in finding the difference between the respective quadratic forms. Hall (2005) mentioned "the unrestricted estimator is just the GMM parameter estimates and the restricted estimator of $\boldsymbol{\beta}$ which minimizes the quadratic form subject to $r(\boldsymbol{\beta}) = 0$ and both these minimizations use the same weighting matrix". This being said, I tried imposing the restriction from the null hypotheses to find the quadratic form for the restricted parameter estimates, $Q\left(\widetilde{\boldsymbol{\beta}}\right)$, in two different ways:

Within the first method of calculating the DM statistic, the quadratic form used for fitting the original GMM to all the parameters using 39 moment conditions was also used for $Q\left(\widetilde{\boldsymbol{\beta}}\right)$ by using the unrestricted GMM estimates and imposing the estimate of the BMI parameter to be equal to zero and then calculating the quadratic form. This quadratic term was calculated using the unrestricted parameter estimates from the newly calculated quadratic form based on the restricted parameter estimates. Then, the difference between

two quadratic terms was found and finally multiplied by the sample size used. These values were very large resulting in always rejecting the null hypothesis for different sample sizes.

The second method, which is believed to correctly calculate the DM statistics, involves writing another GMM function excluding all the moment conditions related to the variable of interest which imposes the restrictions, BMI here, from the quadratic form. Therefore, this quadratic form was written using 33 moment conditions and then the new GMM function was applied to all 3,600 data sets within each sample size, estimating every parameter except from the one for the BMI that was excluded from the model. Then the newly constrained estimates were substituted into the original quadratic form that included 39 moment conditions and were used for the unrestricted parameter estimation. This value was then saved as the restricted quadratic value and the unrestricted quadratic value was subtracted from it and then multiplied by the sample size. So, the DM statistic for this analysis was calculated as

$$T_{DM} = n[Q(\beta_0, \beta_1, \beta_2, 0, \beta_4, \beta_5) - Q(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)].$$

Even though, these values are more reasonable, they do not seem to give the same results compared to the Wald test. The summary of these statistics can be found later but all in all the DM statistics do not seem to behave similarly to the Wald statistics and do not seem to have the same distribution as the Wald statistics for this study. This could be due to not having large enough sample sizes in order for the two statistics to have the same asymptotic distributions or some of the assumptions might not be met for them to have identical chi-square distributions.

These assumptions are mentioned in Hall (2005). It is believed that the first assumption might not be met in the current study due to the use of real data in the process of simulating the data. These 13 assumptions are:

1. Strict stationary process to be formed by the random vectors. This implies all expectations of functions of the random variables to be independent of time.
2. Regularity conditions for the function of the moments and the ability to measure them.
3. The population moment condition assumption which refers to the random vector and the parameter vector satisfying the population moment condition: $E[f(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)] = 0$.
4. Global identification which is $E[f(\boldsymbol{x}_{it}, \overline{\boldsymbol{\beta}})] \neq 0$ for all $\overline{\boldsymbol{\beta}}$ such that $\overline{\boldsymbol{\beta}} \neq \boldsymbol{\beta}_0$.
5. Regularity condition on $\frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$ which refers to this derivative matrix to exist and be continuous for each of the random vectors, $\boldsymbol{\beta}_0$ being an interior point of the set, and $E\left[\frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right]$ existing and being finite.
6. Assumptions regarding the weighting matrix.
7. Ergodicity of the random process.
8. The set being compact.
9. Domination of $f(\boldsymbol{x}_{it}, \boldsymbol{\beta})$.
10. Assumptions regarding the variance of the sample moment.
11. Continuity of $E\left[\frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right]$.
12. Uniform convergence of $\boldsymbol{G}_n(\boldsymbol{\beta})$.

13. Regularity condition for $r(.)$ which includes its being a vector of continuous differentiable functions and $rank\{R(\boldsymbol{\beta}_0)\} = s$ where $(\boldsymbol{\beta}) = \frac{\partial r(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$.

Power Estimation Procedure

After the data were simulated and the method for estimating the model parameters using GMM was developed, it was time to figure out the power estimation procedure using GMM and figure out how the theoretically developed power estimation methods from Chapter III compare to the empirical results from Chapter IV.

Considering the repeated measures, explained in previous chapters of this dissertation, at T time points for n subjects, in order to find the power and then the required sample size of a statistical test, first the hypothesis needs to be specified and tested as shown before,

$$\begin{cases} H_0: r(\boldsymbol{\beta}) = \mathbf{0} \\ H_1: r(\boldsymbol{\beta}) \neq \mathbf{0}' \end{cases}$$

where this hypothesis can be simplified to the hypothesis mentioned below for this study,

$$\begin{cases} H_0: \beta_2 = \mathbf{0} \\ H_1: \beta_2 \neq \mathbf{0}' \end{cases}$$

in which, β_2 is tested to see whether the effect of the type II time-dependent covariate, BMI, in predicting the transformed WOMAC score is significant.

Then the statistic which is used to test this hypothesis needs to be specified. Because the GMM approach was adopted as the estimation method for this study, the Wald statistic from Equation 4.1 and the DM statistic from Equation 4.4 were used to test the hypothesis mentioned above.

In order to estimate the statistical power of these tests, the distributions of these statistics under the null and alternative hypothesis need to be specified. From Chapter III

and Hall (2005), we know the asymptotic distributions of Wald and DM statistics under the null hypothesis are equivalent as below

$$H_0: T_W^* \xrightarrow{d} \chi^2_{(s)}, \qquad T_{DM}^* \xrightarrow{d} \chi^2_{(s)}.$$

Newey and West (1987) showed that the asymptotic equivalence of the statistics extends to the alternative hypothesis. As discussed before, under the alternative hypothesis, the Wald and DM statistics have an asymptotic non-central chi-square distribution of $\chi^2_{(s),\lambda}$ with the non-centrality parameter that could be calculated using Equation 4.5.

In order to estimate the power, assuming that α represents the type I error, $\chi^2_{(s);1-\alpha}$ is the critical value from the central $\chi^2_{(s)}$ distribution. Using this critical value, power can be calculated using Equation 4.6 by finding the probability of

$$\Pr\left(\chi^2_{s,(\lambda)} \geq \chi^2_{s,1-\alpha}\right),$$

with $\chi^2_{s,1-\alpha}$ denoting the $100(1-\alpha)$th percentile of the central chi-square with s degrees of freedom. Thus, the power associated with the Wald and DM test statistics is

$$1-\gamma = \int_{\chi^2_{(s);1-\alpha}}^{\infty} f(x_t; s, \lambda)dx,$$

where γ represents the type II error and $f(x_t; s, \lambda)$ is the probability density function of $\chi^2_{(s),\lambda}$. This process is explained in detail for the model fitted in this study.

Calculating the Theoretical Powers

There are multiple steps I developed to calculate the theoretical power for the pilot data set with 2,456 subjects. First, the mixed effect model was fitted to the OAI dataset using the newly generated normalized WOMAC score after increasing the effect of BMI. The coefficients are listed in Table 4.4 for the sake of comparison to the GEE

and GMM estimates of the same models. As explained before, these estimates are referred to as the "true" parameter values.

Table 4.4

Mixed-Effects Model Summary

Parameters	Intercept	Sex	Age	BMI	t_2	t_3
Coefficients	-1.51988	0.19867	0.00248	0.72956	-0.11230	-0.10446

Table 4.5 shows the coefficients of the model fitted to the data using GEE with the independence covariance structure. These values were used as the initial values of the unknown parameters within the GMM function to get the GMM estimates. GEE estimates with the independent covariance structure are believed to be the closest to the GMM estimates, making them the best option as the initial values to be used in the process of optimization of quadratic form within the GMM function.

Table 4.5

GEE Model Summary

Parameters	Intercept	Sex	Age	BMI	t_2	t_3
Coefficients	-1.48183	0.19881	0.00241	0.72840	-0.11226	-0.10424

These initial values were used within the GMM function to find the GMM estimates of the parameters used in the model. The estimated parameters using GMM are listed in Table 4.6. These values are close to the estimated values using the GEE method and the estimated parameters from the mixed effect model. The effects should be similar

for all three methods; it is the standard errors that change for different models fitted to Equation 4.7, which is the model used for this study. These results show the accuracy of the GMM function I wrote and are the assurance for moving forward with the rest of the power estimation procedure using the pilot data.

Table 4.6

GMM Model Summary

Parameters	Intercept	Sex	Age	BMI	t_2	t_3
Coefficients	-1.48913	0.20018	0.00255	0.72824	-0.11059	-0.10393

The next step involved extracting the quadratic form at the GMM estimated parameters, which was equal to 0.004 for the entire pilot data. This value needs to be used in the process of calculating the non-centrality parameter of the non-central chi-square distribution, which is the distribution of Wald and DM statistics under the alternative hypothesis. The non-centrality parameter was equal to 5.219863, resulting in a power of .627 for the entire data set using *N*=2,456.

I originally believed that by changing the sample sizes and using different sample size values in the process of calculating the non-centrality parameters, I could estimate the power using the quadratic forms and estimate the parameters from the pilot data set. Instead, after using different sample sizes, calculating the power, and comparing them to the post-hoc powers calculated for each sample size, I learned this process could not be done in this way within GMM even though it is the common way of calculating power for other models. The calculated power for sample size of 25 using the estimated parameters and quadratic form from the large pilot dataset was .056. Power calculated the

same way for sample sizes of 50, 100, and 200 were .062, .075, and .099, respectively. However, the post-doc power calculated for the data sets with those sample sizes appeared to be a lot higher when the GMM was fitted to smaller sample sizes. This showed that the GMM power calculation procedure is tied to the size of the pilot data, the estimated parameters, and the magnitude and the number of response variables used in the process of calculating the non-centrality parameter of the non-central chi-square distribution. The estimated parameters and the number of subjects within each data set used in the GMM estimation procedure, directly reflect the summation of the moment conditions and hence the quadratic form of a GMM function. This makes the non-centrality parameter of the non-central chi-square distribution very sensitive to the number of subjects used in the study.

Below, it is shown theoretically how the non-centrality parameter is influenced by the size of the pilot data used in the process of power calculation. Because only one parameter was tested within this study, the non-centrality parameter can be simplified to

$$\lambda = n\beta^2 \boldsymbol{G}^T \boldsymbol{W} \boldsymbol{G}.$$

After substituting the simplified versions of $\boldsymbol{G}$ and weighting matrix for this model, the non-centrality parameter can be written as

$$\lambda = n\beta^2 \left(n^{-1} \sum_{i=1}^{n} \frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T \left(\frac{1}{n} \sum_{i} f(\boldsymbol{x}_{it}, \boldsymbol{\beta}) f(\boldsymbol{x}_{it}, \boldsymbol{\beta})^T \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} \frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right),$$

where all n terms will be canceled out and the magnitude of the resulting non-centrality parameter will increase as the number of terms added together increases by the increase of sample size. Here is the final non-centrality parameter,

$$\lambda = \beta^2 \left(\sum_{i=1}^{n} \frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T \left(\sum_{i} f(\boldsymbol{x}_{it}, \boldsymbol{\beta}) \, f(\boldsymbol{x}_{it}, \boldsymbol{\beta})^T \right)^{-1} \left(\sum_{i=1}^{n} \frac{\partial f(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right).$$

The dependence between the quadratic forms and moment conditions used in the calculation of non-centrality parameters are to the number of subjects, makes it inappropriate to use the quadratic form from the pilot data with a specific number of subjects to calculate the power for future samples with a different number of subjects from the pilot data. This is because the quadratic forms of the pilot data sets with different number of subjects will not be representative of the new dataset with a different number of subjects. In addition, response values are part of the quadratic form and must reflect different effect sizes.

After conducting some theoretical work and testing them on the real data, the final answer for finding the theoretical power for this study can be described in the six steps below:

Step1. Multiple subsamples of the pilot data set need to be taken for each of the sample sizes considered as future sample sizes for future studies. For this study, 100 randomly selected data sets of 25 subjects were selected. The same process was carried out for the sample sizes of 50, 100, and 200 meaning that 100 data sets of each size were randomly selected from the original pilot dataset of size 2,456 subjects.

Step 2. The GEE was fitted to each of the data sets and the parameter estimates were extracted as the initial values to be used within the GMM optimization process.

Step 3. The GMM estimation was applied within each data set and the estimated parameter of interest as well as the value of the quadratic form at the GMM estimated parameters were extracted from each dataset.

Step 4. The non-centrality parameter was calculated for each dataset using the GMM estimates and quadratic value coming from that dataset with the same sample size, which was used in the calculation of the non-centrality parameter.

Step 5. Now, there are 100 non-centrality parameters for samples with size of 25, 100 non-centrality parameters for samples with size of 50, 100 non-centrality parameters for samples with size of 100, and finally 100 non-centrality parameters for samples with size of 200. In order to get one non-centrality parameter for each sample size, the 100 non-centrality parameters of each sample size were averaged.

Step 6. Power was calculated for each sample size using the averaged non-centrality parameters as below

$$1 - \gamma = \int_{\chi^2_{(1)}}^{\infty} f_{\chi^2}(1, \bar{\lambda}) dx.$$

The results of the theoretical power for the four sample sizes considered for this study are summarized in Table 4.7.

Table 4.7

Theoretical Powers for Different Sample Sizes (Using GMM at Each Sub-Sample)

Sample Size	Averaged Non-centrality Parameter	Power
n=25	5.75638551	.6697782
n=50	6.070270952	.6928137
n=100	7.296046533	.770702
n=200	7.479371536	.780796

These non-centrality parameters need to be averaged at the end and the power needs to be calculated for the averaged non-centrality parameter. This means the

integration should happen at the end rather than integrating the non-central chi-square distribution for each sub-sample, finding the power 100 times and at the end averaging the powers.

The reason for averaging the non-centrality parameters first and then finding the theoretical power for them, rather than finding the power multiple times for each sub-sample and then averaging them, which I also tried, is the sensitivity of the power to the non-centrality parameter of each sub-sample. This sensitivity produces higher variance of power than the real power values. This higher variance of the multiple calculated powers results in skewing the mean of the powers when trying to find one theoretical power at the end. Looking at the power calculation process,

$$1-\gamma=\int_{\chi^2_{(1)}}^{\infty}\frac{e^{-\frac{x+\lambda}{2}}}{2}\left(\frac{x}{\lambda}\right)^{-\frac{1}{2}}I_{-\frac{1}{2}}\left(\sqrt{\lambda x}\,\right)dx,$$

where $I_v(y)$ is a modified Bessel function, clarifies the relationship between the non-centrality parameter, λ, and the power, $1-\gamma$, and how when λ gets smaller, power gets extremely small. Once all non-centrality parameters of the representative sub-samples were averaged and then one theoretical power for the mean of the non-centrality parameters was calculated, the power was representative of the actual power and close to the post-hoc power of the 3,600 simulated data, which is reported later.

Another way to calculate the theoretical power is taking the same steps except for step 4. Instead, step 4 is done using the estimated parameters and quadratic form which come from the GMM estimation of the entire population or pilot dataset. This is faster as the GMM estimation procedure is conducted only once and then the non-centrality parameter is calculated for each dataset using the GMM estimates and quadratic value

coming from the pilot dataset. These then are used in the calculation of the non-centrality parameter for each sub-sample. The resulting power values using this method are higher than the ones calculated above and also higher than the post-hoc power but they are closer to the rejection rate of the simulation study explained later. These powers are listed in Table 4.8.

When the size of the pilot data is smaller than the sizes of data sets considered for future studies, multiple data sets of the desired sizes need to be simulated using the characteristics of the data. Then the same steps should be applied to them to calculate the powers for different sample sizes following the rules from Lyles et al. (2007).

Table 4.8

Theoretical Powers for Different Sample Sizes (Using GMM Estimates of Population)

Sample Size	Averaged Non-centrality Parameter	Power
n=25	9.033433	.8521282
n=50	10.0974	.8883268
n=100	10.49592	.8996884
n=200	10.82056	.9081583

This section shows the application of the proposed power estimation techniques on one real data set to provide a guideline for applied researchers and practitioners in how to apply these methods and estimate the power as well as evaluate the performance of the proposed models using Wald and DM statistics. Two methods for calculating the theoretical power have been provided from many methods, which I have been examined. For larger sample sizes, these two power estimation methods should perform similarly.

For the smaller sample sizes used for this study, they performed slightly differently but one was close to the post-hoc powers and the other was close to the rejection rates. The first method is the one I recommend.

As explained these powers increased by the increase of the sample size and effect size. To consider it, not only was the theoretical power calculated for different sample sizes, but it was also calculated for different effect sizes. These effect sizes were implemented by multiplying BMI by 5 and 10 as well as 15, which is the effect size used within this study. The calculated theoretical powers were then plotted against each other. Figure 4.5 shows these theoretical powers for different effect sizes across four sample sizes of interest on the OAI population dataset. It shows the increase in the effect size and sample size of a study result in an increase in the estimated GMM based power.

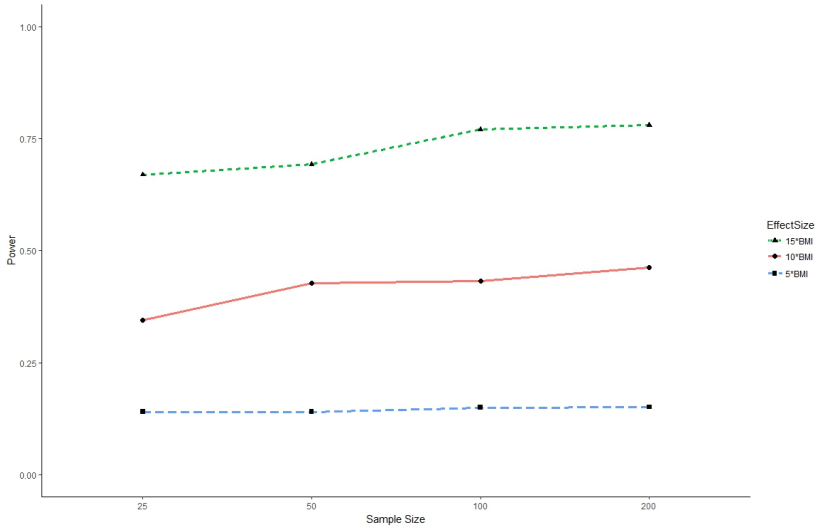

Figure 4.5. Line Chart of the Theoretical Powers for Three Effect Sizes and Four Sample Sizes

Hypothesis Testing for Each Simulated Data

The process of simulating 3,600 datasets for each sample size of 25, 50, 100, and 200 was explained earlier. Once each dataset was created, within each data set, the hypothesis about the significance of BMI needed to be tested. For each of those data sets, after testing the hypothesis, the power was also calculated. Even though this sounds like a post-hoc power calculation and is also called post-hoc power throughout this paper, it was not calculated for the same purpose as some practitioners calculate the post-hoc power which is not supported by some statisticians (Hoenig & Heisey, 2001). In this study, each of the simulated data sets were also considered as a pilot data set of one of the desired four sizes and the power was calculated for each of the data sets to figure out the distribution of the powers for different sample sizes and compare the theoretical powers to them. Below, multiple steps of hypothesis testing and power calculation for the simulated data are explained:

Step 1. In order to save time while running the analysis, the entire data set was simulated separately and saved in one master data set. This master simulated data set included 3,600 data sets of sample size 25 (270,000 rows), 3,600 data sets of sample size 50 (540,000 rows), 3,600 data sets of sample size 100 (1,080,000 rows), and 3,600 data sets of sample size 200 (2,160,000 rows). Altogether, the master simulated data set had 4,050,000 records. At each run, one of the simulated data sets from each sample size was selected and used to perform the power analysis.

Step 2. A GEE was fitted to the selected data set using the “independence” correlation structure to find the initial, estimated points that needed to be used within the GMM function to estimate the model parameters using the GEE package in R.

Step 3. The written GMM function was then used within each data set to find the GMM estimates for the parameters of the model and to test for the effect of the BMI later. The estimated parameters were extracted from each data set to be used in the future steps.

Step 4. The quadratic form for the GMM estimated parameters was then extracted for use in the process of calculating the Wald and DM statistics to test the BMI effect. It was also used in calculating the post-hoc powers of each data set. One additional quadratic form needed to be calculated for the DM statistic calculation, which is explained later in a separate section.

Step 5. The Wald and DM statistics were then calculated using the quadratic form at the estimated values of the parameters and the GMM estimates of the BMI. The Wald and DM statistics were then extracted to be compared to the critical value later.

Step 6. Each Wald and DM statistic was compared to the critical value, which comes from the distribution of the Wald statistic under the null hypothesis that is chi-square with the degree of freedom of one. This value is equal to 3.841459 here.

Step 7. The final decision regarding rejecting or not rejecting the null hypothesis of $\beta_{BMI} = 0$ was made for each data set after comparing the calculated Wald and DM test statistics to the critical value.

Step 8. Within each sample size, there were 3,600 data sets and the rejection rate was calculated for each of the sample sizes by dividing the number of the rejected hypotheses by 3,600. Then the 95% confidence interval was found and reported for each rejection rate. The rejection rates and their confidence intervals are reported in the simulation results section.

Distribution of Powers for Each Simulated Data

As explained above, to figure out the distribution of the powers for different sample sizes and to see how they improve by the increase of sample size, after each hypothesis was tested, the power for that data set was also calculated. Once again, the idea of post-hoc power is not recommended here and this calculation is only being made to see how these powers are distributed across sample sizes by considering each of the simulated data as pilot data sets.

To find out about the distribution of statistical powers of the simulated data, steps 1 through 6, which were used to test the hypothesis about each data set, remain the same and a few more steps were added to the analysis. These steps are as below:

Step 9. The non-centrality parameter was calculated for each data set, as below, using the GMM estimated parameters and the quadratic value at the GMM estimated parameters. This non-centrality parameter for the model considered in this study simplifies to

$$\lambda = n\hat{\beta}_{BMI} Q(\widehat{\boldsymbol{\beta}}) \hat{\beta}_{BMI}.$$

Step 10. The power was then calculated by integrating the non-central chi-square distribution of the Wald statistic under the alternative hypothesis and the area under the non-central chi-square curve was calculated from 3.841459 which is the value of the central chi-square distribution with one degree of freedom, that is the distribution of the Wald-statistic under the null hypotheses, to infinity.

Step 11. All these 3,600 powers within each simulated data were then averaged to find the mean of all these powers (called post-hoc power here). The median power was

also calculated and reported as well as some other descriptive statistics in the simulation results section.

Simulation Results

The results for the simulation studies of four sample sizes of 25, 50, 100, and 200 are reported in four sub-sections below.

Summary of Simulation Results for Sample Size of 25

The simulation for sample size of 25 for the hypothesis test of BMI effect on 2,973 out of 3,600 data sets resulted in an 82.6 % null hypothesis rejection rate using the Wald test. The rejection rate for the same simulated data sets using the DM statistic was 91.58%, which is a lot higher than the Wald test results. As explained before, I do not recommend using the DM statistic. The average post hoc power was .628 and the median was .635. The average GMM estimated BMI parameter was close to the average GEE estimated BMI parameters and was around .73. These estimates were close to the population parameter estimate for BMI listed in Table 4.4. All simulation results for data sets of size 25 are summarized in Table 4.9.

As obvious from Table 4.9, for sample size of 25, the theoretical power of .6698 is very close to the post-hoc power and it falls into the 95% bootstrap confidence interval of the post-hoc power with the lower confidence limit of .3450 and upper confidence limit of .879. However, the rejection rate of the simulated data using the Wald test is much larger than the theoretical power showing us that for smaller sample sizes, the rejection rates do not line up with the calculated powers using the Wald statistic.

Table 4.9

Simulation Results for 3,600 Data Sets of Size 25 (Theoretical Power=.6698)

Rejection Rate using Wald Test	Rejection Rate using DM Test	Post-hoc Power of 3,600 Simulated Data	Average BMI Parameter Estimate
.8260	.9158	Mean: .62837 Q1: .54064 Q2 (Median): .63509 Q3: .72214	$\bar{\hat{\beta}}_{GMM} = .7349$ $\bar{\hat{\beta}}_{GEE} = .7331$
Confidence Interval: (.8136, .8384)	Confidence Interval: (.9067, .9249)	Bootstrap CI: (.3450, .8790)	

As mentioned before, the DM statistic, which was claimed by Hall (2005) to have the same asymptotic distribution as the Wald statistics, produced very high rejection rates and therefore is not recommended at least for smaller sample sizes and under circumstances in which any of the 13 assumptions mentioned before might not be met.

Summary of Simulation Results for Sample Size of 50

The simulation for sample sizes of 50 for the hypothesis test of BMI effect on 3,389 out of 3,600 data sets resulted in a 94.1 % rejection rate of the null hypothesis using the Wald test. The rejection rate for the same simulated data sets using the DM statistic was 99.7%, which is higher than the Wald test results. As explained before, I do not recommend using the DM statistic. The average post hoc power was .71 and the median was .72. The average GMM estimated BMI parameter was close to the average GEE estimated BMI parameters and was around .73. All results are summarized in Table 4.10.

Table 4.10

Simulation Results for 3,600 Data Sets of Size 50 (Theoretical Power=.6928)

Rejection Rate using Wald Test	Rejection Rate using DM Test	Post-hoc Power of 3,600 Simulated Data	Average BMI Parameter Estimate
. 9410	. 9970	Mean: .7105 Q1: .6328 Q2 (Median): .7213 Q3: .8029	$\bar{\hat{\beta}}_{GMM} = .7321$ $\bar{\hat{\beta}}_{GEE} = .7319$
Confidence Interval: (.9333, .9487)	Confidence Interval: (.9952, .9988)	Bootstrap CI: (.4368, .9115)	

As is obvious from Table 4.10, for sample size of 50, the theoretical power of .6928 is very close to the post-hoc power and it falls into the 95% bootstrap confidence interval of the post-hoc power with the lower confidence limit of .4368 and upper confidence limit of .9115. However, the rejection rate of the simulated data using the Wald test is much larger than the theoretical power showing that for smaller sample sizes, the rejection rates do not line up with the calculated powers using the Wald statistic. As mentioned before, the DM statistic, which was claimed by Hall (2005) to have the same asymptotic distribution as the Wald statistics, produced very high rejection rates and therefore is not recommended at least for smaller sample sizes and under circumstances in which any of the 13 assumptions mentioned before might not be met.

Summary of Simulation Results for Sample Size of 100

The simulation for sample sizes of 100 for the hypothesis test of BMI effect on 3,469 out of 3,600 data sets resulted in a 96.4 % rejection rate of the null hypothesis using the Wald test. The rejection rate for the same simulated data sets using the DM

statistic was 100%, which is somewhat higher than the Wald test results. As explained before, I do not recommend using the DM statistic. The average post hoc power was .76 and the median was .77. The average GMM estimated BMI parameter is close to the average GEE estimated BMI parameters and is around .73. All the results are summarized in Table 4.11.

As is obvious from Table 4.11, for sample size of 100, the theoretical power of .7707 is very close to the mean of the post-hoc power and almost equal to their median. It falls into the 95% bootstrap confidence interval of the post-hoc power with the lower confidence limit of .4556 and upper confidence limit of .9381. However, the rejection rate of the simulated data using the Wald test is much larger than the theoretical power, showing us that for smaller sample sizes, the rejection rates do not line up with the calculated powers using the Wald statistic.

Table 4.11

Simulation Results for 3,600 Data Sets of Size 100 (Theoretical Power=.7707)

Rejection Rate using Wald Test	Rejection Rate using DM Test	Post-hoc Power of 3,600 Simulated Data	Average BMI Parameter Estimate
.9640	1. 0000	Mean: .7569 Q1: .6878 Q2 (Median): .7750 Q3: .8488	$\bar{\hat{\beta}}_{GMM} = .7322$ $\bar{\hat{\beta}}_{GEE} = .7322$
Confidence Interval: (.9579, .9701)	Confidence Interval: NA	Bootstrap CI: (.4556, .9381)	

As mentioned before, the DM statistic, which was claimed by Hall (2005) to have the same asymptotic distribution as the Wald statistics, is producing very high rejection

rates, representing Type I error, and therefore is not recommended at least for smaller sample sizes and under circumstances in which any of the 13 assumptions mentioned before might not be met.

Summary of Simulation Results for Sample Size of 200

The simulation for sample sizes of 200 for the hypothesis test of BMI effect on 3,494 out of 3,600 data sets resulted in a 97.06 % rejection rate of the null hypothesis using the Wald test. The rejection rate for the same simulated data sets using the DM statistic was 100%, which is slightly higher than the Wald test results. As explained before, I do not recommend using the DM statistic for this sample size either. The post hoc power of the simulated data sets ranged from .2728 to .9964. The average post hoc power was .78 and the median was .798. The average GMM estimated BMI parameter was close to the average GEE estimated BMI parameters and was around .73. All results are summarized in Table 4.12.

As is obvious from Table 4.12, for sample size of 200, the theoretical power of .7807 is very close to the median of the post-hoc powers and almost equal to their mean. It falls into the 95% bootstrap confidence interval of the post-hoc power with the lower confidence limit of .4860 and upper confidence limit of .9584. However, the rejection rate of the simulated data using the Wald test is much larger than the theoretical power showing us that for smaller sample sizes, the rejection rates do not line up with the calculated powers using the Wald statistic. So, it appears the sample size of 200 is still too small for the Wald test to perform as it is theoretically expected to behave while using the data with characteristics of the OAI data. As mentioned before, the DM statistic, which was claimed by Hall (2005) to have the same asymptotic distribution as the Wald

statistics, produced very high rejection rates and therefore is not recommended at least for smaller sample sizes and under circumstances in which any of the 13 assumptions mentioned before might not be met.

Table 4.12

Simulation Results for 3,600 Data Sets of Size 200 (Theoretical Power=.7807)

Rejection Rate using Wald Test	Rejection Rate using DM Test	Post-hoc Power of 3,600 Simulated Data	Average BMI Parameter Estimate
.9706	1. 0000	Mean: .7788 Q1: .7030 Q2 (Median): .7981 Q3: .8725	$\bar{\hat{\beta}}_{GMM} = .7323$ $\bar{\hat{\beta}}_{GEE} = .7322$
Confidence Interval: (.9651, .9761)	Confidence Interval: NA	Bootstrap CI: (.4860, .9584)	

These results clearly show that the post-hoc powers are right in line with the calculated theoretical powers showing the accuracy of the power calculation technique developed in this dissertation. It is obvious that by an increase in sample size, the theoretical powers get much closer to the measures of central tendency of the post-hoc powers. Figure 4.6 shows the box plots of the post-hoc powers for different sample sizes displaying the distribution of the post-hoc powers. As is clear from these box plots, by increase sample sizes, the post-hoc power values move higher. The theoretical powers are indicated on the box plots, using circles, and connected to each other, using a solid line. They show an increasing trend by the increase of sample size and they obviously are very close to the center of the box plots. The rejection rates based on the Wald test are also indicated on the box plots and connected using dotted lines. They also show an increasing trend by the increase in sample size but they do not fall within the 25^{th} and 75^{th}

percentiles of each box plot showing they are much higher than the mean of the post hoc powers and the theoretical powers.

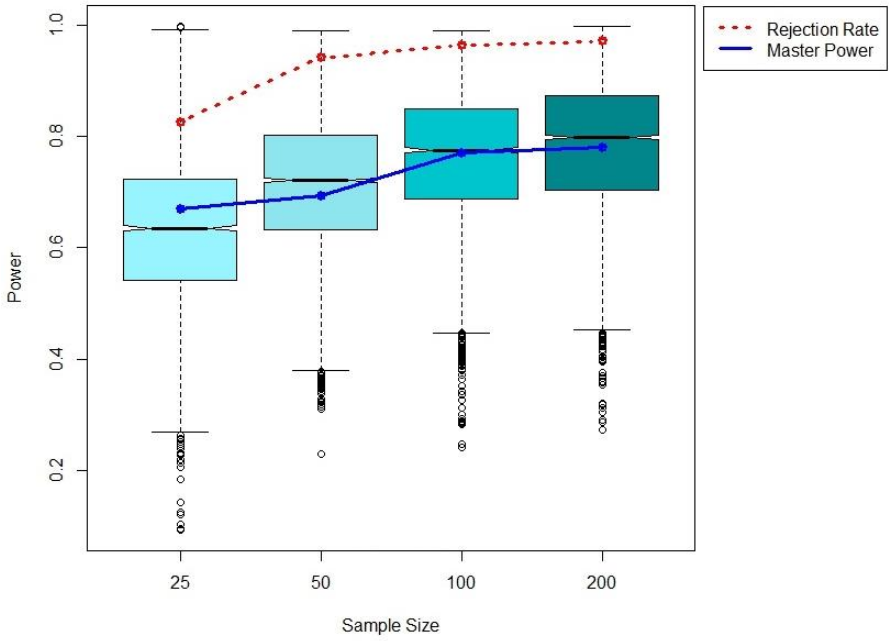

Figure 4.6. Distributions of Post-Hoc Powers for Different Sample Sizes Using Wald Test

Considering the main test investigated in this study is the Wald test, the rejection rates of the Wald were explored. The rejection rates being higher than the theoretical powers and the average of the post-hoc powers for each sample size shows that the hypothesis tests are rejected more often than what they should be. This is due to the high values of the Wald statistics calculated for each of the 3,600 simulated data for each of the four sample sizes considered for this study. It shows that the Wald statistics probably do not follow the non-central chi-square distribution that they should follow under the alternative hypothesis, according to Hall (2005). To investigate this possibility, the Wald test statistics from the simulated data were examined to see what parameter of their distribution is different from the non-central chi-square distribution mentioned in Hall

(2005). Table 4.13 summarizes the Wald statistics calculated for the simulated data sets of different size.

Table 4.13

Wald Statistics for 3,600 Data Sets of Sizes 25, 50, 100, and 200

	Sample size of 25	Sample size of 50	Sample size of 100	Sample size of 200
Mean	5.51652	6.67186	7.52108	8.05377
Variance	3.861	3.921	4.899	6.504
25^{th} Percentiles	4.25167	5.28609	6.00113	6.21453
50^{th} Percentiles	5.31447	6.48601	7.37399	7.81154
100^{th} Percentiles	6.49833	7.90512	8.94792	9.60096

These Wald statistics were then plotted for each sample size using histograms. Figures 4.7, 4.8, 4.9, and 4.10 show the histograms of the Wald statistics for the simulated data of sample sizes of 25, 50, 100, and 200, respectively. The non-central distributions of the Wald statistics they theoretically are supposed to follow are plotted on the histograms using a dashed curve. The solid curve in to the left shows the central chi-square distribution these statistics are supposed to follow under the null hypotheses. Not having most of the histogram bars even close to the null curves, clearly suggests the null hypotheses should be rejected most of the time, which is true. Three vertical lines are also indicated on the histogram of the 3,600 Wald statistic values for each sample size. The first line from the left, which is in thicker than the rest of the lines, shows the critical value to which each Wald statistic was being compared and if the Wald statistics were higher than this critical value, the null hypothesis was rejected. It is obvious that most of

the Wald statistics were much higher than the critical value; this explains the high frequency of the times the null hypotheses were rejected. The second vertical line from the left specifies the average of the 3,600 Wald statistics. The third line from the left, which is dotted, shows the mean of the non-central chi-square distribution the Wald statistics should theoretically follow. It is obvious that the second and the third line are slightly different from each other but not too far away from one another. This shows that the means of the non-central chi-square distributions Wald statistics follow theoretically and empirically are almost the same. Looking at Figures 4.7 through 4.10, as sample size increases, the shape of the distribution of the Wald statistics clearly becomes wider. It should be noted that no matter how much their variance increases based on increasing sample size, the non-central chi-square distribution the Wald statistic should theoretically follow under the alternative hypothesis does not seem to fit well to the actual values resulting from the Wald test on the 3,600 replications. As displayed on these histograms, what leads to a high rejection rate is that most of the Wald statistics are more concentrated around the area which is to the right of the critical value. This shows that even though the mean of the test statistics seems to be close to the mean of the hypothetical non-central chi-square distribution, their variances are not equal to the variance of the non-central chi-square distribution they theoretically should follow.

As clearly observed, by the increase of the sample sizes, the variances of the population of Wald statistics seem to increase as well. The variances for sample sizes of 25, 50, 100, and 200 are, respectively, equal to 3.861, 3.921, 4.899, and finally 6.504, which agrees with what the histograms in Figures 4.7, 4.8, 4.9, and 4.10 illustrate. The mean-variance relationship that exists for the theoretical non-central chi-square

distribution does not exist in the same way for the current values of the Wald statistics but the trend shows the Wald statistics are getting closer to what it should be by increasing the sample size. This theoretical mean-variance relationship shows if there is a χ_1^2 distribution with the non-centrality parameter of $\lambda = 5.75$, which is the average non-centrality value for the datasets with sample sizes of 25. The variance in this case should be equal to 2(1+2(5.75))=25; however, this variance is much larger than the variance of 3.861 that is what the population of the calculated Wald statistics from data sets of size 25 produced. This is why the Wald statistic values were mostly larger than the critical value and were not spread enough toward the tails of the curves of the non-central chi-square distributions shown in Figure 4.7. The increase in the variance associated with the increase in sample size is promising and informative in providing the reason for having the empirical Wald statistics from the simulation study to not to follow the exact non-central distributions they should follow based on the proof by Newey and West (1987).

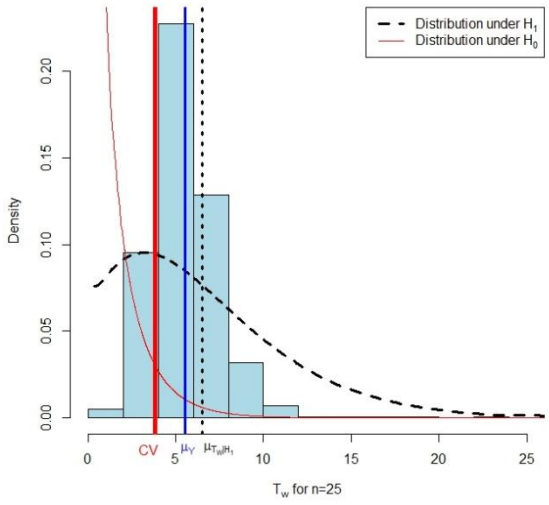

Figure 4.7. Distributions of the Wald Statistics for Sample Size of 25

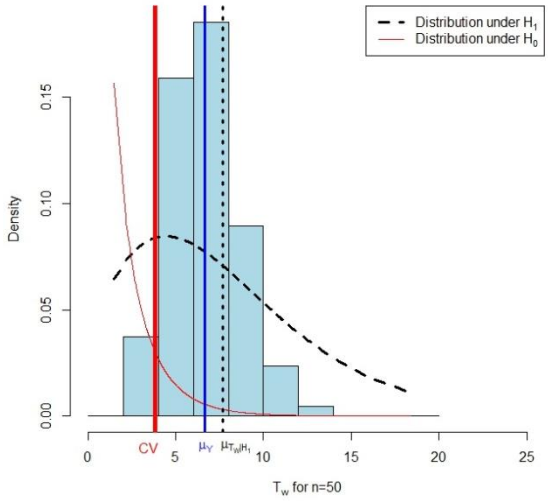

Figure 4.8. Distributions of the Wald Statistics for Sample Size of 50

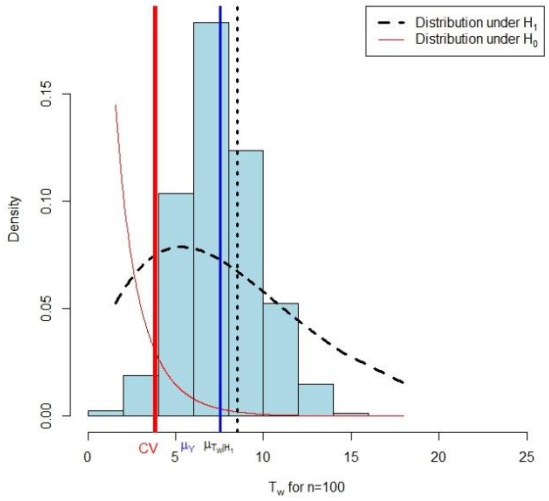

Figure 4.9. Distributions of the Wald Statistics for Sample Size of 100

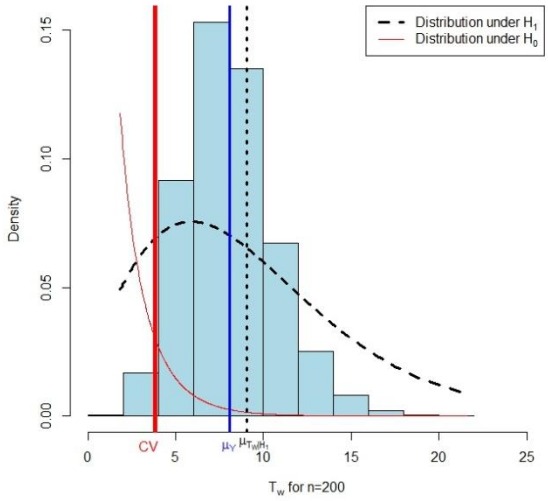

Figure 10. Distributions of the Wald Statistics for Sample Size of 200

Summary and Implications for the Power Estimation of Longitudinal Data with Time-Dependent Covariates Using Generalized Method of Moments Estimation

Estimating the GMM-based power is tough. At the same time, the methodology needed to be developed in a way that response values, effect sizes, parameter estimates, and the number of subjects were reflected in the estimated power since GMM-based power depends on all these criteria.

Two methods were developed in this dissertation for calculating the theoretical power of pilot data using GMM and due to the results each method provided and their comparison to the post-hoc powers calculated from a subsequent simulation study, the first method is the method I recommend especially when working with smaller sample sizes. The results of the simulation study clearly showed that the post-hoc powers were consistent with the calculated theoretical powers showing the accuracy of the power calculation technique developed in this dissertation. It is obvious that by the increase in

sample size, the theoretical powers get much closer to the measures of central tendency of the post-hoc powers.

According to the results from the simulation study, the estimated post-hoc powers increased with sample size and effect size and as expected, their average values were close to the theoretical powers calculated using the first method I proposed. This shows powers calculated based on the Wald statistic are distributed and behave similarly to the population theoretical powers. In contrast, the rejection rates were not close to the theoretical powers, which is due to not having a large enough sample size.

The DM statistic, which was adopted due to the claim by Hall (2005) regarding its having the same asymptotic distribution as the Wald statistic, did not perform the same as the Wald statistic and did not provide similar results to the Wald test. The rejection rates using the DM test were higher than the rejection rates using a Wald test. As explained above, this might be due to smaller sample sizes or violation of assumptions that were specified in Hall (2005). By smaller sample sizes, I mean the sample size of 200, with the data characteristics of OAI data, had not yet reached the size necessary to satisfy the asymptotic distributional assumptions for the two statistics to perform similarly. This causes the power calculation to be higher than it should be for lower sample sizes. Power still increases, as expected, with increase sample size, but it is inflated for lower sample sizes.

In summary, there need to be a much higher sample sizes for the empirical results to perform the same as the proposed theoretical methods and for now I recommend using the Wald statistic over the DM statistic for performing tests within longitudinal data with time-dependent covariates using the GMM estimation method.

CHAPTER V

CONCLUSIONS

Methods for estimation of statistical power for longitudinal data with time-dependent covariates using generalized method of moments (GMM) were developed in this dissertation. GMM was adopted within the power estimation techniques as the estimation method in order to provide more efficient estimates than generalized estimating equations (GEE) or restricted maximum likelihood (REML) used in the power estimation procedure when dealing with varying types of covariates.

The developed power estimation methods mainly focused on the use of the Wald statistic, which was proven to follow a chi-square distribution. The centrality or non-centrality of this distribution depends on whether the distribution of the Wald statistic is considered under the null or alternative hypothesis. Under the null hypothesis, this statistic follows an asymptotic central chi-square distribution; however, it follows a non-central chi-square distribution under the alternative hypothesis. The other statistic evaluated in this study is the distant metric (DM) statistic, which according to Hall (2005) should have the identical asymptotic chi-square distribution as the Wald statistic. Therefore, theoretically, the power estimation procedures developed in this dissertation based on the two statistics should perform similarly when the sample size is large and all the assumptions given by Hall (2005) are met.

The objective of the proposed methods and the results presented in this dissertation was to help applied researchers and practitioners, who design studies using real data, make valid decisions in terms of sample size selection. In turn, such decisions should result in an optimal sample size for their study, which results in an acceptable range of statistical power according to their discipline. The contribution of the proposed power estimation methods in this dissertation is that it is a new technique capable of coping with the use of time-dependent covariates in longitudinal modeling. A review of the literature on longitudinal modeling, GMM techniques, and power estimation techniques in Chapter II indicated that no known work had been done that applies the GMM estimation technique in the process of estimating power for repeated measures. This gap negatively affected this field of research in a way that the existing power estimation techniques were not general enough to efficiently involve time-dependent as well as time-independent covariates in a model. In Chapter III, these methods were theoretically developed and in Chapter IV, the performance of the proposed methods was evaluated. After validating these methods through real data analyses and simulation studies, the limitations of the proposed methodology were illustrated.

The power estimation technique introduced in this dissertation is different in the sense that there had not been any developed power estimation procedure that uses the GMM estimation technique and its related test statistics to estimate power for hypothesis tests for longitudinal data when dealing with time-dependent covariates. In previously developed techniques, covariates were assumed to stay constant throughout the study, which is not always realistic. The power estimation methods established in this paper, however, give researchers and practitioners the opportunity to use varying types of

covariates and still be able to estimate the statistical power of tests in their studies and predict optimal sample sizes for desired levels of power. The developed power estimation algorithm improves upon the other methods in that the process takes advantage of using a more efficient estimation technique (i.e., GMM) to capture the changes of the covariates over time.

The power estimation approach developed in this dissertation has two major advantages over previously developed power estimation methods for longitudinal models. First, the current power estimation method uses an estimation technique within its procedure that does not require any distributional assumptions, which can be helpful when dealing with data that do not meet the usual distributional assumptions. Second, GMM, which was used in this study, uses a set of moment conditions to take into account the autocorrelation among subjects and the time varying nature of some of the covariates. On the other hand, the GEE-based power estimation approach is subject to some criticisms because of forcing all the covariates used in a model to remain constant throughout the study, which will result in some loss of information, hence the reduced efficiency of the results.

The performance of the proposed methodology was tested in a simulation study as well as in applications using a pre-existing data set consisting of osteoarthritis initiative (OAI) data from a multi-center study on osteoarthritis of the knee. The results regarding the accuracy of performance of the developed power techniques and the situations that would affect their performance in application were tabulated and discussed in Chapter IV.

I developed the simulation scheme by borrowing information from the real OAI data to ensure that the results of the simulation study are generalizable to real data analysis; hence, the methodology could be adopted by researchers in different fields when using real data with unexpected behavior over time. Using this scheme rather than controlling for the distribution of all the covariates used in the simulation study made it more difficult to meet all the assumptions but, on the other hand, it resulted in a valuable gain in generalizability of the methodology when evaluating the performance of the theoretically developed methods in dealing with real data. This aided in providing helpful guidelines for practitioners regarding the situations that might arise in real data analysis when the methods might not perform as well as what was claimed in theory.

Furthermore, the simulation study clarified the accuracy of the power estimation method using the Wald test and the fact that using the DM technique may be erroneous when sample sizes are smaller and the random vectors of data do not necessarily form a strict stationary and ergodic process. In such cases, these techniques do not necessarily perform as expected. In order to improve the accuracy of the estimated power, different recommendations, such as increasing the sample size and sub-sampling or simulating data in the process of calculating the statistical power for future studies following certain steps, are provided.

To validate the results obtained from the simulation process, the hypothesis tests using the Wald statistic as well as the DM statistic were conducted on 3,600 simulated data sets. Different results from the simulation study, such as the rejection rates, test statistic values, and post-hoc power, were compared to the values calculated from the population data and across sample sizes within the simulation. The simulation study

showed that the average post-hoc power lines up with the theoretical power, using the procedure developed in this dissertation, for different sample sizes. However, the rejection rates are much higher than the theoretical powers for the smaller sample sizes considered in this study.

The first research question addressed the process of estimating the statistical power for longitudinal data in the presence of time dependent covariates using the Wald approach within a GMM estimation technique. The GMM estimation used within the Wald test was combined with the power estimation process to find the power of hypothesis tests using such data. This question was theoretically answered in Chapter III and the main steps leading to the results of applying the methods to real data are demonstrated in Chapter IV. The results obtained from the post-hoc power calculation of the simulated data and by comparing their distribution to the theoretical powers calculated from the pilot data showed that the powers calculated based on the Wald statistic are distributed and behave, similarly to the population theoretical powers. As expected, the estimated post-hoc powers increased with the increase of sample size and effect size. The accuracy of these powers was enhanced by the increase of sample size.

The applied methods developed in this study to address the second and fourth research questions were the biggest contributions made in Chapter IV of this dissertation to provide easy directions for applied researchers to find out the optimal sample size and power for their studies. Two methods of estimating the theoretical power for different sample sizes, leading to optimal sample size for the desired power for each study, were developed. Then, the first method was assessed and it provided results that were closer to

the results from the simulation study, making it the preferred option to be adopted by researchers. Briefly, the process is as follows.

After defining the appropriate model for each study and determining the hypothesis to be tested, the true effects of the alternative and sample sizes of interest need to be determined. If the pilot data set is larger than the sample sizes of interest, sub-samples of covariates and outcomes need to be taken. If effect sizes for the study were decided to alter from the original model, the new outcomes must be generated to reflect these effects. On the other hand, if the pilot data set is smaller than the sample sizes of interest, data sets of the sizes of interest need to be randomly generated using similar characteristics of the pilot data set. So, either pilot data or generated data are always needed within this method. Then the programs I wrote, or any other software, can be used to obtain the non-centrality parameters for all sub-samples, or simulated samples for the second scenario.

What differentiates the first method from the second method is the use of the GMM estimated parameters from each data set in the process of finding the non-centrality parameter for the respective data set. The first method produces estimates for each sub-sampled or simulated data. On the other hand, the second method uses the parameter estimates from the original pilot dataset in finding the non-centrality parameter for all different sub-sampled or simulated data sets. Using the average of all the non-centrality parameters, the power of the study can be calculated using the same procedure that was used when answering the first and third research questions.

The third research question was addressed the same way as the first question with the only difference being the different statistic used in the testing process. Within this

procedure, the DM statistic was adopted instead of the Wald statistic and the rest of the steps stayed the same due to the identical asymptotic distributions of these two statistics. The results obtained from the post-hoc power calculation of the simulated data revealed that the DM statistic, which was adopted from Hall (2005), did not perform the same as the Wald statistic and did not provide similar results to the Wald statistic and Wald test. These differences might be due to smaller sample sizes or violation of assumptions. If assumption violation is the case, the violated assumptions most likely are the violation of ergodicity or stationarity of data. I recommend using the Wald statistic over the DM statistic for performing tests within longitudinal data with time-dependent covariates using the GMM estimation method.

The simulation study was also used to answer the last research question regarding the behavior of the proposed method under varying sample sizes and the comparison of its results to the empirical results regarding power. Comparisons of the rejection rates of the simulated study and the estimated theoretical powers of the pilot dataset for different sample sizes was used to evaluate the behavior of the developed power estimation methodology. The results varied depending upon the sample sizes used within this study but they all agreed in one respect, which is the need for a higher sample size for the empirical results to perform exactly the same as the proposed theoretical methods. It is concluded that the methods which were theoretically proven to work in Chapter III for estimating the power do not perfectly work for sample sizes of 200 or smaller but it is shown that, even within these smaller sample sizes, as sample size increases, the results get closer to the theoretical expectations.

For the application considered in this dissertation, the developed methods were applied to a biomedical data set. However, these methods can be applied to any discipline or area of research as long as the model and hypothesis tests are correctly specified, the assumptions are met, and the sample sizes are large enough for the statistical tests to follow the asymptotic distributions they are supposed to follow in line with the theoretical proofs.

Limitations and Future Research

Though this dissertation investigated the power estimation methods of a specific type of longitudinal data, the methodology can be applied to a wide range of data types and models. This encourages future work in this area due to its potential generalizability to different models. The results also highlight the fact that the research line on power estimation and sample size calculation using GMM within longitudinal models that deal with varying types of covariates is not closed.

Limitations such as smaller sample sizes used for this study and the lengthy run time are acknowledged and therefore are areas of future research to explore. The sample size limitation is believed to be the main reason for the differences in the final results in Chapter IV compared to what was expected based on the theory developed in Chapter III. The run time was the main reason I could not extend the work to larger sample sizes for the current study; however, advances in technology and using more powerful computers will help in investigating the performance of the developed methods for larger sample sizes.

I aim to continue with the extension of this research line in several areas, including, but not limited to, extending the current methods to varying types of response

variables such as binary and categorical responses, applying these models to other types of time-dependent covariates, as well as developing R packages that can handle balanced data, meaning the circumstances where different numbers of follow-up times for the repeated measurements of the longitudinal studies are involved.

Extending these models to unbalanced data would enable researchers to estimate statistical power for circumstances where not every subject's measurement is recorded for every follow-up time. These situations arise in different areas such as biomedical studies when patients do not show up for every follow-up visit to their physician's office or hospital; in education when students drop out of school or do not take every exam while being evaluated at the end of a school year; in social research when not everyone fills out every survey throughout a study and, in general, in every field that involves multiple measurement of the same subject and not every measurement can be recorded over the period of study.

Extending this methodology to different types of outcome variables is another area of interest that can greatly benefit applied practitioners working with varying types of responses. I plan to adopt binary logistic models when dealing with dichotomous responses and borrow the theory from ordinal or multinomial models when predicting categorical responses, then apply them along with the power estimation techniques developed in this dissertation to build models that are more general.

Adopting and extending the developed techniques for data with other types of time-varying covariates could also provide valuable information for researchers testing for types I, III, and IV time-dependent covariates. Within this dissertation, the main hypothesis was tested on a type II time-dependent covariate; however, researchers might

be interested in testing other types of time-dependent covariates. Investigating the performance of the developed methods within this dissertation on the other types of time-dependent covariates and extending them, if necessary, will add to the body of research, making it possible for researchers to test other types of covariates as well while being able to estimate power for their models.

Although GMM power estimation is tough to calculate due to the fact that responses, effect sizes, parameter estimates, and the number of subjects are reflected in the estimated power, the developed methods add options in being able to estimate power for longitudinal data with time-dependent covariates in different fields. Writing packages and manuals in R for each of the models I worked on in the current dissertation and am planning to continue to pursue in the future, will help practitioners to easily use these techniques to design studies with optimal power and minimum sample size.

REFERENCES

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley.

Bain, L. J., & Engelhardt, M. (2009). *Introduction to probability and mathematical statistics*. Belmont, CA: Brooks/Cole Cengage Learning.

Chaussé, P. (2010). Computing generalized method of moments and generalized empirical likelihood with R. *Journal of Statistical Software, 34*(11), 1-35.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155.

Cullen, A. C., & Frey, H. C. (1999). *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*. New York, NY: Plenum Press.

Diggle, P., Liang, K. Y., & Zeger, S. L. (1994). *Longitudinal data analysis.* New York: Oxford University Press, 5, 13.

Erickson, T., & Whited, T. M. (2002). Two-step GMM estimation of the errors-in-variables model using high-order moments. *Econometric Theory, 18*(03), 776-799.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.). (2009). *Longitudinal data analysis*. Boca Raton: CRC Press.

Fitzmaurice, G. M., Laird, N. M., & Rotnitzky, A. G. (1993). Regression Models for Discrete Longitudinal Responses. *Statistical Science, 8*(3), 284-299.

Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling: An International Journal, 1*(3), 177-193.

Gupta, A. K., & Nadarajah, S. (2004). *Handbook of beta distribution and its applications*. New York, NY: Marcel Dekker.

Hall, A. R. (2005). Generalized method of moments. Oxford: Oxford University Press.

Hanfelt, J. J., & Liang, K. Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika,* 82(3), 461-477.

Hansen, L. P. (1982), Large-Sample Properties of Generalized Method of Moment Estimators, *Econometrica,* 50, 1029–1054.

Hansen, L. P. (2007). Generalized Method of Moments Estimation. *The New Palgrave Dictionary of Economics,* 1-10.

Hansen, L. P., Heaton, J., & Yaron, A. (1996). Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business & Economic Statistics, 14*(3), 262.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association,* 72(358), 320-338.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*(1), 19-24.

Hu, F.C. (1993) A statistical methodology for analyzing the causal health effect of a time-dependent exposure from longitudinal data. ScD dissertation. Department of Biostatistics, Harvard School of Public Health, Boston.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), 696-791.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 805-820.

Kraemer, H. C., & Blasey, C. (2015). How many subjects?: Statistical power analysis in research. Sage Publications.

Lai, T. L., & Small, D. (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1), 79-99.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 963-974.

Lalonde, T. L., Wilson, J. R., & Yin, J. (2014). GMM logistic regression models for longitudinal data with time-dependent covariates and extended classifications. *Statistics in medicine*, 33(27), 4756-4769.

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

Liu, G., & Liang, K. Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics*, 937-947.

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

Liang, K. Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3-40.

Lipsitz, S. R., Fitzmaurice, G. M., Orav, E. J., & Laird, N. M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics*, 270-278.

Lyles, R. H., Lin, H. M., & Williamson, J. M. (2007). A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Statistics in Medicine*, 26(7), 1632-1648.

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (Vol. 37). CRC press.

Newey, W. and McFadden, D. (1994), "Estimation in Large Samples," in The Handbook of Econometrics, Vol. 4, eds. D. McFadden and R. F. Engle, Amsterdam: North Holland.

Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777-787.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 1033-1048.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52(3/4), 447-458.

Ravishanker, N., & Dey, D. K. (2002). A first course in linear model theory. CRC Press.

Robert, C., & Casella, G. (2013). Monte Carlo statistical methods. Springer Science & Business Media.

Rochon, J. (1998). Application of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in medicine*, 17(14), 1643-1658.

Self, S. G., & Mauritsen, R. H. (1988). Power/sample size calculations for generalized linear models. *Biometrics*, 79-86.

Sullivan Pepe, M., & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in Statistics-Simulation and Computation, 23(4), 939-951.

Thompson, S. K. (2012). Sampling. Hoboken, NJ: Wiley.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models and the Gauss—Newton method. *Biometrika*, 61(3), 439-447.

Williams, R. L. (1995). Product-limit survival functions with correlated survival times. Lifetime data analysis, 1(2), 171-186.

Wishart, J. (1938). Growth-rate determinations in nutrition studies with the bacon pig and their analysis. *Biometrika*, 30(1/2), 16-28.

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 121-130.

Zeger, S. L. and Liang, K.Y. (1991) Feedback models for discrete and continuous time series. *Statist. Sin.*, 1, 51–64.

Zorn, C. J. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 470-490.

APPENDIX A

DATA GENERATION CODE

```
N=c(25,50,100,200)

OAIModel = lmer(WOMAC1 ~ Sex+Age+BMI+t2+t3+(1|ID), data= Dataset, REML=TRUE)
beta<-(summary(OAIModel))$coefficients[,1]

ids<-sqldf::sqldf("select distinct `ID` from Dataset")

samples<-data.frame('n'=c(0), 'run'=c(0), 'ID'=c(0), 'Sex'=c(0), 'Index1'=c(0), 'Age'=c(0), 'BMI'=c(0), 'WOMAC'=c(0), 'Sex'=c(0),
'WOMAC1'=c(0), 't2'=c(0), 't3'=c(0), 'predicted2'=c(0), 'predicted3'=c(0), 'predicted5'=c(0), 'predicted10'=c(0), 'predicted15'=c(0),
'predicted20'=c(0), 'predicted25'=c(0), 'predicted30'=c(0), 'predicted40'=c(0), 'predicted50'=c(0), 'time'=c(0), 'ID2'=c(0))

for (j in 1:4)
{
 for (i in 1:3,600)
 {
  set.seed(seed=i^2+7)
  subjects<-sample(x = t(ids), size=as.numeric(N[[j]]), replace=FALSE)
  X<-Dataset[which(Dataset$`ID` %in% subjects),]
  errors<-rnorm(n=length(X[,1]), mean=0, sd=0.5444)
  RandomIntercept<-rnorm(n=length(X[,1])/3, mean=0, sd=0.7503)
  errorREP<-rep(RandomIntercept, each=3)

  values1<-beta[1]+ beta[2]*X[["Sex"]]+beta[3]*X[["Age"]]+2*beta[4]*X[["BMI"]]+beta[5]*X[["t2"]]+beta[6]*X[["t3"]]
  values2<-beta[1]+ beta[2]*X[["Sex"]]+beta[3]*X[["Age"]]+3*beta[4]*X[["BMI"]]+beta[5]*X[["t2"]]+beta[6]*X[["t3"]]
  values3<-beta[1]+ beta[2]*X[["Sex"]]+beta[3]*X[["Age"]]+5*beta[4]*X[["BMI"]]+beta[5]*X[["t2"]]+beta[6]*X[["t3"]]
  values4<-beta[1]+ beta[2]*X[["Sex"]]+beta[3]*X[["Age"]]+10*beta[4]*X[["BMI"]]+beta[5]*X[["t2"]]+beta[6]*X[["t3"]]
  values5<-beta[1]+ beta[2]*X[["Sex"]]+beta[3]*X[["Age"]]+15*beta[4]*X[["BMI"]]+beta[5]*X[["t2"]]+beta[6]*X[["t3"]]
  values6<-beta[1]+ beta[2]*X[["Sex"]]+beta[3]*X[["Age"]]+20*beta[4]*X[["BMI"]]+beta[5]*X[["t2"]]+beta[6]*X[["t3"]]
  values7<-beta[1]+ beta[2]*X[["Sex"]]+beta[3]*X[["Age"]]+25*beta[4]*X[["BMI"]]+beta[5]*X[["t2"]]+beta[6]*X[["t3"]]
  values8<-beta[1]+ beta[2]*X[["Sex"]]+beta[3]*X[["Age"]]+30*beta[4]*X[["BMI"]]+beta[5]*X[["t2"]]+beta[6]*X[["t3"]]
  values9<-beta[1]+ beta[2]*X[["Sex"]]+beta[3]*X[["Age"]]+40*beta[4]*X[["BMI"]]+beta[5]*X[["t2"]]+beta[6]*X[["t3"]]
  values10<-beta[1]+ beta[2]*X[["Sex"]]+beta[3]*X[["Age"]]+50*beta[4]*X[["BMI"]]+beta[5]*X[["t2"]]+beta[6]*X[["t3"]]

  predicted2<-values1+errors+errorREP
  predicted3<-values2+errors+errorREP
  predicted5<-values3+errors+errorREP
  predicted10<-values4+errors+errorREP
  predicted15<-values5+errors+errorREP
  predicted20<-values6+errors+errorREP
  predicted25<-values7+errors+errorREP
  predicted30<-values8+errors+errorREP
  predicted40<-values9+errors+errorREP
  predicted50<-values10+errors+errorREP

  sample.run<-data.frame(cbind('n'=c(rep(N[[j]], N[[j]]*3)), 'run'=c(rep(i, N[[j]]*3)), X, predicted2, predicted3, predicted5,
            predicted10, predicted15, predicted20, predicted25, predicted30, predicted40, predicted50))
  samples=rbind(samples, sample.run)
 }
}
toc()
```

APPENDIX B

GENERALIZED METHOD OF MOMENTS FUNCTION

```
QuadForm2=function(beta1, sampleset, n) # the new quadform
{
 G=c(rep(0,39))
 S=matrix(0,39,39)
 #n=nrow(sampleset)/3
 for (i in 1:n)
 {
  gets<-numeric()
  g11<-sampleset[((i*3) -2), ]; g11 <- unlist(g11)
  g1<-unlist(c(1, g11[c("Sex","Age", "BMI", "t2", "t3")])); g1 <- unlist(g1)
  g22<-sampleset[((i*3) -1), ]; g22 <- unlist(g22)
  g2<-unlist(c(1, g22[c("Sex","Age", "BMI", "t2", "t3")])); g2 <- unlist(g2)
  g33<-sampleset[((i*3)), ]; g33 <- unlist(g33)
  g3<-unlist(c(1, g33[c("Sex","Age", "BMI", "t2", "t3")])); g3 <- unlist(g3)

  #Intercept
  gets[1]=g11["predicted"]-g1%*%beta1
  gets[2]=g22["predicted"]-g2%*%beta1
  gets[3]=g33["predicted"]-g3%*%beta1

  #Age
  gets[4]=g1["Age"]*(g11["predicted"]-g1%*%beta1)
  gets[5]=g1["Age"]*(g22["predicted"]-g2%*%beta1)
  gets[6]=g1["Age"]*(g33["predicted"]-g3%*%beta1)

  gets[7]=g2["Age"]*(g11["predicted"]-g1%*%beta1)
  gets[8]=g2["Age"]*(g22["predicted"]-g2%*%beta1)
  gets[9]=g2["Age"]*(g33["predicted"]-g3%*%beta1)

  gets[10]=g3["Age"]*(g11["predicted"]-g1%*%beta1)
  gets[11]=g3["Age"]*(g22["predicted"]-g2%*%beta1)
  gets[12]=g3["Age"]*(g33["predicted"]-g3%*%beta1)
```

```
 #Sex
gets[13]=g1["Sex"]*(g11["predicted"]-g1%*%beta1)
gets[14]=g2["Sex"]*(g22["predicted"]-g2%*%beta1)
gets[15]=g3["Sex"]*(g33["predicted"]-g3%*%beta1)

#BMI
gets[16]=g1["BMI"]*(g11["predicted"]-g1%*%beta1)
gets[17]=g2["BMI"]*(g11["predicted"]-g1%*%beta1)
gets[18]=g3["BMI"]*(g11["predicted"]-g1%*%beta1)

gets[19]=g2["BMI"]*(g22["predicted"]-g2%*%beta1)
gets[20]=g3["BMI"]*(g22["predicted"]-g2%*%beta1)

gets[21]=g3["BMI"]*(g33["predicted"]-g3%*%beta1)

#t2
gets[22]=g1["t2"]*(g11["predicted"]-g1%*%beta1)
gets[23]=g1["t2"]*(g22["predicted"]-g2%*%beta1)
gets[24]=g1["t2"]*(g33["predicted"]-g3%*%beta1)

gets[25]=g2["t2"]*(g11["predicted"]-g1%*%beta1)
gets[26]=g2["t2"]*(g22["predicted"]-g2%*%beta1)
gets[27]=g2["t2"]*(g33["predicted"]-g3%*%beta1)

gets[28]=g3["t2"]*(g11["predicted"]-g1%*%beta1)
gets[29]=g3["t2"]*(g22["predicted"]-g2%*%beta1)
gets[30]=g3["t2"]*(g33["predicted"]-g3%*%beta1)

#t3
gets[31]=g1["t3"]*(g11["predicted"]-g1%*%beta1)
gets[32]=g1["t3"]*(g22["predicted"]-g2%*%beta1)
gets[33]=g1["t3"]*(g33["predicted"]-g3%*%beta1)

gets[34]=g2["t3"]*(g11["predicted"]-g1%*%beta1)
gets[35]=g2["t3"]*(g22["predicted"]-g2%*%beta1)
gets[36]=g2["t3"]*(g33["predicted"]-g3%*%beta1)

gets[37]=g3["t3"]*(g11["predicted"]-g1%*%beta1)
```

```
    gets[38]=g3["t3"]*(g22["predicted"]-g2%*%beta1)
    gets[39]=g3["t3"]*(g33["predicted"]-g3%*%beta1)

    G=G + (gets)
    S=S + gets%*%t(gets)
  }
  G=G/n
  W=MASS::ginv((1/n)*S)
  QF=t(G)%*%W%*%G
}
```

APPENDIX C

POWER FUNCTION

```
rep <- 100
n <- 25
pb <- txtProgressBar(style = 3)
results25 <- data.frame(Lambdabhatn = rep(NA,rep), powerbhatn = rep(NA,rep))
for(i in 1:rep){
   ids <- sample(x = unique(Dataset$ID), size = n, replace = FALSE)
   sampleset <- Dataset[Dataset$ID %in% ids,] #working data
   beta1 <- gee::gee(predicted ~ Sex+Age+BMI+t2+t3, id = ID, data = sampleset, corstr = "independence")$coefficients
   betahat <- optim(beta1, QuadForm2)$par
   QQ1<-QuadForm2(betahat)
   results25$Lambdabhatn[i]<-n*betahat[4]*QQ1*betahat[4]
   results25$powerbhatn[i]<-pchisq(q=CV, ncp=results25$Lambdabhatn[i], df=1, lower.tail = F)
   setTxtProgressBar(pb, i/rep)
 }
close(pb)
head(results25); mean(results25$powerbhatn)

# Randomly sample 50 observations 100 times
n <- 50
pb <- txtProgressBar(style = 3)
results50 <- data.frame(Lambdabhatn = rep(NA,rep), powerbhatn = rep(NA,rep))
for(i in 1:rep){
   ids <- sample(x = unique(Dataset$ID), size = n, replace = FALSE)
   sampleset <- Dataset[Dataset$ID %in% ids,] #working data
   beta1 <- gee::gee(predicted ~ Sex+Age+BMI+t2+t3, id = ID, data = sampleset, corstr = "independence")$coefficients
   betahat <- optim(beta1, QuadForm2)$par
   QQ1<-QuadForm2(betahat)
   results50$Lambdabhatn[i]<-n*betahat[4]*QQ1*betahat[4]
   results50$powerbhatn[i]<-pchisq(q=CV, ncp=results50$Lambdabhatn[i], df=1, lower.tail = F)
   setTxtProgressBar(pb, i/rep)
 }
close(pb)
head(results50); mean(results50$powerbhatn)

# Randomly sample 100 observations 100 times
n <- 100
pb <- txtProgressBar(style = 3)
results100 <- data.frame(Lambdabhatn = rep(NA,rep), powerbhatn = rep(NA,rep))
for(i in 1:rep){
   ids <- sample(x = unique(Dataset$ID), size = n, replace = FALSE)
   sampleset <- Dataset[Dataset$ID %in% ids,] #working data
   beta1 <- gee::gee(predicted ~ Sex+Age+BMI+t2+t3, id = ID, data = sampleset, corstr = "independence")$coefficients
   betahat <- optim(beta1, QuadForm2)$par
   QQ1<-QuadForm2(betahat)
   results100$Lambdabhatn[i]<-n*betahat[4]*QQ1*betahat[4]
   results100$powerbhatn[i]<-pchisq(q=CV, ncp=results100$Lambdabhatn[i], df=1, lower.tail = F)
   setTxtProgressBar(pb, i/rep)
 }
close(pb)
head(results100); mean(results100$powerbhatn)
# Randomly sample 200 observations 100 times
n <- 200
pb <- txtProgressBar(style = 3)
results200 <- data.frame(Lambdabhatn = rep(NA,rep), powerbhatn = rep(NA,rep))
for(i in 1:rep){
   ids <- sample(x = unique(Dataset$ID), size = n, replace = FALSE)
   sampleset <- Dataset[Dataset$ID %in% ids,] #working data
   beta1 <- gee::gee(predicted ~ Sex+Age+BMI+t2+t3, id = ID, data = sampleset, corstr = "independence")$coefficients
   betahat <- optim(beta1, QuadForm2)$par
   QQ1<-QuadForm2(betahat)
   results200$Lambdabhatn[i]<-n*betahat[4]*QQ1*betahat[4]
```

```
   results200$powerbhatn[i]<-pchisq(q=CV, ncp=results200$Lambdabhatn[i], df=1, lower.tail = F)
   setTxtProgressBar(pb, i/rep)
  }
close(pb)
head(results200); mean(results200$powerbhatn)
printout <- (cbind(results25, results50, results100, results200))
colnames(printout) <- c("25_Lambdabhatn",
              "25_powerbhatn","50_Lambdabhatn","50_powerbhatn","100_Lambdabhatn","100_powerbhatn","200_Lambdabhatn","200
              _powerbhatn")
write.csv(printout, file = "theoretical power results rep 5 by BMI 100_n25to200_GEE quadratic each sample_6.15.2017.csv
```